Modèles discrets et schémas itératifs

APPLICATION AUX ALGORITHMES MULTIGRILLES ET MULTIDOMAINES

Jean-Antoine DÉSIDÉRI
desideri@sophia.inria.fr

7 septembre 1998

Table des matières

Avant-propos

Ces vingt-cinq dernières années, on a assisté à une extraordinaire explosion du calcul scientifique et de ses applications, dont l'impact dans différents milieux industriels a largement dépassé le cadre des études amont [26]. On peut se demander quels ont été les progrès scientifiques et techniques, ou même ceux liés à la technologie des ordinateurs qui ont permis de telles avancées.

Toutes disciplines applicatives confondues, la remarquable montée en puissance des super-calculateurs, et aujourd'hui des stations de travail, constitue un élément incontestable. Cependant, les progrès de l'analyse numérique peuvent légitimement revendiquer une part aussi grande du mérite dans ce résultat.

En « Mécanique des Fluides Numérique », de nombreux progrès ont été réalisés grâce aux schémas d'approximation de type « volumes finis » ou s'en inspirant, permettant de résoudre de manière quasi-systématique la question de la conservativité pour les problèmes hyperboliques non linéaires, et la gestion beaucoup plus générale des maillages non structurés, élément clé de l'impact en milieu industriel [35]. Les divers « solveurs de Riemann » approchés aujourd'hui quasiment construits à la carte en fonction de l'application envisagée, contribuent également à qualifier le calcul scientifique comme outil majeur d'aide à la modélisation, et plus généralement à la validation des grands projets de « R & D » (recherche et développement), un thème resté central tout au long du projet européen Hermès, par exemple.

Parallèlement, les progrès des techniques d'approximation ont encouragé le développement d'algorithmes performants pour la résolution des grands systèmes linéaires ou non linéaires qui résultent de la discrétisation d'équations aux dérivées partielles (EDP) de la physique ou de l'ingénierie. Ces algorithmes sont de plus en plus « implicites » et généralisent dans un sens de plus en plus large le concept de « préconditionneur ». A ce titre, on doit citer les *méthodes implicites d'intégration pseudo-temporelle*, notamment celles de type « *Defect-Correction* » qui ont certaines propriétés de convergence indépendante du maillage, les *algorithmes multigrilles*, dont le coût théorique peut être proportionnel au nombre de degrés de liberté du problème discret, et les *méthodes par décomposition de domaine*. Si l'assise théorique des mé-

thodes multigrilles est désormais solidement bâtie, notamment dans un cadre variationnel [66], les méthodes multidomaines, bien qu'antérieures à l'origine, bénéficient aujourd'hui d'un regain d'intérêt, particulièrement en mécanique des structures et des fluides [61] [62] [63] [14], grâce au calcul parallèle [83] [11] [32] [39] [1] [2] [56]. Ces diverses approches ont entre elles des liens très étroits, comme l'indique plusieurs travaux récents dont [23] [9] [75].

Ces progrès en potentiel de résolution ouvrent la voie d'un champ élargi non seulement dans les disciplines spécialisées de l'ingénierie et comme outil d'aide à la modélisation, mais aussi en optimisation et en couplage multi-disciplinaire.

Cet ouvrage vise l'apprentissage de l'étudiant en analyse numérique ou de l'ingénieur en calcul scientifique aux algorithmes multigrilles et aux méthodes de résolution par décomposition de domaine. On y adopte le point de vue de l'algèbre linéaire et de l'analyse de Fourier. On a cherché à mettre en évidence les phénomènes numériques essentiels par l'analyse de modèles simplifiés le plus souvent linéaires. Plus de quarante exercices sont proposés avec leurs corrigés complets incluant de nombreuses illustrations. On pense ainsi permettre au lecteur moins averti de contrôler graduellement son assimilation du cours, notamment en l'alertant sur certaines difficultés que le praticien rencontre dans la mise en œuvre d'une expérimentation numérique ou l'analyse des observations faites dans son déroulement.

On a inclus un chapitre présentant des applications récentes de méthodes multigrilles en mécanique des fluides. Les calculs que l'on y présente ont été réalisés à l'INRIA Sophia-Antipolis au sein du Projet SINUS et on remercie leurs auteurs d'avoir autorisé leur reproduction partielle.

On espère ainsi que l'ouvrage sera utile à la pédagogie des méthodes, ainsi qu'à son utilisation par les praticiens.

L'auteur tient à remercier vivement les Professeurs G. Dhatt et O. Pironneau pour leurs conseils, ainsi que ses collègues de l'INRIA, en particulier A. Dervieux, H. Guillard, L. Hascoët et S. Lantéri et plus généralement tous les membres passés ou actuels du Projet SINUS, R. Peyret à l'Université de Nice et J. Périaux de Dassault Aviation, dont les commentaires scientifiques ont été très appréciés. Les étudiants de l'Université de Nice–Sophia-Antipolis qui ont « rodé » le cours ont beaucoup contribué aussi et ils en sont ici chaleureusement remerciés. R. Savalle a créé l'environnement informatique pour l'élaboration du manuscrit ; je le remercie sincèrement ainsi que P. Maleyran et C. Mercier pour leur aide dans la finalisation du document. Enfin, l'auteur souhaite témoigner de l'efficacité des services éditoriaux d'Hermès, ainsi que de son appréciation des encouragements de S. Gazel.

Nice, le 7 septembre 1998
Jean-Antoine Désidéri

Chapitre 1

Introduction

1.1. Aspects du calcul scientifique

Le traitement d'un problème de calcul scientifique présente généralement les principaux aspects suivants :

Modélisation physique

Le physicien ou l'ingénieur associe à une description qualitative d'un phénomène physique un modèle mathématique quantitatif dit « modèle physique » ou « modèle continu » qui consiste très souvent en un jeu d'équations aux dérivées partielles (EDP) soumis à des conditions aux limites (CL), initiales (CI) ou mixtes. Dans certains cas plus complexes, des équations ou inéquations d'autre nature (algébrique, intégrale, etc.) peuvent être adjointes à ce jeu sous la forme de contraintes supplémentaires. Dans ce cours, on considère presqu'exclusivement le cas du « problème aux limites » suivant :

$$\boxed{A\,u = f} \tag{1.1}$$

dans lequel A symbolise un opérateur aux dérivées partielles elliptique et linéaire, et f est une fonction donnée. La fonction u désigne la solution exacte de ce modèle continu.

Au chapitre 6, consacré à certaines applications en mécanique des fluides, on considérera notamment des opérateurs plus généraux.

Etude mathématique du modèle continu

Le « mathématicien appliqué » cherchera à « débroussailler » le problème en identifiant le cadre mathématique dans lequel l'existence, l'unicité et les propriétés essentielles de la solution peuvent être établies, au moins pour un modèle simplifié. Cet aspect relève de l'analyse fonctionnelle.

Construction et analyse d'une approximation discrète

En général, la solution exacte du modèle continu appartient à un espace de dimension infinie (généralement un sous-espace de L^2), et on sait rarement la représenter formellement au moyen d'un nombre fini de paramètres. En conséquence, pour se ramener à un problème de dimension finie résolvable sur ordinateur, on construit un système discret algébrique dit « modèle numérique » ou « modèle discret » qui approche l'EDP (et les CL/CI) dans un certain sens :

$$\boxed{A_h\, u_h = f_h} \tag{1.2}$$

Dans cette équation,

$$u_h = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \tag{1.3}$$

est le vecteur des N inconnues ou « degrés de liberté » associés à la représentation approchée de l'inconnue sur une discrétisation du domaine spatial dite « maillage » ou « grille » $\mathcal{M}_h$. L'indice h qui réfère à cette discrétisation, est un paramètre, parfois appelé « maille », caractéristique de la finesse du maillage, tel que le pas d'espace Δx dans le cas d'un problème unidimensionnel uniformément discrétisé. Dans le cas où plusieurs grilles seront considérées, on les distinguera par cet indice.

A_h est la *matrice d'approximation* (dimension $N \times N$).

f_h est un vecteur connu contenant la discrétisation de la fonction f et/ou des termes reflétant les conditions aux limites.

Parmi les méthodes d'approximation les plus classiques, notons les différences, volumes ou éléments fini(e)s, les méthodes spectrales et certaines méthodes probabilistes (méthodes de Monte-Carlo). On renvoie à d'autres cours pour une description

détaillée des méthodologies possibles d'approximation. Quelle que soit l'approche utilisée, l'analyste numéricien cherchera à démontrer (dans un cadre simplifié) la *convergence de l'approximation*, c'est-à-dire le fait que la solution du problème discret u_h a pour limite, au sens d'une norme appropriée, la solution du problème continu u lorsque la maille tend vers 0 ou lorsque le nombre de degrés de liberté tend vers l'infini (dans tout sous-domaine du calcul) :

$$\boxed{\lim_{h \to 0} u_h = u} \tag{1.4}$$

Cet aspect relève aussi de l'analyse fonctionnelle.

Résolution du problème discret

Une fois le système discret construit et justifié, il convient d'élaborer un algorithme de résolution de ce système sur ordinateur. Si le système algébrique est linéaire et bien conditionné, ce qui sous-entend notamment un nombre modéré de degrés de liberté, une inversion directe par élimination de Gauss est envisageable. Dans tout autre cas, on aura recours à une méthode itérative (Jacobi, Gauss-Seidel, GMRES, Newton, etc.) permettant de construire une suite infinie $\{u_h^n\}$ ($n = 0, 1, ...$) d'estimations de u_h. L'analyste numéricien s'attachera alors à prouver la *convergence itérative* de l'algorithme de résolution :

$$\boxed{\lim_{n \to \infty} u_h^n = u_h \quad (h \text{ fixé})} \tag{1.5}$$

Il cherchera également à évaluer la performance en coût, dite *efficacité de l'algorithme*. Cette tâche consiste à établir la relation entre le coût (en temps calcul ou en nombre d'opérations élémentaires) et la tolérance sur le degré de convergence itérative. Un problème distinct mais voisin est celui du contrôle des erreurs d'arrondi. Ces questions relèvent principalement de l'algèbre linéaire et l'analyse de Fourier discrète.

Adaptation aux architectures et validation des logiciels

Il s'agit ici d'aspects de nature informatique. Cependant aujourd'hui, dans le cas de problèmes « dimensionnants », c'est-à-dire proches des limites de la puissance de calcul disponible, la disponibilité d'ordinateurs vectoriels et/ou parallèles oriente systématiquement le choix ou la construction même de l'algorithme de résolution.

Dépouillement des résultats, CAO et synthèse

Le responsable du projet établira une synthèse des résultats contenant notamment une analyse critique des options prises concernant les aspects précédents.

Un point mérite d'être souligné : si ces différents aspects du problème viennent d'être présentés selon une logique séquentielle, en pratique, dans le cas d'un grand projet pour lequel des développements mais aussi de la recherche sont nécessaires, la « solution » ne peut être atteinte que par une coopération interactive des différents experts intervenant. Par exemple, le modèle physique n'est pas forcément arrêté *a priori* : l'étude mathématique peut révéler une insuffisance du modèle proposé initialement. En outre, la résolution numérique qui a plus facilement le potentiel de prendre en compte des modélisations fines hors de portée de l'analyse mathématique, est un atout majeur pour guider le choix du modèle nécessaire et/ou suffisant. Il est clair également que le choix de la méthode d'approximation est conditionné par celui de la technique de résolution, lui-même guidé par les caractéristiques de l'ordinateur utilisé. Enfin, l'analyse critique des résultats conduit généralement à une deuxième campagne d'étude dans laquelle les choix méthodologiques sont affinés.

Dans cet ouvrage, on supposera que des modèles continu et discret bien posés ont été construits, et on se focalisera sur l'analyse de certaines techniques de résolution. Afin de rendre l'exposé aussi concret que possible, on examinera presque toujours le cas particulier du modèle discret unidimensionnel que l'on présente dans la section suivante, avant de discuter des différentes formes d'erreur introduites dans les phases successives de la mise en œuvre d'un problème.

1.2. Modèle discret fondamental unidimensionnel

Le « problème de Poisson » constitué de l'« équation de Laplace » (ici avec second membre),

$$-\Delta u = f \quad (u \in H^1(\Omega),\ f \in L^2(\Omega)) \tag{1.6}$$

soumise à des conditions de Dirichlet,

$$u = g \quad (\text{sur } \partial\Omega) \tag{1.7}$$

fournit un prototype couramment utilisé dans l'analyse des problèmes aux limites. Dans le cas d'une seule variable d'espace x, si les conditions aux limites sont homo-

gènes, ce prototype se réduit au suivant :

$$\boxed{\begin{array}{c} -u_{xx} = f \quad (0 < x < 1) \\ u(0) = u(1) = 0 \end{array}} \tag{1.8}$$

que nous considérerons comme le « modèle continu unidimensionnel fondamental ».

La discrétisation par différences finies centrées de ce modèle unidimensionnel sur un maillage uniforme :

$$x_j = j\,h\,,\ h = \Delta\,x = \frac{1}{M+1} \quad (j = 0, 1, 2, ..., M+1) \tag{1.9}$$

est très classique :

$$\boxed{\begin{array}{c} \dfrac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} = f_j \quad (j = 1, 2, ..., M) \\ u_0 = u_{M+1} = 0 \end{array}} \tag{1.10}$$

Le système discret qui en résulte lorsqu'on rassemble ces équations linéaires a la forme indiquée en (1.2) dans laquelle la matrice d'approximation A_h a la structure tridiagonale suivante :

$$A_h = \frac{1}{h^2}\,\mathrm{Trid}_{D\,D}\,(-1, 2, -1) = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \tag{1.11}$$

où l'indice « *DD* » affecté au symbole de matrice tridiagonale a pour but de rappeler que l'on a supposé des conditions aux limites de Dirichlet aux 2 limites du domaine spatial [-1,1] de l'étude.

Au chapitre 2, on examinera les propriétés spectrales des opérateurs de différences finies dans de nombreux cas. Sans attendre certains résultats généraux, dans le but d'illustrer certaines notions grâce au modèle fondamental discret unidimensionnel, on établit dès à présent la diagonalisation de la matrice d'approximation A_h de (1.11) par le théorème suivant :

Théorème 1.1 (Diagonalisation du modèle discret fondamental unidimensionnel)

La matrice d'approximation tridiagonale A_h, explicitée en (1.11), qui est associée au modèle discret fondamental unidimensionnel, (1.10), admet la diagonalisation suivante :

$$\boxed{A_h = \frac{1}{h^2}\, S_h\, \Lambda_h\, S_h^{-1}} \tag{1.12}$$

où la matrice S_h est orthogonale et symétrique (donc involutive) :

$$S_h^{-1} = S_h^T = S_h = \{S_{j,m}\} \tag{1.13}$$

la matrice Λ_h est diagonale :

$$\Lambda_h = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_M \end{pmatrix} \tag{1.14}$$

et l'on a précisément :

$$\boxed{\begin{array}{c} S_{j,m} = \sqrt{2h}\, \sin\, j\theta_m \\ \lambda_m = 2 - 2\, \cos\, \theta_m \\ (j, m = 1, 2, ..., M) \end{array}} \tag{1.15}$$

où :

$$\boxed{\theta_m = \frac{m\pi}{M+1} = m\pi h} \tag{1.16}$$

est communément appelé « paramètre de fréquence ».

DÉMONSTRATION : vérifions d'abord que la matrice S_h définie dans l'énoncé du théorème est une matrice orthogonale. Pour cela, notons

$$S^{(m)} \stackrel{\text{déf}}{=} (S_{j,m}) \quad (j, m = 1, 2, ..., M) \tag{1.17}$$

ses vecteurs colonnes et vérifions qu'ils sont 2 à 2 orthogonaux et de norme euclidienne égale à 1. On calcule donc le produit scalaire suivant pour tout couple d'indices

(m,ℓ) $(m,\ell = 1,2,...,M)$:

$$\begin{aligned}(S^{(m)}, S^{(\ell)}) &= \sum_{j=1}^{M} S_j^{(m)}\, S_j^{(\ell)} \\ &= \sum_{j=1}^{M} S_{j,m}\, S_{j,\ell} \\ &= 2h \sum_{j=1}^{M} \sin\, j\,\theta_m \,\sin\, j\,\theta_\ell \\ &= h \sum_{j=1}^{M} \Big(\cos\, j\,(\theta_m - \theta_\ell) - \cos\, j\,(\theta_m + \theta_\ell)\Big) \\ &= h\,(\mathcal{C}_{m-\ell} - \mathcal{C}_{m+\ell})\end{aligned} \tag{1.18}$$

où l'on a posé pour tout indice k :

$$\mathcal{C}_k \overset{\text{déf}}{=} \sum_{j=1}^{M} \cos\, j\theta_k \tag{1.19}$$

On pose également :

$$\mathcal{S}_k \overset{\text{déf}}{=} \sum_{j=1}^{M} \sin\, j\theta_k \tag{1.20}$$

et

$$\mathcal{E}_k \overset{\text{déf}}{=} \sum_{j=1}^{M} \exp\,(ij\theta_k) = \mathcal{C}_k + i\,\mathcal{S}_k \tag{1.21}$$

Il vient, pour $\theta_k \neq 0\ [\text{mod}\, 2\pi]$:

$$\begin{aligned}\mathcal{E}_k &= \exp\,(i\,\theta_k) \sum_{p=0}^{M-1} \exp\,(pi\theta_k) \\ &= \exp\,(i\,\theta_k)\ \frac{1-\exp\,(Mi\theta_k)}{1-\exp\,(i\theta_k)} \\ &= \exp\Big(i(M+1)\frac{\theta_k}{2}\Big)\ \frac{\sin\,\frac{M\theta_k}{2}}{\sin\,\frac{\theta_k}{2}}\end{aligned} \tag{1.22}$$

puis

$$\mathcal{C}_k = \cos\,(M+1)\frac{\theta_k}{2}\ \frac{sin\, M\frac{\theta_k}{2}}{\sin\,\frac{\theta_k}{2}} \quad (\theta_k \neq 0\ [\text{mod}\, 2\pi]) \tag{1.23}$$

Enfin, notons que :

$$\cos(M+1)\frac{\theta_k}{2} = \cos\ \frac{k\pi}{2} \tag{1.24}$$

et

$$\begin{aligned} sin\, M\tfrac{\theta_k}{2} &= \sin\ \left((M+1)\frac{\theta_k}{2} - \frac{\theta_k}{2}\right) \\ &= \sin\ \left(\frac{k\pi}{2} - \frac{\theta_k}{2}\right) \\ &= \sin\ \frac{k\pi}{2}\ \cos\ \frac{\theta_k}{2} - \cos\ \frac{k\pi}{2}\ \sin\ \frac{\theta_k}{2} \end{aligned} \tag{1.25}$$

Par conséquent, pour $\theta_k \neq 0\ [\text{mod}\ 2\pi]$, on a :

$$\mathcal{C}_k = \cos\ \frac{k\pi}{2}\ \frac{\sin\ \frac{k\pi}{2}\ \cos\ \frac{\theta_k}{2} - \cos\ \frac{k\pi}{2}\ \sin\ \frac{\theta_k}{2}}{\sin\ \frac{\theta_k}{2}} \tag{1.26}$$

ce qui donne :

$$\mathcal{C}_k = \begin{cases} 0 & \text{si } k \text{ est impair} \\ -1 & \text{si } k \text{ est pair (et non nul)} \end{cases} \tag{1.27}$$

Par contre si $\theta_k = 0\ [\text{mod}\ 2\pi]$, on a immédiatement :

$$\mathcal{C}_k = \mathcal{C}_0 = M \tag{1.28}$$

Revenons au produit scalaire de (1.18). Si les indices m et ℓ sont distincts, les entiers $m-\ell$ et $m+\ell$ sont non nuls et de même parité : par conséquent :

$$\mathcal{C}_{m-\ell} = \mathcal{C}_{m+\ell} \tag{1.29}$$

et l'on a :

$$\boxed{(S^{(m)}, S^{(\ell)}) = 0} \tag{1.30}$$

ce qui établit bien la propriété d'orthogonalité des vecteurs colonnes. A l'inverse, si $\ell = m$, on a :

$$(S^{(m)}, S^{(m)}) = h\ (\mathcal{C}_0 - \mathcal{C}_{2m}) = \frac{1}{M+1}\ \Big(M - (-1)\Big) = 1 \tag{1.31}$$

ce qui équivaut à :

$$\boxed{\|\ S^{(m)}\ \|_2 = 1} \tag{1.32}$$

et prouve que ces vecteurs colonnes sont bien de norme euclidienne égale à 1.

En conséquence des résultats établis, la matrice S_h est bien orthogonale :

$$S_h^T \, S_h = I \tag{1.33}$$

De plus, cette matrice est symétrique,

$$\forall \, j, m \; : \; S_{j,m} = \sqrt{2h} \, \sin \, j\theta_m = \sqrt{2h} \, \sin \, m\theta_j = S_{m,j} \tag{1.34}$$

donc involutive :

$$S_h^{-1} = S_h^T = S_h \tag{1.35}$$

Il reste à prouver que les vecteurs colonnes $\{S^{(m)}\}$ ($m = 1, 2, ..., M$) (qui sont forcément linéairement indépendants) sont bien *des*, donc *les* vecteurs propres de la matrice A_h. A cette fin, on calcule la quantité :

$$\begin{aligned} h^2 \, \left(A_h \, S^{(m)}\right)_j \quad &= h^2 \sum_{k=1}^{M} \left(A_h\right)_{j,k} \, S_k^{(m)} \\ &= h^2 \sum_{k=1}^{M} \left(A_h\right)_{j,k} \, S_{k,m} \\ &= -S_{j-1,m} + 2 \, S_{j,m} - S_{j+1,m} \\ &= 2 \, S_j^{(m)} - \sqrt{2h} \, [\, \sin \, (j-1)\theta_m + \sin \, (j+1)\theta_m \,] \\ &= 2 \, S_j^{(m)} - \sqrt{2h} \, 2 \sin \, j\theta_m \, \cos \, \theta_m \\ &= \lambda_m \, S_j^{(m)} \end{aligned} \tag{1.36}$$

où λ_m a bien l'expression donnée dans l'énoncé. L'équation (1.12) est donc bien vérifiée et la démonstration est désormais complète. □

Une conséquence immédiate de cette diagonalisation est que la forme quadratique

$$q(u) = u^T \, A_h \, u \quad (u \, \in \, \mathbb{R}^M) \tag{1.37}$$

est définie positive. En effet, les valeurs propres de la matrice A_h étant réelles et strictement positives, la matrice suivante est réelle, diagonale et strictement positive :

$$\sqrt{\Lambda_h} \stackrel{\text{déf}}{=} \text{Diag} \, (\sqrt{\lambda_m}) \tag{1.38}$$

En posant,

$$W = \frac{1}{h} \sqrt{\Lambda_h} \, S_h \tag{1.39}$$

et

$$v = W\, u \tag{1.40}$$

il vient :

$$\boxed{A_h = W^T\, W} \tag{1.41}$$

et

$$q(u) = v^T\, v = \left(\| \, v \, \|_2 \right)^2 \geq 0 \quad (\forall\, u \, \in \mathbb{R}^M\,) \tag{1.42}$$

car les vecteurs u et v ont des composantes réelles. □

1.3. Différentes formes d'erreur – Critères d'arrêt

Erreur de modélisation : $\| \, u_{\text{réel}} - u \, \|$ (u : solution exacte du modèle continu).

C'est l'écart qui existe entre une représentation symbolique quantitative du système physique étudié, $u_{\text{réel}}$, et la solution exacte du modèle continu, u. Souvent, on peut estimer *a priori* ou *a posteriori* les principales composantes des phénomènes physiques négligés.

Erreur d'approximation : $\| \, u - u_h \, \|$ (u_h : solution exacte du modèle discret).

C'est l'écart qui existe entre les solutions (exactes) des modèles continu et discret. Par définition, un schéma d'approximation est d'ordre α ssi :

$$\boxed{\| \, u - u_h \, \| = O\,(h^\alpha) \quad (h \to 0)} \tag{1.43}$$

Malheureusement, cette erreur est rarement accessible car la fonction u n'est pas connue. En pratique on étudie plutôt l'*erreur de troncature* dont la définition usuelle est la suivante :

Définition 1.1 (Erreur de troncature)

On appelle « erreur de troncature » le résultat de l'application de l'opérateur d'approx-

imation discrète à la solution exacte du problème continu ; avec les notations précédentes, il vient :

$$\boxed{E_T = A_h\, u - f_h} \tag{1.44}$$

REMARQUE : on peut noter que dans le cas de *problèmes évolutifs* tels que l'équation parabolique de la chaleur, ou l'équation hyperbolique des ondes, certains auteurs [93] [4] préfèrent à l'inverse définir l'erreur de troncature comme le résultat de l'application l'opérateur continu à un interpolant fonctionnel de la solution discrète, u_h (dont il est délicat de donner une définition précise) :

$$E_T = A\, u_h - f\,. \tag{1.45}$$

Ce point de vue n'est pas adopté dans cet ouvrage.

Revenant à (1.44), on exprime habituellement l'erreur de troncature par un développement limité suivant les puissances croissantes du paramètre h, les différents termes faisant intervenir les dérivées successives de la fonction inconnue u. Dans les cas les plus simples, ce développement se réduit au reste de la formule de Taylor-Lagrange. Le développement permet alors d'identifier simplement l'ordre de l'approximation. Par exemple, dans le cas d'un problème linéaire, il vient :

$$E_T = A_h\ (u - u_h) \tag{1.46}$$

Par conséquent, pour que l'erreur d'approximation soit d'ordre h^α, il suffit qu'il en soit ainsi de l'erreur de troncature et que l'opérateur A_h^{-1} soit borné.

Exercice 1.1 (Erreur d'approximation du modèle discret fondamental)

L'exercice a pour but d'établir la forme de l'erreur d'approximation dans le cas du modèle discret unidimensionnel (1.10) pour lequel la matrice d'approximation est donnée par (1.11).

(1) Afin de résoudre explicitement le système discret, montrer d'abord que pour tout indice $\ell = 2, 3, ..., M+1$, on a :

$$u_1 + u_{\ell-1} - u_\ell = h^2\,(f_1 + f_2 + ... + f_{\ell-1}) \tag{1.47}$$

puis pour tout indice $j = 2, 3, ..., M+1$ on a :

$$u_j = j\, u_1 - h^2\,[(j-1)f_1 + (j-2)f_2 + ... + f_{j-1}] \tag{1.48}$$

D'où :

$$u_1 = \frac{h^2}{M+1}\,[M f_1 + (M-1) f_2 + ... + f_M] \tag{1.49}$$

et enfin :

$$\begin{aligned} u_j \;&= \frac{M+1-j}{M+1}\,h^2\,[f_1 + 2f_2 + ... + (j-1) f_{j-1}] \\ &+ \frac{j\,h^2}{M+1}\,[(M+1-j) f_j + (M-j) f_{j+1} + ... + f_M] \end{aligned} \tag{1.50}$$

En déduire $\| A_h^{-1} \|_\infty$.

(2) Par ailleurs, on fait l'hypothèse que $f \in C^2([0,1])$. Trouver une majoration de l'erreur de troncature faisant intervenir le nombre :

$$\mu = \max_{x \in [0,1]} |f''(x)| \tag{1.51}$$

(3) Etablir la majoration uniforme suivante de l'erreur d'approximation :

$$\|u - u_h\|_\infty \leq \frac{\mu\,h^2}{96} \tag{1.52}$$

Erreur itérative – Résidu – Itération de Jacobi - Rayon spectral - Raideur

L'erreur itérative, e_h^n, est l'écart entre le n-ième itéré u_h^n de l'algorithme itératif de résolution et la solution (exacte) du modèle discret $u_h = A_h^{-1}\,f_h$ (également notée u_h^∞ en raison de l'hypothèse faite de convergence itérative) :

$$\boxed{e_h^n = u_h^n - u_h^\infty} \tag{1.53}$$

On définit également le *résidu itératif* :

$$\boxed{r_h^n = A_h\,u_h^n - f_h} \tag{1.54}$$

On voit que si le problème est linéaire, on a la relation suivante importante entre l'erreur et le résidu :

$$\boxed{A_h\,e_h^n = r_h^n} \tag{1.55}$$

HYPOTHÈSE : faisons maintenant l'hypothèse supplémentaire que la matrice A_h est réelle-symétrique strictement définie positive de sorte que son spectre est constitué de valeurs propres réelles strictement positives,

$$\sigma\,(A_h) = \{\lambda_{h\,m}\} \quad (m = 1, 2, ..., M) \quad (0 < \lambda_{h\,1} \leq \lambda_{h\,2} \leq ... \leq \lambda_{h\,M}) \tag{1.56}$$

(*Nota Bene* : par rapport aux notations du théorème 1.1, $\lambda_{h\,m} = \lambda_m / h^2$.)

On peut alors résoudre le système discret par l'« Itération de Jacobi » définie comme suit :

$$\boxed{u_h^{n+1} = u_h^n - \tau\, r_h^n} \tag{1.57}$$

où τ est un paramètre de relaxation ($\tau \in \mathbb{R}^+$) que l'on optimise, en anticipant les résultats du chapitre 3, de manière à minimiser le rayon spectral de l'itération :

$$\boxed{\rho = \max_{m=\,1,\,2,\,...,\,M} |1 - \tau\, \lambda_{h\,m}|} \tag{1.58}$$

Exercice 1.2 (Paramètres optimaux de l'itération de Jacobi)

Montrer que la valeur optimale du paramètre τ est la suivante

$$\tau^* = \left(\frac{\lambda_{h\,M} + \lambda_{h\,1}}{2} \right)^{-1} \tag{1.59}$$

et que la valeur correspondante du rayon spectral est :

$$\rho^* = \frac{\lambda_{h\,M} - \lambda_{h\,1}}{\lambda_{h\,M} + \lambda_{h\,1}}\,. \tag{1.60}$$

L'effet de n applications de l'algorithme itératif est de réduire la norme de l'erreur d'un facteur de l'ordre de ρ^n :

$$\boxed{\frac{\| e_h^n \|}{\| e_h^0 \|} = O\,(\rho^n)} \tag{1.61}$$

(Nous ferons une analyse plus fine de cette estimation au chapitre 3.) Par conséquent, afin que n applications de l'algorithme itératif aient pour effet de réduire la norme de l'erreur d'un facteur ε, il faut que :

$$n \geq \frac{\ln 1/\varepsilon}{\ln 1/\rho} \tag{1.62}$$

En introduisant le nombre de conditionnement de la matrice A_h [1]

$$\kappa = \kappa_2 (A_h) = \frac{\lambda_{h\,M}}{\lambda_{h\,1}} \tag{1.63}$$

on aboutit à :

$$\rho = 1 - \frac{2}{\kappa} + O\left(\frac{1}{\kappa^2}\right) \tag{1.64}$$

de sorte que

$$\ln 1/\rho \approx \frac{2}{\kappa} \tag{1.65}$$

et l'inéquation précédente prend la forme suivante pour $\varepsilon = 10^{-1}$:

$$n \geq 1.15\,\kappa \tag{1.66}$$

Ce résultat bien que spécifique à l'itération de Jacobi révèle néanmoins que le paramètre κ est à 15% près, égal au nombre d'itérations nécessaires à une réduction de l'erreur itérative d'un facteur 10. On comprend donc que la difficulté à résoudre le problème itérativement, aussi appelée « raideur du système », est d'autant plus grande que le nombre de conditionnement est grand.

On peut généraliser la notion de « nombre de conditionnement » en introduisant la définition suivante :

Définition 1.2 (Nombre de conditionnement)

On appelle nombre de conditionnement de la matrice A (supposée inversible) au sens de la norme-p induite, le nombre :

$$\kappa = \kappa_p (A) = \| A \|_p \| A^{-1} \|_p \tag{1.67}$$

(On renvoie à l'annexe A pour les définitions des normes usuelles.)

Exercice 1.3 (Propriétés des nombres de conditionnement)

(1) Montrer que le nombre de conditionnement de toute matrice carrée inversible $A \in \mathcal{M}_{M\times M}$, au sens de toute norme induite $\|.\|$, est supérieur ou égal à 1 :

$$\forall\, A \in \mathcal{M}_{M\times M}\,,\ \kappa (A) \geq 1 \tag{1.68}$$

1. Voir exercice 1.3.

L'égalité peut-elle se produire ?

(2) Exprimer $\kappa_2(A)$ en fonction des valeurs propres de la matrice $A^* A$. Caractériser les matrices A pour lesquelles on a $\kappa_2(A) = 1$. Simplifier l'expression de $\kappa_2(A)$ dans le cas où la matrice A est réelle-symétrique définie positive.

(3) Calculer $\kappa_2(A_h)$ et $\kappa_\infty(A_h)$ en fonction de M dans le cas de la matrice A_h du modèle discret unidimensionnel explicitée en (1.11).

Exercice 1.4 (Encadrement de résidus)

On suppose que $e_h^0 \neq 0$. Démontrer alors l'encadrement suivant :

$$\frac{1}{\kappa} \frac{\| r_h^n \|}{\| r_h^0 \|} \leq \frac{\| e_h^n \|}{\| e_h^0 \|} \leq \kappa \frac{\| r_h^n \|}{\| r_h^0 \|} \tag{1.69}$$

Erreurs d'arrondi

En pratique, la résolution directe ou itérative (à convergence complète) sur ordinateur ne fournit pas exactement la solution u_h du modèle discret, mais une solution voisine, $u_h + \delta u_h$, où la perturbation δu_h est le résultat de l'accumulation des erreurs d'arrondi. Dans le cas linéaire, la théorie de Wilkinson [96] permet d'estimer cette erreur en supposant que les arrondis équivalent à une définition imparfaite de la matrice A_h ou du second membre f_h. Par exemple, en considérant d'abord le cas d'une perturbation δA_h de la matrice, il vient :

$$(A_h + \delta A_h)(u_h + \delta u_h) = f_h \tag{1.70}$$

dont on retranche (1.2) pour ne retenir que les termes linéaires par rapport aux perturbations :

$$A_h \delta u_h + \delta A_h u_h = 0 \tag{1.71}$$

ce qui donne :

$$\delta u_h = -A_h^{-1} \delta A_h u_h \tag{1.72}$$

dont on tire la majoration suivante :

$$\| \delta u_h \| \leq \| A_h^{-1} \| \| \delta A_h \| \| u_h \|, \tag{1.73}$$

ce qui s'écrit aussi :

$$\frac{\| \delta u_h \|}{\| u_h \|} \leq \kappa(A_h) \frac{\| \delta A_h \|}{\| A_h \|} \tag{1.74}$$

où $\kappa\,(A_h)$ est à nouveau le nombre de conditionnement de la matrice A_h (voir exercice 1.3). De manière analogue, dans le cas d'une perturbation du membre de droite, on a :

$$A_h\,\delta u_h = \delta f_h \tag{1.75}$$

d'où :

$$\|\,\delta u_h\,\| \leq \|\,A_h^{-1}\,\|\,\|\,\delta f_h\,\|\,. \tag{1.76}$$

D'autre part, la relation $f_h = A_h\,u_h$ implique l'inégalité suivante :

$$\|\,f_h\,\| \leq \|\,A_h\,\|\,\|\,u_h\,\|\,, \tag{1.77}$$

équivalente à la suivante :

$$\frac{1}{\|\,u_h\,\|} \leq \|\,A_h\,\|\,\frac{1}{\|\,f_h\,\|}\,, \tag{1.78}$$

de sorte que finalement :

$$\frac{\|\,\delta u_h\,\|}{\|\,u_h\,\|} \leq \kappa\,(A_h)\,\frac{\|\,\delta f_h\,\|}{\|\,f_h\,\|} \tag{1.79}$$

Au vu de (1.74) et (1.79), le nombre de conditionnement $\kappa\,(A_h)$ apparaît comme le facteur par lequel on doit multiplier les erreurs relatives portant sur la matrice ou le membre de droite pour obtenir une majoration de l'erreur relative sur la solution du système discret due aux erreurs d'arrondi. Ces erreurs relatives sont de l'ordre de la précision arithmétique ε :

$$\frac{\|\,\delta A_h\,\|}{\|\,A_h\,\|} \sim \varepsilon\,,\ \frac{\|\,\delta f_h\,\|}{\|\,f_h\,\|} \sim \varepsilon\,. \tag{1.80}$$

En conséquence, l'erreur d'arrondi $\|\,\delta u_h\,\|$, est (au pire) de l'ordre de :

$$\boxed{\varepsilon\,\kappa\,(A_h)\,\|\,u_h\,\|} \tag{1.81}$$

où ε est un nombre représentatif de la précision arithmétique. Par exemple si $\kappa\,(A_h) = 10^6$, si on a normalisé les équations de sorte que $\|\,u_h\,\| = 1$ et si les erreurs d'arrondi locales affectent la k-ième décimale (des mantisses) à chaque opération élémentaire, le résultat final n'est plus fiable à partir de la $k-6$-ième décimale.

Notons enfin que le nombre de conditionnement d'une matrice A_h de structure donnée, e.g. $A_h \sim \text{Trid}\,(-1,2,-1)$, est une fonction croissante de sa dimension c'est-à-dire du nombre N de degrés de liberté. Par conséquent, lorsque N augmente, la résolution devient plus difficile à cause de la raideur accrue du système : de plus, la

difficulté peut le cas échéant augmenter à cause de la nécessité de passer à une précision arithmétique supérieure.

Il existe par ailleurs d'autres formes d'erreur moins facilement quantifiables : les erreurs humaines. Ce sont principalement les erreurs de programmation, d'utilisation non optimale ou même erronée d'un logiciel, etc. Nous ne traiterons pas ce vaste sujet qui relève du thème général de la « vérification », de la « calibration » et de la « validation ».

En définitive, l'écart existant entre le résultat du calcul tel qu'il est fourni par l'ordinateur et symbolisé par la fonction $u_{\text{calculé}}$ et le phénomène réel, décrit par la fonction $u_{\text{réel}}$ est égal à la somme des erreurs précédemment introduites :

$$\begin{aligned} u_{\text{calculé}} - u_{\text{réel}} \quad &= \underbrace{\left(u_{\text{calculé}} - u_h^n\right)}_{\text{erreur d'arrondi}} + \underbrace{\left(u_h^n - u_h\right)}_{\text{erreur itérative}} \\ &+ \underbrace{\left(u_h - u\right)}_{\text{erreur d'approximation}} + \underbrace{\left(u - u_{\text{réel}}\right)}_{\text{erreur de modélisation}} \end{aligned} \tag{1.82}$$

En utilisant les estimations précédentes et l'inégalité triangulaire, on aboutit à la majoration suivante :

$$\| u_{\text{calculé}} - u_{\text{réel}} \| \leq \varepsilon \, \kappa \left(A_h\right) \, \| u_h \| + \rho^n \, \| u_h^0 \| + O\left(h^\alpha\right) + E_m \tag{1.83}$$

où E_m est l'erreur de modélisation.

En pratique, on a plus ou moins de contrôle sur les quatre termes qui apparaîssent à droite de cette inégalité. Néanmoins, il est naturel de chercher à les rendre du même ordre par le choix des paramètres numériques h et n. Théoriquement, la finesse du maillage (contrôlée par h) est choisie de telle sorte que l'erreur d'approximation soit de l'ordre de l'erreur de modélisation :

$$\text{Contrôle de la discrétisation } (h) : C_a \, h^\alpha \sim E_m \tag{1.84}$$

(C_a : constante de l'ordre de 1).

D'autre part, on doit poursuivre l'itération jusqu'à ce que l'erreur itérative ait atteint cet ordre de grandeur : un critère naturel d'arrêt de l'itération est donc le suivant :

$$\boxed{\text{Contrôle de l'itération } (n) : \rho^n \, \kappa(X_h) \, \| \, e_h^0 \, \| \leq C_a \, h^\alpha} \tag{1.85}$$

où $\kappa(X_h) = \| \, X_h \, \| \, \| \, X_h^{-1} \, \|$ est le nombre de conditionnement de la matrice de passage de la base canonique à la base des vecteurs propres : pour un système symétrique ($A_h^T = A_h$) ou antisymétrique ($A_h^T = -A_h$), en norme-2 induite $\kappa_2(X_h) = 1$ car la matrice X_h est alors orthogonale ou unitaire. Par ailleurs, il est fortement recommandé d'« adimensionner » les variables et les équations afin de mettre en évidence les échelles physiques importantes et de « normaliser » le système algébrique en conséquence. Dans ce cas, on peut supposer que $\| \, e_h^0 \, \| = O(1)$. Lorsque ces hypothèses sont légitimes, un critère naturel de l'itération prendra la forme suivante :

$$\rho^n \sim h^\alpha \tag{1.86}$$

Enfin, on s'assure *a posteriori* que l'erreur d'arrondi ne dépasse pas les autres formes d'erreur. Lorsqu'on sait estimer le rayon spectral ρ et le nombre de conditionnement $\kappa \, (A_h)$, on vérifie pour cela que le paramètre ε qui caractérise la précision arithmétique satisfait la condition suivante :

$$\boxed{\text{Contrôle de la précision arithmétique } (\varepsilon) : \varepsilon \, \kappa \, (A_h) \, \| \, u_h \, \| \leq \rho^n \, \kappa(X_h) \, \| \, e_h^0 \, \|} \tag{1.87}$$

A nouveau, grâce à l'adimensionnement et la normalisation, on peut supposer que $\| \, u_h \, \| = O(1)$ également.

Il est rarement facile en pratique d'appliquer ces règles idéalisées. Nous retiendrons néanmoins le critère d'arrêt (1.85) dans l'évaluation de la performance théorique de divers algorithmes de résolution.

Exercice 1.5 (Estimation de convergence)

On se place dans le cadre de l'exercice 1.1 et on suppose que $f(x) = 4x(1-x)$.

(1) Comment choisir h (ou M) pour que la solution numérique approche uniformément la solution exacte à 10^{-4} près ?

Désormais on fixe M à la plus petite valeur satisfaisant ce critère.

(2) Estimer le nombre de conditionnement de la matrice d'approximation A_h (au sens de la norme infinie).

(3) Estimer le nombre d'itérations de Jacobi nécessaires à une réduction relative de l'erreur itérative de 10^{-4}.

(4) Avec quelle précision arithmétique faut-il effectuer les calculs pour que les erreurs d'arrondi aient un effet négligeable sur le résultat ?

(5) Si l'on n'a pas accès à l'erreur itérative mais seulement au résidu, quelle condition portant sur le rapport $\| r_h^n \| / \| r_h^0 \|$ permet de garantir une réduction relative de l'erreur itérative de 10^{-4} ?

1.4. Objectifs et plan de l'ouvrage

Cet ouvrage s'adresse à un public ayant une connaissance au moins succincte des principaux schémas d'approximation des EDP servant de modèles classiques, généralement linéaires, des équations de la physique : équation de Laplace ou de Poisson (cas elliptique), équation de la chaleur (cas parabolique), équation d'advection ou de convection (cas hyperbolique ; on adopte ici le terme d'« advection » pour la convection linéaire). Le but du cours est de discuter des principales techniques de résolution numérique des systèmes algébriques qui résultent de la discrétisation. Il s'agit généralement de grands systèmes ayant des propriétés spectrales connues, au moins dans les cas simplifiés.

Au chapitre 2, on fait quelques rappels sur les propriétés spectrales des opérateurs discrets qui interviennent lorsqu'on approche ces équations sur un maillage régulier par des méthodes de type différences finies.

Ensuite, au chapitre 3, on étudie certaines méthodes itératives dans un cadre linéaire général indépendant de la discrétisation d'une EDP. En particulier, on met en évidence comment les méthodes de relaxation classiques (Jacobi, Gauss-Seidel, etc.) varient en performance suivant la fréquence des modes que l'on cherche à atténuer. On aboutit en particulier à la notion de lisseur qui est l'ingrédient principal des méthodes multigrilles.

Au chapitre 4, on introduit le concept d'enrichissement progressif de maillage et on montre comment, en supposant des interpolations suffisamment précises, la technique permet de gagner en efficacité.

Au chapitre 5, on définit dans le cadre d'un problème elliptique les méthodes bigrille et multigrille « idéales » dont on évalue l'efficacité. On y introduit enfin le concept de « méthode multigrille complète ».

Le chapitre 6 donne un aperçu de techniques d'approximation des équations de la mécanique des fluides compressibles en maillages non structurés et des algorithmes multigrilles construits dans un souci d'efficacité en temps calcul et éprouvés par une abondante expérimentation numérique. On s'attache à mettre en évidence les problèmes que posent la construction d'une hiérarchie de grilles non structurées et à illustrer différentes méthodes qui ont permis de les résoudre.

Au chapitre 7, on présente quelques méthodes multidomaines qui s'apparentent abstraitement aux méthodes multigrilles.

Chapitre 2

Propriétés spectrales des modèles continus et discrets

Dans ce chapitre, on rappelle brièvement les principales propriétés spectrales des opérateurs qui interviennent dans les analyses de modèles linéaires. Dans le cas d'un modèle continu, il s'agit d'opérateurs aux dérivées partielles (par rapport à la variable d'espace) définis sur un espace fonctionnel approprié ; celui-ci est généralement un sous-espace de L^2 dont les éléments sont des fonctions satisfaisant certaines conditions aux limites intervenant dans la définition du modèle et dont la structure est celle d'un espace de Hilbert de dimension infinie. Dans le cas d'un modèle discret, il s'agit des analogues discrets des précédents auxquels on associe des « matrices d'approximation ». On s'intéresse au « spectre » de ces opérateurs, c'est-à-dire à l'ensemble de ses valeurs propres ainsi qu'aux « modes propres » c'est-à-dire les fonctions ou vecteurs propres associés. On s'attache en particulier à mettre en évidence le rapport très étroit qui existe entre les « diagonalisations » d'un modèle continu et de son modèle discret associé.

2.1. Quelques rappels sur les espaces de Hilbert

Afin d'expliquer aussi brièvement que possible les raisons principales pour lesquelles les modèles linéaires usuels continus ou discrets ont des propriétés spectrales bien spécifiques, on rappelle ici certaines définitions classiques essentielles. On réfère aux ouvrages élémentaires d'algèbre linéaire (e.g. [80]) pour un rappel plus complet des définitions et théorèmes sur les espaces vectoriels de dimension finie, et aux ouvrages d'analyse fonctionnelle (e.g. [24] [86]) pour une introduction rigoureuse aux espaces de Hilbert et aux spectres des opérateurs compacts. Ici, on a choisi une introduction intuitive au sujet qui évite délibérément de discuter la notion de complétude.

On pourra aussi consulter certains ouvrages de physique théorique (e.g. [27]) qui introduisent le sujet un peu dans le même esprit.

On considère un espace vectoriel H sur le corps K des complexes ($K = \mathbb{C}$) ou des réels ($K = \mathbb{R}$).

Définition 2.1 (Forme sesquilinéaire)

Une application $H \times H \longrightarrow \mathbb{C}$ qui au couple $u, v \in H \times H$ fait correspondre un nombre complexe noté (u, v), est dite sesquilinéaire ssi :
(a) pour tout $v \in H$ fixé, l'application partielle $u \in H \longrightarrow (u, v)$ est linéaire, soit :

$$\begin{aligned} &\forall\, v \in H\,,\ \forall\, u_1, u_2 \in H\,,\ \forall\, \alpha_1, \alpha_2 \in \mathbb{C}\,, \\ &\qquad (\alpha_1 u_1 + \alpha_2 u_2, v) = \alpha_1 (u_1, v) + \alpha_2 (u_2, v) \end{aligned} \tag{2.1}$$

(b) pour tout $u \in H$ fixé, l'application partielle $v \in H \longrightarrow (u, v)$ est semi-linéaire, soit :

$$\begin{aligned} &\forall\, u \in H\,,\ \forall\, v_1, v_2 \in H\,,\ \forall\, \beta_1, \beta_2 \in \mathbb{C}\,, \\ &\qquad (u, \beta_1 v_1 + \beta_2 v_2) = \overline{\beta_1} (u, v_1) + \overline{\beta_2} (u, v_2) \end{aligned} \tag{2.2}$$

(où le sur-lignement des symboles correspond à la conjugaison).

Définition 2.2 (Forme hermitienne)

Une forme sesquilinéaire $(.,.)$ est dite hermitienne ssi :

$$\forall\, u \in H\,,\ \forall\, v \in H\,,\ (v, u) = \overline{(u, v)} \tag{2.3}$$

Une conséquence immédiate de cette propriété est que pour tout $u \in H$, le nombre complexe (u, u) est réel. Cette observation conduit à poser la définition suivante :

Définition 2.3 (Forme quadratique q associée à la forme hermitienne $(.,.)$)

On désigne ainsi l'application q de $H \longrightarrow \mathbb{R}$ définie par :

$$q(u) = (u, u) \tag{2.4}$$

Définition 2.4 (Formes positives, non dégénérées)

On dit que la forme q est positive ssi

$$\forall\, u \in H\,,\ q(u) \geq 0 \tag{2.5}$$

et non dégénérée ssi

$$q(u) = 0 \Longrightarrow u = 0 \tag{2.6}$$

EXEMPLES :

1. Cas d'un espace de dimension finie : $H = \mathbb{C}^M$
   On assimile les vecteurs de $\mathbb{C}^M$ aux vecteurs colonnes à M composantes, on note

   $$u = (u_j)\ ,\ v = (v_j) \quad (j = 1, 2, ..., M) \tag{2.7}$$

   et on pose :

   $$(u, v) = \sum_{j=1}^{M} u_j\, \overline{v_j} = u^T\, \overline{v} \tag{2.8}$$

2. Cas d'un espace de dimension infinie :

   $$H = \left\{ u \in L^2\Big([a, b] \to \mathbb{C}\Big) \ / \ u(a) = u(b) \right\} \tag{2.9}$$

   On pose alors :

   $$G(u, v) = \int_a^b u(x)\, \overline{v(x)}\, dx \tag{2.10}$$

Exercice 2.1 (Formes hermitiennes)

Vérifier que les axiomes de définition d'une forme hermitienne sont bien satisfaits dans ces deux cas. Vérifier de plus que les formes quadratiques associées sont positives non dégénérées.

Théorème 2.1 (Cauchy-Schwarz)

Soit q la forme quadratique hermitienne associée à une forme hermitienne non dégénérée positive ; on a :

1. *« Inégalité de Cauchy-Schwarz » :*

   $$\forall\, u, v \in H\ ,\ |(u, v)|^2 \leq (u, u)\,(v, v) \tag{2.11}$$

2. *On désigne ainsi l'application de H dans $\mathbb{R}+$ définie par :*

   $$u \longrightarrow \sqrt{q(u)} \tag{2.12}$$

   est une norme sur H.

DÉMONSTRATION : les vecteurs u et v sont quelconques mais fixés. Pour tout $\lambda \in \mathbb{C}$, on a :

$$q(u+\lambda v) = (u+\lambda v, u+\lambda v) = \alpha + \overline{\lambda}\,\beta + \lambda\,\overline{\beta} + \lambda\,\overline{\lambda}\,\gamma \geq 0 \tag{2.13}$$

où l'on a posé :

$$\begin{array}{c} \alpha = q(u) = (u,u) \geq 0 \\ \beta = (u,v) \quad \text{(complexe non réel en général)} \\ \gamma = q(v) = (v,v) \geq 0 \end{array} \tag{2.14}$$

car la forme quadratique q est par hypothèse définie-(semi-)positive. Dans le cas très particulier où $\gamma = 0$, l'inéquation (2.13) exige que l'on ait aussi $\beta = 0$, auquel cas l'inégalité de Cauchy-Schwarz énonce un résultat trivial. Dans le cas inverse, $\gamma > 0$; on a alors :

$$\forall\, \lambda \in \mathbb{C},\ \alpha\,\gamma + \overline{\lambda}\,\beta\,\gamma + \lambda\,\overline{\beta}\,\gamma + \lambda\,\overline{\lambda}\,\gamma^2 \geq 0 \tag{2.15}$$

soit encore

$$\forall\, \lambda \in \mathbb{C},\ (\lambda\,\gamma + \beta)\left(\overline{\lambda}\,\gamma + \overline{\beta}\right) + \alpha\,\gamma - \beta\,\overline{\beta} = |\lambda\,\gamma + \beta|^2 + \alpha\,\gamma - |\beta|^2 \geq 0 \tag{2.16}$$

Le résultat recherché s'obtient enfin en considérant le cas particulier $\lambda = -\beta/\gamma$. □

Définition 2.5 (Espace de Hilbert)

Lorsqu'on est dans le cadre des hypothèses du théorème précédent si on suppose de plus que l'espace H est *complet*, ce qui est acquis en dimension finie mais pas nécessairement vrai sinon, la structure ainsi construite est celle d'un « espace de Hilbert ».

Désormais, on supposera qu'on se place toujours dans le cadre d'un espace de Hilbert.

Définition 2.6 (Opérateur adjoint)

On dit que l'opérateur A défini sur H admet un opérateur adjoint A^* ssi :

$$\forall\, u, v \in H\,,\ (Au, v) = (u, A^* v) \tag{2.17}$$

Dans le cas d'un espace de dimension finie précédent, on a les identités suivantes :

$$(Au,v) = (Au)^T\ \overline{v} = u^T\,A^T\,\overline{v}\,,\ (u, A^*v) = u^T\,\overline{(A^*\,v)} = u^T\,\overline{A^*}\,\overline{v} \tag{2.18}$$

de sorte que A^* est l'adjoint de A ssi $\overline{A^*} = A^T$ ce qui équivaut à :

$$\boxed{A^* = \overline{A}^T} \tag{2.19}$$

Définition 2.7 (Opérateur auto-adjoint)

Tout opérateur A égal à son adjoint

$$\boxed{A = A^*} \tag{2.20}$$

Dans ce cas :

$$\forall\, u \in H\,,\ \forall\, v \in H\,,\ (Au, v) = (u, Av) \tag{2.21}$$

Théorème 2.2 (Spectre d'un opérateur auto-adjoint)

Toute valeur propre d'un opérateur auto-adjoint est réelle.

DÉMONSTRATION : soit λ une telle valeur propre et u_0 un vecteur propre associé non nul (ce qui implique que $q(u_0) = (u_0, u_0) \neq 0$ dans le cas d'une forme quadratique non dégénérée). Alors, en remplaçant u et v par u_0 dans l'équation précédente on obtient

$$\lambda = \overline{\lambda} \tag{2.22}$$

(après simplification par $q(u_0)$), ce qui implique le résultat :

$$\boxed{\lambda \in \mathbb{R}} \tag{2.23}$$

□

Définition 2.8 (Opérateur antisymétrique)

Tout opérateur A égal à l'opposé de son adjoint

$$\boxed{A = -A^*} \tag{2.24}$$

Dans ce cas :

$$\forall\, u \in H\,,\ \forall\, v \in H\,,\ (Au, v) = -(u, Av) \tag{2.25}$$

Théorème 2.3 (Spectre d'un opérateur antisymétrique)

Toute valeur propre d'un opérateur antisymétrique est purement imaginaire.

DÉMONSTRATION : soit λ une telle valeur propre et u_0 un vecteur propre associé non nul (ce qui implique que $q(u_0) = (u_0, u_0) \neq 0$ dans le cas d'une forme quadratique non dégénérée). Alors, en remplaçant u et v par u_0 dans l'équation précédente on obtient

$$\lambda = -\overline{\lambda} \tag{2.26}$$

(après simplification par $q(u_0)$), ce qui implique le résultat :

$$\boxed{\lambda = \mu\, i\,,\ \mu \in \mathbb{R}} \tag{2.27}$$

□

Définition 2.9 (Opérateur normal)

Un opérateur qui commute avec son adjoint :

$$A\, A^* = A^*\, A \tag{2.28}$$

C'est le cas en particulier d'opérateurs auto-adjoints ou antisymétriques. Cette propriété est la clé principale d'un théorème portant sur la « décomposition spectrale » de l'espace de Hilbert H (voir [34], Théorème 4) que nous omettons ici. Dans les applications qui nous concernent, en particulier celles faisant intervenir l'opérateur laplacien et son inverse, nous admettrons que l'existence d'un opérateur normal (compact) permet l'identification d'une suite dénombrable de fonctions propres $\{\phi_m(x)\}$ ($m = 1, 2, ...$) deux à deux orthogonales formant une base (topologique) de H. Autrement dit, toute fonction u de H peut alors être décomposée en une série de ces fonctions propres :

$$\forall\, u \in H\,,\ u(x) = \sum_m c_m\, \phi_m(x) \tag{2.29}$$

$$A\, \phi_m(x) = \lambda_m\, \phi_m(x) \tag{2.30}$$

$$\lambda_m \neq \lambda_k \implies (\phi_m, \phi_k) = 0 \tag{2.31}$$

ce qui ramène l'étude linéaire à une analyse modale. (L'équation (2.29) qui est une égalité au sens de L^2 implique une convergence « en moyenne ».)

2.1.1. Application au cas de la dérivée première

On se place à nouveau dans le cas de l'espace de Hilbert

$$H = \Big\{\, u \in L^2\Big([a,b] \to \mathbb{C}\Big) \ /\ u(a) = u(b) \Big\} \tag{2.32}$$

muni de la forme hermitienne :

$$(u, v) = \int_a^b u(x)\,\overline{v(x)}\,dx \tag{2.33}$$

Exercice 2.2 (Opérateur de dérivée première)

Montrer que l'opérateur de dérivée première A

$$u \in H \quad \xrightarrow{A} \quad v = A\,u \overset{\text{déf}}{=} \frac{du}{dx} \tag{2.34}$$

est antisymétrique, que ses fonctions propres sont (et sont exclusivement) les « modes de Fourier » suivants,

$$\phi_m(x) = \exp\left(2\pi\,i\,m\,\frac{x-a}{b-a}\right) \quad (m \in \mathbb{Z}) \tag{2.35}$$

associés aux *valeurs propres purement imaginaires* suivantes :

$$\lambda_m = \frac{2\pi\,i\,m}{b-a} \tag{2.36}$$

Ces fonctions propres constituent une « *base topologique* » des fonctions $u(x)$ de période $b-a$ développables en *séries de Fourier* :

$$u(x) = c_0 + \sum_{m=1}^{\infty} \Big(c_m\,\phi_m(x) + c_{-m}\,\phi_{-m}(x)\Big) \quad (c_m \in \mathbb{C},\ \forall\, m) \tag{2.37}$$

Une telle fonction est à valeurs dans $\mathbb{R}$ ssi

$$\forall\, m \in \mathbb{Z}\,,\ c_{-m} = \overline{c_m} \tag{2.38}$$

dans ce cas en posant,

$$c_m = \frac{a_m - i\,b_m}{2} \quad (a_m\,,\ b_m \in \mathbb{R},\ \forall\, m) \tag{2.39}$$

on obtient :

$$u(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} \left(a_m\,\cos\left(2\pi\,m\,\frac{x-a}{b-a}\right) + b_m\,\sin\left(2\pi\,m\,\frac{x-a}{b-a}\right)\right) \tag{2.40}$$

Exercice 2.3 (Equation d'advection pure)

Utiliser la décomposition en série de Fourier précédente pour résoudre formellement le problème suivant :

$$\left\{ \begin{array}{l} u_t + c\, u_x = 0 \quad (x \in \mathbb{R};\ t \in \mathbb{R}+) \\ u(x,0) = u_0(x) \quad (\forall\, x) \end{array} \right. \tag{2.41}$$

dans lequel la condition initiale $u_0(x)$ est une fonction donnée de période L ($a = 0\,,\ b = L$).

2.1.2. Application au cas de la dérivée seconde

On se place ici dans le cas de l'espace de Hilbert suivant

$$H_0^1\left([a,b]\right) = \left\{ u \in L^2\Big([a,b] \to \mathbb{R}\Big)\ ,\ u' \in L^2\Big([a,b] \to \mathbb{R}\Big)\ /\ u(a) = u(b) = 0 \right\} \tag{2.42}$$

muni du produit scalaire

$$(u,v) = \int_a^b u(x)\, v(x)\, dx \tag{2.43}$$

Exercice 2.4 (Opérateur dérivée seconde)

Montrer que l'opérateur de dérivée seconde A

$$u \in H_0^1 \quad \stackrel{A}{\longrightarrow} \quad v = A\, u \stackrel{\text{déf}}{=} \frac{d^2 u}{dx^2} \tag{2.44}$$

est auto-adjoint, défini-négatif, que ses fonctions propres sont (et sont exclusivement) les fonctions sinusoïdales suivantes,

$$\psi_m(x) = \sin\left(m\,\pi\,\frac{x-a}{b-a} \right) \quad (m = 1, 2, \ldots) \tag{2.45}$$

associées aux *valeurs propres réelles négatives* suivantes :

$$\mu_m = -\left(\frac{m\,\pi}{b-a} \right)^2 \tag{2.46}$$

Ces fonctions propres constituent une « *base topologique* » de fonctions $u(x)$ nulles aux limites dont les extensions par périodicité sont de période $2(b-a)$.

Exercice 2.5 (Equation de la chaleur)

Utiliser une décomposition en série de fonctions sinusoïdales pour résoudre formellement le problème suivant :

$$\left\{ \begin{array}{l} u_t = \sigma\, u_{xx} \quad (x \in [0, L]\,;\; t \in \mathbb{R}+) \\ u(0,t) = u(L,t) = 0 \quad (\forall\, t > 0) \\ u(x,0) = u_0(x) \quad (\forall\, x) \end{array} \right. \tag{2.47}$$

dans lequel la condition initiale $u_0(x)$ est une fonction donnée sur $[0, 1]$.

2.2. Opérateurs discrets

2.2.1. Les opérateurs aux différences en périodique

On s'intéresse ici à l'analyse spectrale d'opérateurs linéaires en dimension finie lorsque des conditions de périodicité sont appliquées à tout vecteur u_h de $\mathbb{R}^M$,

$$u_h = \left(\begin{array}{c} u_1 \\ u_2 \\ \vdots \\ u_M \end{array} \right) \tag{2.48}$$

(M : nombre de degrés de liberté) que l'on interprète comme le discrétisé sur un maillage uniforme dont les nœuds ont pour abscisses

$$x_j = jh \quad (h = \frac{L}{M}) \tag{2.49}$$

d'une certaine fonction périodique $u_h(x)$:

$$\begin{array}{c} j = 1, 2, ..., M\,,\; u_j = u_h(jh) \\ \forall\, x \in \mathbb{R},\; u_h(x + L) = u_h(x) \end{array} \tag{2.50}$$

(L : période spatiale du problème continu). On développe l'outil classique d'« *Analyse de Fourier* » par le biais de l'algèbre des *matrices circulantes*.

Dans le contexte périodique, il est commode d'adopter la convention de notation consistant à étendre la définition de u_j à tout entier relatif j par périodicité ; en particulier près des bords, on posera :

$$\begin{array}{l} u_{M+1} = u_1\,,\; u_{M+2} = u_2\,,\; ... \\ u_0 = u_M\,,\; u_{-1} = u_{M-1}\,,\; ... \end{array} \tag{2.51}$$

Plus précisément, on s'intéresse au cas d'un endomorphisme A_h de $\mathbb{R}^M$,

$$\boxed{u_h \in \mathbb{R}^M \quad \xrightarrow{A_h} \quad v_h = A_h\, u_h = \{v_j\} \in \mathbb{R}^M} \tag{2.52}$$

« spatialement uniforme » c'est-à-dire tel qu'il existe des coefficients $\{\alpha_k\}$ pour lesquels

$$\boxed{\forall\, j\,,\ v_j \stackrel{\text{déf}}{=} \sum_{k=-K}^{K} \alpha_k\, u_{j+k}} \tag{2.53}$$

K étant le plus petit entier pour lequel cette équation est valable. Dans cette formule, les coefficients $\{\alpha_k\}$ ne dépendent pas de l'indice j, mais seulement de l'indice k. D'autre part, lorsque l'indice $j+k$ n'appartient pas à l'ensemble $\{1, 2, ..., M\}$ la convention (2.51) est utilisée. Interprétant le vecteur v_h comme le discrétisé sur le maillage d'une certaine fonction $v_h(x)$, on voit que cette condition équivaut à imposer que la dépendance fonctionnelle de v_h sur u_h est linéaire, locale, uniforme et périodique. Notons que si l'opérateur A_h est un opérateur de différence finie, les coefficients α_k ne sont pas indépendants puisque la condition de consistance impose notamment que l'on ait :

$$\sum_{k=-K}^{K} \alpha_k = 0 \tag{2.54}$$

Cette hypothèse sur A_h n'est pas faite ici, bien que nous appliquerons les résultats généraux à des cas particuliers où elle est satisfaite.

On introduit en particulier les opérateurs de différence première décentrée « amont », ∇_h, et « aval », Δ_h du premier ordre en posant pour tout j :

$$\boxed{\begin{array}{l} \nabla_h u_j = u_j - u_{j-1} \\ \Delta_h u_j = u_{j+1} - u_j \end{array}} \tag{2.55}$$

Ces opérateurs sont représentés par les matrices pseudo-bidiagonales suivantes :

$$\nabla_h = \begin{pmatrix} 1 & & & -1 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}, \quad \Delta_h = \begin{pmatrix} -1 & 1 & & \\ & -1 & \ddots & \\ & & \ddots & 1 \\ 1 & & & -1 \end{pmatrix} \tag{2.56}$$

où les éléments qui rompent la structure bidiagonale stricte sont la marque de la condition de périodicité. D'une manière générale, on utilise toujours la même notation pour un opérateur linéaire quelconque et la matrice qui le représente.

Exercice 2.6 (Opérateurs de différences finies périodiques usuels)

Identifier les coefficients $\{\alpha_k\}$ correspondant aux opérateurs suivants :
– ∇_h et ∇_h^2 : opérateurs de différence première et seconde décentrées « amont » du premier ordre,
– Δ_h et Δ_h^2 : opérateurs de différence première et seconde décentrées « aval » du premier ordre,
– δ_h : opérateur de différence première centrée (donc du second ordre), défini par

$$\delta_h u_j = \frac{u_{j+1} - u_{j-1}}{2} \tag{2.57}$$

– δ_h^- : opérateur de différence première décentrée « amont » du second ordre, défini par

$$\delta_h^- u_j = \frac{3u_j - 4u_{j-1} + u_{j-2}}{2} \tag{2.58}$$

– δ_h^+ : opérateur de différence première « aval » du second ordre, défini par

$$\delta_h^+ u_j = \frac{-3u_j + 4u_{j+1} - u_{j+2}}{2} \tag{2.59}$$

Définition 2.10 (Matrice circulante)

On dit qu'une matrice carrée est circulante lorsque chacune de ses lignes s'obtient de la précédente par permutation circulaire.

Exercice 2.7 (Structure matricielle d'opérateur périodique)

(1) Identifier la structure de la matrice représentant l'opérateur A_h de (2.52)-(2.53) en fonction des coefficients $\{\alpha_k\}$ dans le cas où $K = 2$ et $M = 7$.

(2) Est-elle circulante ?

On considère maintenant deux opérateurs particuliers, D_h^- et D_h^+ : opérateurs de « translation » ou d'« incrémentation d'indice » définis comme suit :

$$D_h^- \, u_j = u_{j-1} \, , \; D_h^+ \, u_j = u_{j+1} \tag{2.60}$$

et représentés en périodique par les matrices pseudo-bidiagonales suivantes :

$$D_h^- = \begin{pmatrix} 0 & & & 1 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} , \quad D_h^+ = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{pmatrix} \tag{2.61}$$

Il résulte de ces définitions que

$$\nabla_h = I - D_h^- \ , \ \Delta_h = D_h^+ - I \tag{2.62}$$

Exercice 2.8 (Diagonalisation des matrices circulantes)

(1) Montrer que les opérateurs D_h^- et D_h^+ sont
– inverses l'un de l'autre,
– adjoints (ici transposés) l'un de l'autre,
– normaux (donc diagonalisables),
et qu'ils commutent.

(2) En déduire qu'il existe une transformation unitaire Φ_h,

$$\boxed{\Phi_h^* \, \Phi_h = \Phi_h \, \Phi_h^* = I} \tag{2.63}$$

qui diagonalise simultanément ces opérateurs :

$$D_h^- = \Phi_h \, \mathcal{D}_h^- \, \Phi_h^* , \ D_h^+ = \Phi_h \, \mathcal{D}_h^+ \, \Phi_h^* \tag{2.64}$$

On vérifiera spécifiquement que :

$$\Phi_h = \left(\Phi_{h\,j,m} \right) \ , \ \mathcal{D}_h^- = \mathrm{Diag} \left(d_m^- \right) \ , \ \mathcal{D}_h^+ = \mathrm{Diag} \left(d_m^+ \right) \tag{2.65}$$

où :

$$\boxed{\Phi_{h\,j,m} = \frac{1}{\sqrt{M}} \, \exp \left(i \, j \, \theta_m \right)} \tag{2.66}$$

$$d_m^- = \exp \left(- \, i \, \theta_m \right) \ , \ d_m^+ = \exp \left(+ \, i \, \theta_m \right) \tag{2.67}$$

$$\boxed{\theta_m = \frac{2 \pi \, m}{M}} \tag{2.68}$$

La quantité $\theta_m \in [0, 2\pi]$ est appelée « paramètre de fréquence ».

(3) Montrer que les vecteurs propres, c'est-à-dire les vecteurs colonnes de la matrice Φ_h sont (à une normalisation près) les discrétisés des fonctions propres $\phi_m(x)$ de (2.35).

En vertu des résultats de l'exercice 2.8, on simplifie la notation en posant :

$$\boxed{D_h^+ = D_h\ ,\ D_h^- = D_h^{-1}\ ,\ \mathcal{D}_h^+ = \mathcal{D}_h\ ,\ \mathcal{D}_h^- = \mathcal{D}_h^{-1}} \tag{2.69}$$

Revenant à l'opérateur général de (2.52)-(2.53), on voit qu'il s'exprime symboliquement au moyen du seul opérateur D_h et de la fonction rationnelle

$$\boxed{R(X) = \sum_{k=-K}^{K} \alpha_k\, X^k} \tag{2.70}$$

à savoir :

$$\boxed{A_h = R\,(D_h)} \tag{2.71}$$

Par conséquent, la diagonalisation de l'opérateur D_h permet aussi de diagonaliser l'opérateur général A_h indépendamment des coefficients $\{\alpha_k\}$:

$$\boxed{A_h = \Phi_h\, \Lambda_h\, \Phi_h^*} \tag{2.72}$$

où

$$\boxed{\Lambda_h = R\,(\mathcal{D}_h) = \mathrm{Diag}\,(\lambda_{\mathrm{h\,m}}) = \begin{pmatrix} \lambda_{h\,1} & & & \\ & \lambda_{h\,2} & & \\ & & \ddots & \\ & & & \lambda_{h\,M} \end{pmatrix}} \tag{2.73}$$

et les valeurs propres de A_h sont les suivantes :

$$\boxed{\lambda_{h\,m} = R\Big(\exp\,(i\,\theta_m)\Big)} \tag{2.74}$$

Plus explicitement,

$$\boxed{\begin{array}{rl} \lambda_{h\,m} = & \alpha_0 + \alpha_1\, e^{i\,\theta_m} + \alpha_{-1}\, e^{-\,i\,\theta_m} + \alpha_2\, e^{2\,i\,\theta_m} + \alpha_{-2}\, e^{-\,2\,i\,\theta_m} + \ldots \\ & + \alpha_K\, e^{K\,i\,\theta_m} + \alpha_{-K}\, e^{-\,K\,i\,\theta_m} \end{array}} \tag{2.75}$$

Exercice 2.9 (Spectres d'opérateurs périodiques particuliers)

Identifier les spectres des opérateurs de l'exercice 2.6.

Définition 2.11 (Transformé de Fourier discret, composantes fréquentielles)

On appelle transformé de Fourier discret du vecteur u_h le vecteur

$$\boxed{\widehat{u_h} = \Phi_h^* \, u_h} \tag{2.76}$$

Ses composantes sont appelées « composantes fréquentielles » (ou « modales »). Réciproquement :

$$\boxed{u_h = \Phi_h \, \widehat{u_h}} \tag{2.77}$$

Les « relations de réciprocité » entre composantes nodales (de u_h) et modales (de $\widehat{u_h}$) sont les suivantes :

$$\boxed{\begin{aligned} (\widehat{u_h})_m &= \sum_{j=1}^{M} \overline{\Phi_h}_{j,m} \, (u_h)_j = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} \exp\,(-\,i\,j\,\theta_m)\,(u_h)_j \\ (u_h)_j &= \sum_{m=1}^{M} \Phi_{h\,j,m} \, (\widehat{u_h})_m = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \exp\,(+\,i\,j\,\theta_m)\,(\widehat{u_h})_m \end{aligned}} \tag{2.78}$$

Exercice 2.10 (Conservation de la norme par la transformée de Fourier)

Montrer que les vecteurs u_h et $\widehat{u_h}$ ont la même norme euclidienne.

Exercice 2.11 (Transformées de Fourier discrète et continue)

De quelles relations les équations de (2.78) sont-elles les analogues discrètes ?

Ainsi les composantes fréquentielles sont les composantes du vecteur u_h lorsqu'on exprime ce vecteur dans la base des vecteurs colonnes de la matrice Φ_h, c'est-à-dire dans la base des modes de Fourier discrets.

En conclusion, nous venons d'établir que tous les opérateurs linéaires, locaux, uniformes et périodiques sont simultanément diagonalisés par la transformation de Fourier discrète représentée par la matrice unitaire Φ_h. Leurs spectres de valeurs propres s'obtiennent immédiatement par les équations (2.74)-(2.75). De plus, toute expression algébrique ne faisant intervenir que de tels opérateurs se diagonalise aussi simultanément. Par conséquent, toute l'algèbre de ces opérateurs (addition, multiplication par un nombre, composition) se transpose directement par une algèbre équivalente portant sur leurs valeurs propres respectives, chaque opérateur étant remplacé par sa valeur propre générique ou « symbole de Fourier ».

2.2.2. L'opérateur de différence première centrée

$$\boxed{\delta_h \ : \ \delta_h \, u_j = \frac{u_{j+1} - u_{j-1}}{2}} \tag{2.79}$$

Cas périodique

$$\boxed{\lambda_{h\,m} = \frac{e^{i\,\theta_m} - e^{-\,i\,\theta_m}}{2} = i \ \sin \ \theta_m} \tag{2.80}$$

Comme son analogue continu, l'opérateur de dérivée première, cet opérateur discret est antisymétrique et son spectre est purement imaginaire. Ce spectre, constitué de valeurs propres doubles lorsque M est pair, est représenté à la figure 2.1 dans le cas où $M = 32$.

Exercice 2.12 (Spectre de la différence centrée)

Etablir le lien entre ce spectre et le spectre de l'opérateur de dérivée première.

Cas non périodique

Pour analyser un modèle d'opérateur de différence centrée non périodique, on considère la variante suivante du problème d'advection pure, (2.41), dans laquelle le domaine spatial est borné et les fonctions non périodiques spatialement :

$$\left\{ \begin{array}{l} u_t + c\,u_x = 0 \quad (c > 0\,;\ x \in [a,b]\,;\ t \in \mathbb{R}+) \\ u(x,0) = u_0(x) \quad (\forall\, x) \\ u(a,t) = \alpha \quad (\forall\, t) \end{array} \right. \tag{2.81}$$

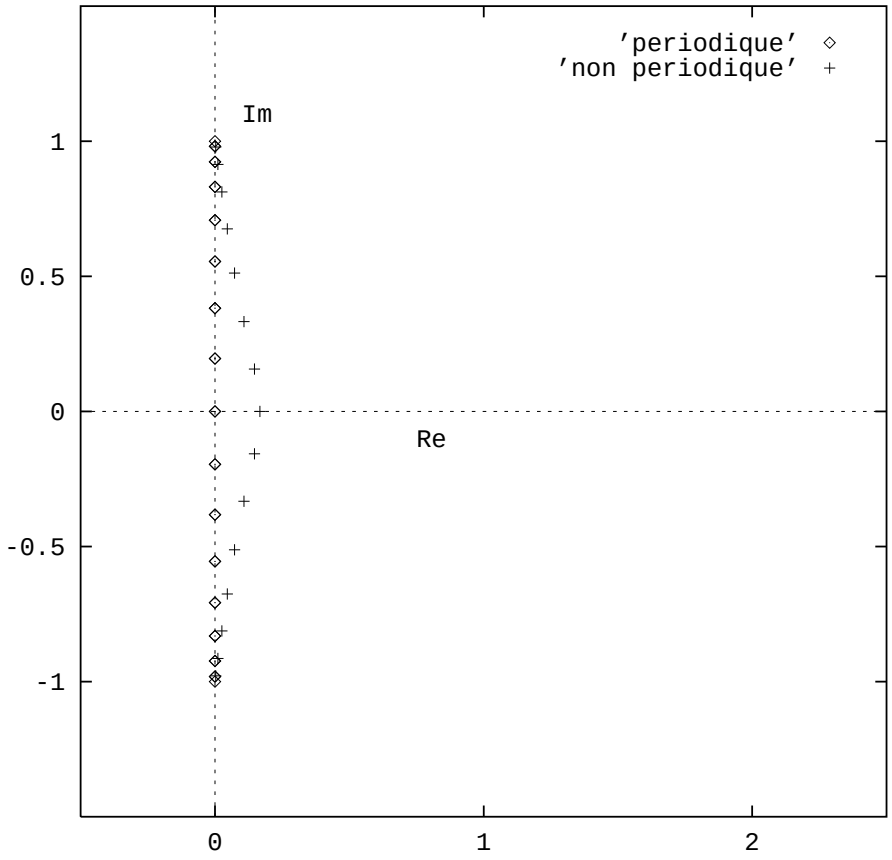


Figure 2.1. Spectre de l'opérateur de différence centrée δ_h dans les cas périodique ($M = 32$) et non périodique ($M = 15$)

Dans ce problème, on a supposé $c > 0$ pour fixer les idées. En conséquence, la fonction $u(x,t) = u_0(x-ct)$ représente une onde se déplaçant dans le sens des x croissants. Ceci justifie d'une part qu'on impose une condition de Dirichlet à gauche ($x = a$) et aucune condition à droite ($x = b$) qui constitue un bord libre calculé par intégration de l'EDP. Pour imposer la condition de Dirichlet, on introduit un point fictif en $x = a$ auquel on affecte l'indice 0, de sorte qu'au premier degré de liberté, $x = x_1$, la quantité $\delta_h\, u_1$ se réduit à :

$$\begin{aligned} \delta_h\, u_1 &= \frac{u_2 - u_0}{2} \\ &= \underbrace{\tfrac{1}{2}\, u_2}_{\text{contribution à } A_h\, u_h} - \underbrace{\tfrac{1}{2}\, \alpha}_{\text{contribution à } f_h} \end{aligned} \tag{2.82}$$

Par conséquent, la valeur limite α n'affecte pas la matrice A_h associée à δ_h mais seulement le vecteur f_h de conditions aux limites. A droite, $x = b = x_M$, on ne peut calculer de différence centrée puisqu'on ne dispose d'aucune condition légitime pour extrapoler une valeur de u en un point fictif. On choisit alors de dégrader l'ordre de l'approximation, en remplaçant δ_h en ce point par une différence décentrée amont du premier ordre seulement, sachant que cette perturbation n'est perçue que localement en raison du sens de propagation de l'onde. Ceci donne :

$$\delta_h\, u_M = \nabla_h u_M = u_M - u_{M-1} \tag{2.83}$$

REMARQUE : la procédure au bord droit équivaut à extrapoler linéairement une valeur au point fictif $x_{M+1} = x_M + h$:

$$u_{M+1} = 2\, u_M - u_{M-1} \tag{2.84}$$

avant d'appliquer l'opérateur de différence centrée au point limite :

$$\delta_h\, u_M = \frac{u_{M+1} - u_{M-1}}{2} = u_M - u_{M-1} \tag{2.85}$$

Par conséquent, cette procédure équivaut à imposer discrètement que la dérivée seconde est nulle au bord.

Compte tenu de ces choix, le schéma discret dans le cas non périodique est complètement défini par le couple matrice-vecteur suivant :

$$A_h = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & -1 & 1 \end{pmatrix}, \; f_h = \begin{pmatrix} \frac{1}{2}\,\alpha \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \tag{2.86}$$

pour lequel $\delta_h\, u_h = A_h\, u_h - f_h$. Le spectre de la matrice A_h est représenté à la figure 2.1 dans le cas où $M = 15$. On y constate la très forte parenté entre les spectres

de l'opérateur de différence centrée dans les cas périodique et non périodique. La modification de procédure en $x = b$, a eu pour effet d'introduire le nombre 1 dans la diagonale de la matrice et donc d'en augmenter la trace, qui est aussi la somme de ses valeurs propres :

$$\text{Trace } (A_h) = \sum_m \lambda_{h\,m} = 1 \tag{2.87}$$

Les valeurs propres en sont visiblement seulement légèrement modifiées, les petites perturbations portant principalement sur les parties réelles, celles-ci devenant strictement positives dans le cas non périodique (voir figure 2.1). On peut d'ailleurs montrer que ces perturbations sont de l'ordre de $\ln\, M/M$.

2.2.3. L'opérateur de différence première décentrée « amont » du premier ordre

$$\boxed{\nabla_h \ :\ \nabla_h\, u_j = u_j - u_{j-1}} \tag{2.88}$$

Cas périodique

$$\boxed{\lambda_{h\,m} = 1 - e^{-\,i\,\theta_m} = 1 - \cos\,\theta_m + i\,\sin\,\theta_m} \tag{2.89}$$

Lorsque m varie, $\lambda_{h\,m}$ décrit le cercle du plan complexe de centre 1 et de rayon 1 représenté à la figure 2.2. Ce cercle est tangent à l'axe des imaginaires qui porte le spectre du modèle continu associé. Les valeurs propres les plus proches de cet axe sont celles qui sont associées à des modes de basses fréquences (θ_m petit).

Cas non périodique

Dans le cas non périodique, on considère à nouveau un modèle dans lequel on applique une condition de Dirichlet à gauche ($x = a$). Aucune condition n'est nécessaire à droite. On aboutit ici au couple matrice-vecteur suivant :

$$A_h = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix},\ f_h = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.90}$$

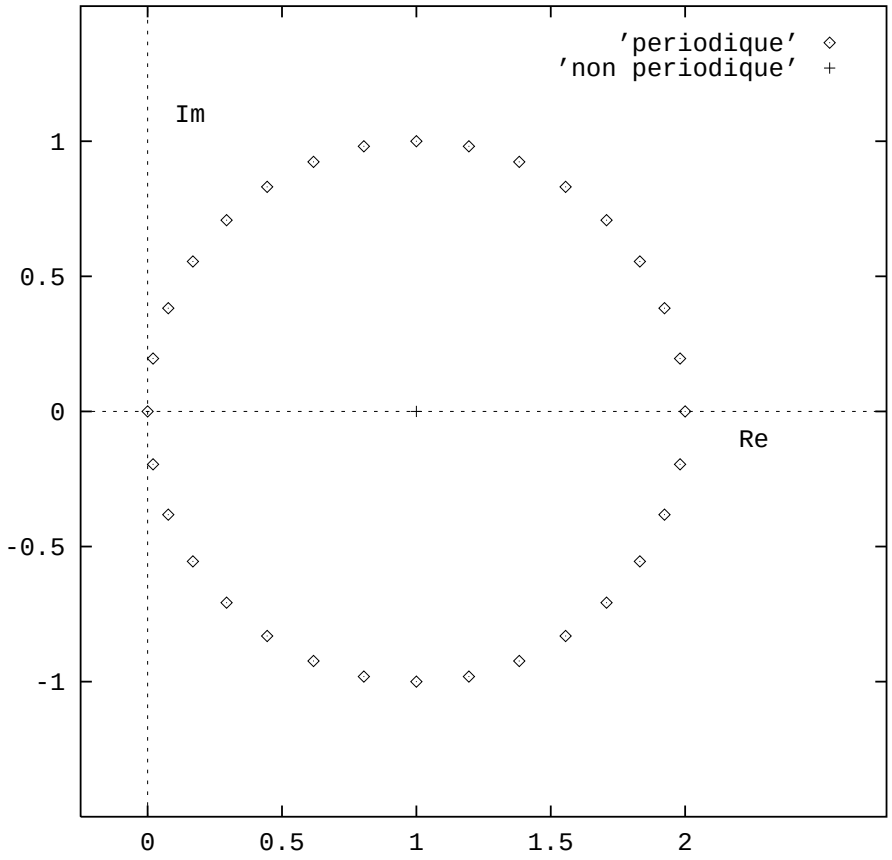


Figure 2.2. Spectre de l'opérateur de différence décentrée amont du premier ordre ∇_h dans les cas périodique ($M = 32$) et non périodique

pour lequel $\nabla_h\, u_h \;=\; A_h\, u_h \,-\, f_h$. La matrice A_h est *triangulaire inférieure*. Ses valeurs propres qui apparaîssent dans la diagonale sont toutes égales :

$$\forall\, m\,,\ \lambda_{h\,m} = 1 \tag{2.91}$$

(voir figure 2.2). Par conséquent, *la matrice n'est pas diagonalisable*. On dit aussi qu'elle est *défective* car des vecteurs propres font défaut ; on ne peut en construire une base. Dans ce cas précis, il n'existe qu'une seule direction propre. Cette anomalie peut avoir pour conséquence les pathologies de convergence de certains algorithmes [36] [38].

2.2.4. L'opérateur de différence première décentrée « amont » du second ordre

$$\boxed{\delta_h^-\ :\ \delta_h^-\, u_j = \frac{3\, u_j - 4\, u_{j-1} + u_{j-2}}{2}} \tag{2.92}$$

Cas périodique

$$\boxed{\lambda_{h\,m} = \frac{3 - 4\, e^{-\,i\,\theta_m} + e^{-\,2\,i\,\theta_m}}{2}} \tag{2.93}$$

Le spectre correspondant est représenté à la figure 2.3 dans le cas où $M \;=\; 32$. Lorsque m varie, $\lambda_{h\,m}$ décrit une courbe fermée tangente à l'axe des imaginaires qui, on le rappelle, porte le spectre du modèle continu associé. L'augmentation du degré de précision a eu pour effet de rendre ce contact sur-osculateur.

Cas non périodique

Pour cet opérateur, on a besoin de 2 conditions à gauche ($x = a$) en amont du nœud 1 où se place le premier degré de liberté. On peut à nouveau imposer la condition de Dirichlet suivante :

$$u_0 = u(a) = \alpha \tag{2.94}$$

ce qui permet la simplification suivante :

$$\begin{aligned} \delta_h^-\, u_2 \;&= \tfrac{1}{2}\,(3\, u_2 - 4\, u_1 + u_0) \\ &= \underbrace{\tfrac{1}{2}\,(3\, u_2 - 4\, u_1)}_{\text{contribution à } A_h\, u_h} + \underbrace{\tfrac{1}{2}\alpha}_{\text{contribution à } f_h} \end{aligned} \tag{2.95}$$

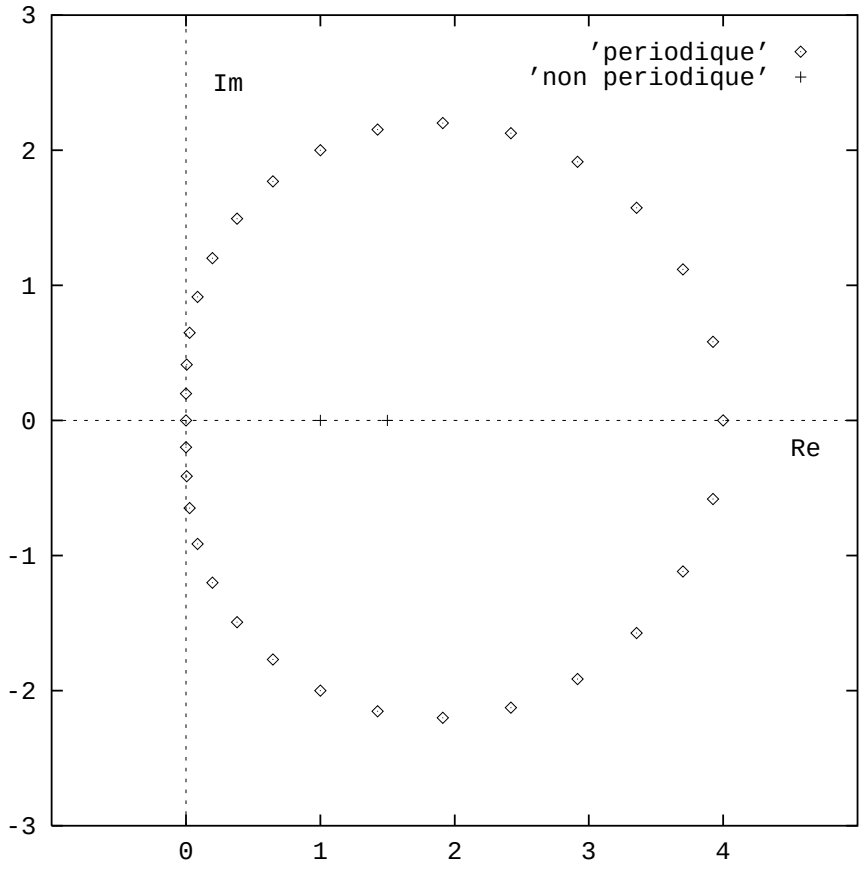


Figure 2.3. Spectre de l'opérateur de différence décentrée amont du second ordre δ_h^- dans les cas périodique ($M = 32$) et non périodique

Cependant, une deuxième condition doit être imposée. On choisit de dégrader l'opérateur au premier ordre au premier point :

$$\begin{aligned} \delta_h^- \, u_1 &= u_1 - u_0 \\ &= \underbrace{u_1}_{\text{contribution à } A_h \, u_h} - \underbrace{\alpha}_{\text{contribution à } f_h} \end{aligned} \tag{2.96}$$

On est donc amené à considérer l'analogue discret constitué par le couple matrice-vecteur suivant :

$$A_h = \begin{pmatrix} 1 & & & & \\ -2 & \frac{3}{2} & & & \\ \frac{1}{2} & -2 & \frac{3}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{2} & -2 & \frac{3}{2} \end{pmatrix}, \; f_h = \begin{pmatrix} \alpha \\ -\frac{1}{2}\,\alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.97}$$

pour lequel $\delta_2^U \, u_h = A_h \, u_h - f_h$. A nouveau la matrice A_h est *triangulaire inférieure, non diagonalisable ou « défective »* ; ses valeurs propres sont évidentes :

$$\lambda_{h\,1} = 1 \, , \; \lambda_{h\,m} = \frac{3}{2} \quad (m = 2, 3, ..., M) \tag{2.98}$$

(voir figure 2.3).

2.2.5. L'opérateur de différence seconde centrée

$$\boxed{\delta_{hh} \; : \; \delta_{hh} \, u_j = u_{j-1} - 2 \, u_j + u_{j+1}} \tag{2.99}$$

Cas périodique

$$\boxed{\lambda_{h\,m} = e^{-\,i\,\theta_m} - 2 + e^{+\,i\,\theta_m} = -2 + 2 \, \cos \, \theta_m} \tag{2.100}$$

Quand m varie, $-\lambda_{h\,m}$ balaye l'intervalle [0,4] (et le décrit complètement dans la limite $h \to 0$).

Cas non périodique

Dans le cas de conditions aux limites de Dirichlet, la diagonalisation a été établie au chapitre 1 (cf. Théorème 1.1). On remarque que la formule ci-dessus pour les valeurs propres est encore valable dans le cas non périodique à condition de modifier la

définition du paramètre de fréquence comme suit :

$$\boxed{\theta_m = \frac{m\,\pi}{M+1}} \tag{2.101}$$

Comme dans le cas périodique, *le spectre est contenu dans l'intervalle [0,4]* et le recouvre dans la limite $h \to 0$. *Les vecteurs propres sont les discrétisés de fonctions sinusoïdales* $\{\psi_m(x)\}$ de (2.45) (voir figures 2.4-2.5).

Exercice 2.13 (Modes propres dans un cas de conditions de Dirichlet-Neumann)

(1) On s'intéresse ici aux fonctions satisfaisant une condition de Dirichlet homogène à gauche ($u(a) = 0$) et une condition de Neumann homogène à droite ($u'(b) = 0$). Autrement dit, l'espace de Hilbert de travail est ici :

$$H = \Big\{ u \in L^2\Big([a,b]\Big),\ u' \in L^2\Big([a,b]\Big),\ u(a) = u'(b) = 0 \Big\} \tag{2.102}$$

Identifier les fonctions et valeurs propres de l'opérateur de dérivée seconde opérant sur H.

(2) Afin d'identifier un analogue discret, on construit un maillage uniforme dont les nœuds ont les abscisses suivantes :

$$x_j = j\,h \tag{2.103}$$

où ici

$$h = \frac{1}{M + \frac{1}{2}} \tag{2.104}$$

Les degrés de liberté sont encore associés aux nœuds d'indice $j = 1, 2, ..., M$. Le point $j = 0$ est celui où l'on applique la condition de Dirichlet

$$u_0 = 0 \tag{2.105}$$

de sorte que

$$\begin{aligned} \delta_{hh}\, u_1 \quad &= u_0 - 2\,u_1 + u_2 \\ &= -2\,u_1 + u_2 \end{aligned} \tag{2.106}$$

La condition de Neumann est appliquée en introduisant un point fictif à l'abscisse x_{M+1} symétrique de x_M par rapport au point d'abscisse $x = 1$ et en imposant que :

$$u_{M+1} - u_M = 0 \tag{2.107}$$

ce qui constitue une discrétisation de la condition précise au second ordre. Cette condition discrète permet de simplifier l'expression de $\delta_{hh}\, u_M$:

$$\begin{aligned} \delta_{hh}\, u_M &= u_{M-1} - 2\, u_M + u_{M+1} \\ &= u_{M-1} - u_M \end{aligned} \tag{2.108}$$

Il en résulte que la matrice A_h associée à l'opérateur de différence seconde centrée dans le cas de conditions mixtes de Dirichlet-Neumann a la structure suivante :

$$A_h = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -1 \end{pmatrix} \tag{2.109}$$

S'inspirer des résultats de la question précédente pour deviner la forme des vecteurs propres de la matrice A_h et en déduire son spectre. Comparer au cas où des conditions de Dirichlet sont appliquées aux deux bords.

2.2.6. Les opérateurs discrets en plusieurs dimensions d'espace

Lorsque la structure des données est celle de produits tensoriels, ce qui est le cas lorsqu'on discrétise une EDP sur un rectangle (en 2D) ou un parallélépipède (en 3D) uniformément discrétisé, l'écriture formelle des opérateurs fait intervenir des « produits » et des « sommes » de Kronecker d'opérateurs unidimensionnels (voir annexe B). Au chapitre suivant, nous utiliserons cette technique pour identifier le spectre du laplacien discret en 2 ou 3 dimensions d'espace.

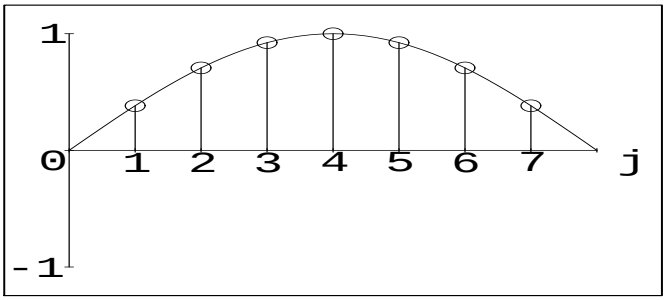
2.3. Localisation de valeurs propres

En analyse numérique, il arrive fréquemment, y compris dans l'étude de modèles théoriques linéaires très simplifiés, que l'on aboutisse à un problème de valeurs propres dont on ne connaît pas la solution exacte formelle. La cause de cette difficulté peut être liée par exemple à l'application de conditions aux limites particulières qui augmente la largeur de bande et détruit la structure régulière de la matrice, ou à la combinaison d'opérateurs qui ne commutent pas, comme l'advection-diffusion dans un cadre non périodique. Il reste cependant essentiel de localiser les valeurs propres dans le plan complexe, en particulier pour établir des conditions suffisantes (et/ou nécessaires) de stabilité de certains algorithmes de résolution itératifs (ou schémas d'intégration pseudo-temporels). La localisation des valeurs propres constitue un thème très vaste des mathématiques numériques, et on réfère en particulier aux ouvrages les plus connus [96] [48] [10] pour un exposé approfondi de la théorie. Dans cette section, deux théorèmes classiques sont rappelés afin d'en illustrer l'application aux modèles discrets fondamentaux.

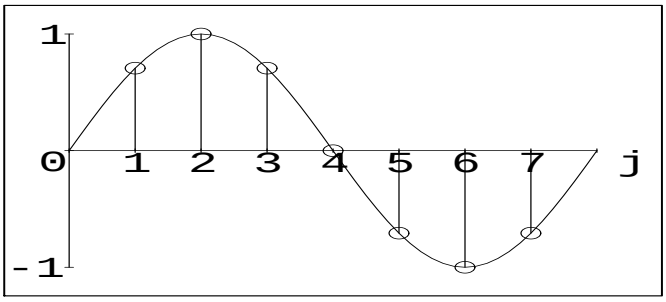
EXEMPLE : $M = 7$ degrés de liberté

1. Basses fréquences : $\theta_m < \frac{\pi}{2}$

$$s_{j,1} = \psi^{(1)}(x_j) \sim \sin(\pi\, x_j)$$



$$s_{j,2} = \psi^{(2)}(x_j) \sim \sin(2\,\pi\, x_j)$$



$$s_{j,3} = \psi^{(3)}(x_j) \sim \sin(3\,\pi\, x_j)$$

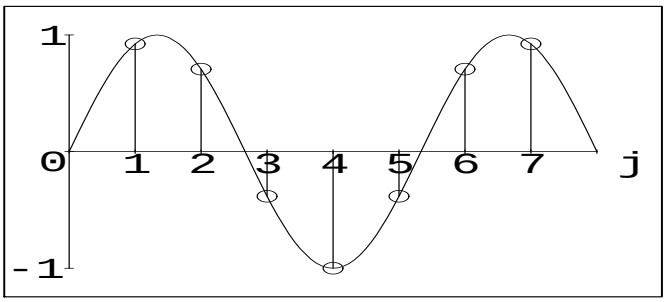


Figure 2.4. Modes de Fourier continus associés à l'opérateur de dérivée seconde, modes de Fourier discrets associés à l'opérateur de différence divisée seconde centrée (conditions de Dirichlet aux limites) – Basses fréquences

2. Hautes fréquences : $\theta_m \geq \frac{\pi}{2}$

$s_{j,4} = \psi^{(4)}(x_j) \sim \sin(4\,\pi\,x_j)$

$s_{j,5} = \psi^{(5)}(x_j) \sim \sin(5\,\pi\,x_j)$

$s_{j,6} = \psi^{(6)}(x_j) \sim \sin(6\,\pi\,x_j)$

$s_{j,7} = \psi^{(7)}(x_j) \sim \sin(7\,\pi\,x_j)$

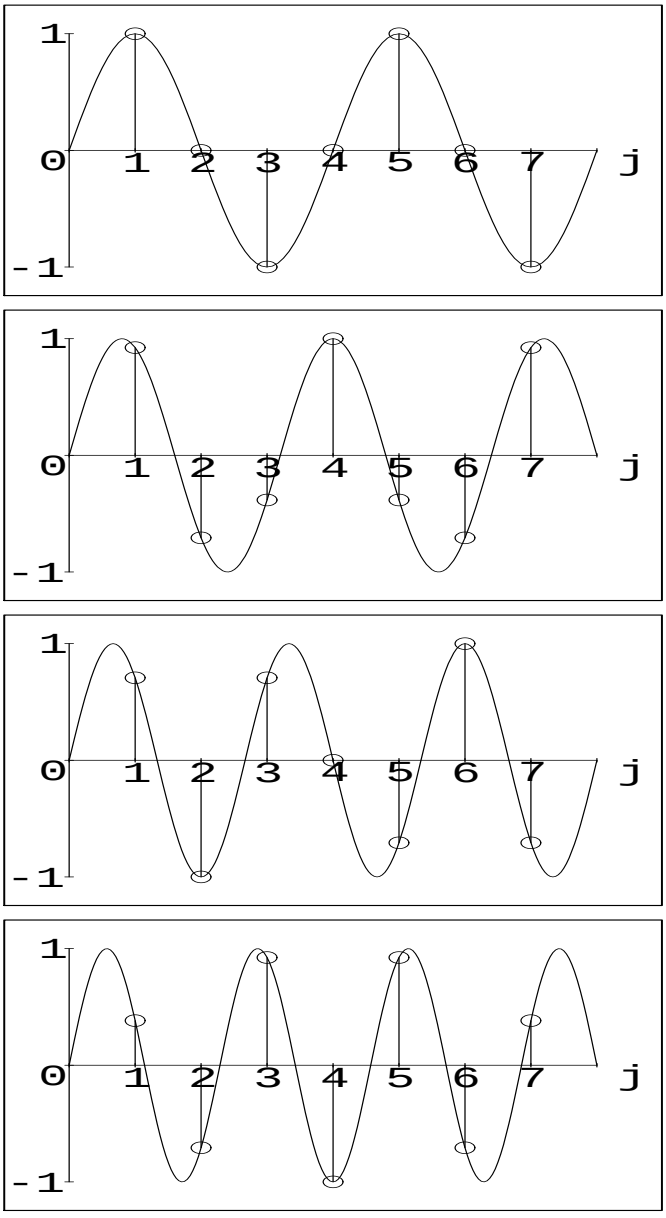


Figure 2.5. Modes de Fourier continus associés à l'opérateur de dérivée seconde, modes de Fourier discrets associés à l'opérateur de différence divisée seconde centrée (conditions de Dirichlet aux limites) – Hautes fréquences

2.3.1. Théorème de Gershgorin

Définition 2.12 (Disques de Gershgorin)

Etant donné une matrice carrée A dont les éléments $\{a_{j,k}\}$ ($j = 1, 2, ..., N\,;\ k = 1, 2, ..., N$) sont des complexes, on associe à chaque ligne j le disque fermé D_j du plan complexe dont le centre est situé à l'affixe $a_{j,j}$ et le rayon est égal à la somme des modules des éléments de la ligne hormis celui de la diagonale $R_j = \sum_{k=1\,;\ k\neq j}^{N} |a_{j,k}|$.

Théorème 2.4 (Gershgorin)

Les notations étant celles de la définition précédente, toute valeur propre λ de la matrice A appartient à l'un au moins des disques de Gershgorin, soit :

$$\forall\, \lambda \in \sigma(A)\,,\ \exists\, j \text{ tel que } \lambda \in D_j\,. \tag{2.110}$$

De manière équivalente, ce théorème précise que le spectre de A, noté $\sigma(A)$, est globalement inclus dans l'union des disques de Gershgorin, ce qui permet d'exclure la région du plan complexe qui est extérieure à cette union :

$$\sigma(A) \subset \bigcup_j D_j\,. \tag{2.111}$$

DÉMONSTRATION : soit $\lambda \in \sigma(A)$ et x un vecteur propre associé non nul :

$$A\,x = \lambda\,x\,,\ x \neq 0 \tag{2.112}$$

de sorte que pour tout indice j :

$$\sum_k a_{j,k}\,x_k = \lambda\,x_j \tag{2.113}$$

Les composantes du vecteur x ne sont pas toutes nulles. Particularisons désormais l'indice j à la valeur de l'indice m pour laquelle $|x_m|$ est maximum :

$$|x_j| = \max_m |x_m| \tag{2.114}$$

En conséquence :

$$|x_j| \neq 0 \text{ et } \forall\, k\,,\ \frac{|x_k|}{|x_j|} \leq 1 \tag{2.115}$$

En réarrangeant (2.113), il vient :

$$\lambda - a_{j,j} = \sum_{k\,,\ k\neq j} a_{j,k}\,\frac{|x_k|}{|x_j|} \tag{2.116}$$

puis :

$$|\lambda - a_{j,j}| \leq R_j \tag{2.117}$$

ce qui prouve que

$$\lambda \in D_j \,. \tag{2.118}$$

□

EXEMPLE :

La matrice réelle

$$A = \begin{pmatrix} 7 & 5 & 0 & 1 \\ 6 & 14 & -2 & 0 \\ 0 & 1 & 18 & 1 \\ -2 & 0 & 1 & 28 \end{pmatrix} \tag{2.119}$$

est à « diagonale dominante » stricte car chaque élément diagonal dépasse strictement la somme (des valeurs absolues) des éléments extra-diagonaux de la même ligne. Les disques de Gershgorin,

D_1 : disque de centre (7,0) de rayon 5+0+1=6,
D_2 : disque de centre (14,0) de rayon 6+2+0=8,
D_3 : disque de centre (18,0) de rayon 0+1+1=2,
D_4 : disque de centre (28,0) de rayon 2+0+1=3,

sont représentés à la figure 2.6 sur laquelle on a ombré le complémentaire de leur union. Il apparaît sur cette figure que les valeurs propres se situant dans cette union, sont toutes à parties réelles strictement positives ; en particulier aucune n'est nulle ; la matrice A est donc inversible. Par ailleurs, un calcul numérique approché de ces valeurs, indiquées sur la figure par le symbole •, a révélé que 2 sont réelles ($\lambda \approx 4.092\,,\ 28.00$) et 2 complexes conjuguées ($\lambda \approx 17.45 \pm 1.241\,i$, ce qui confirme la localisation fournie par le théorème.

REMARQUES :

– Le théorème de Gershgorin admet la conséquence suivante :

Théorème 2.5 (Dominance diagonale)

Toute matrice à diagonale strictement dominante est inversible.

DÉMONSTRATION : le résultat découle directement du fait que 0 n'appartient à aucun des disques de Gershgorin. □

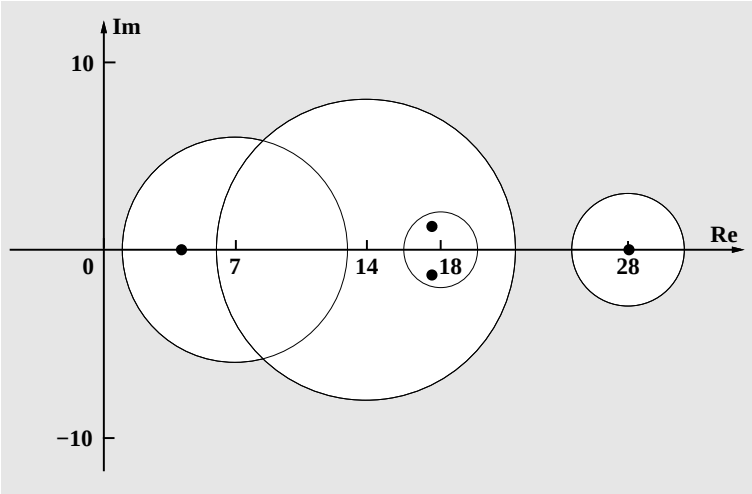


Figure 2.6. Illustration du théorème de Gershgorin – matrice définie en (2.119)

– En appliquant également le théorème à la matrice transposée A^T qui admet le même spectre, il vient :

$$\sigma(A) \subset \bigcup_k \tilde{D}_k \tag{2.120}$$

où $\tilde{D}_k$ est le disque fermé du plan complexe dont le centre est situé à l'affixe $a_{k,k}$ et dont le rayon $\tilde{R}_k$ est égal à la somme des éléments extra-diagonaux de la colonne k de la matrice A : $\tilde{R}_k = \sum_{j\,,\ j\neq k} |a_{j,k}|$. Donc finalement,

$$\sigma(A) \subset \left(\bigcup_j D_j \cap \bigcup_k \tilde{D}_k \right) \tag{2.121}$$

2.3.2. Théorème de Bendixson

Théorème 2.6 (Bendixson)

Soit A une matrice carrée quelconque dont les éléments sont complexes, et A^* la matrice adjointe. On pose :

$$H_1 = \frac{A + A^*}{2}\,, \quad H_2 = \frac{A - A^*}{2i} \tag{2.122}$$

de sorte que H_1 et H_2 sont des matrices hermitiennes ($H_1^* = H_1\,,\ H_2^* = H_2$) dont les valeurs propres sont réelles, et

$$A = H_1 + i\,H_2\,. \tag{2.123}$$

Soient $a,\ b,\ c,\ d$ *les réels suivants :*

$$\begin{aligned} a &= \lambda_{\min}(H_1) = \min\ \sigma(H_1)\,,\ b = \lambda_{\max}(H_1) = \max\ \sigma(H_1) \\ c &= \lambda_{\min}(H_2) = \min\ \sigma(H_2)\,,\ d = \lambda_{\max}(H_2) = \max\ \sigma(H_2) \end{aligned} \tag{2.124}$$

Alors, les parties réelle et imaginaire de toute valeur propre λ *de la matrice* A *vérifient les encadrements suivants :*

$$a \leq \Re(\lambda) \leq b\,,\ c \leq \Im(\lambda) \leq d\,. \tag{2.125}$$

DÉMONSTRATION : les matrices H_1 et H_2 étant hermitiennes, il existe des matrices unitaires U_1 et U_2,

$$U_1^*\,U_1 = I\,,\ U_2^*\,U_2 = I\,, \tag{2.126}$$

et des matrices diagonales réelles $\Lambda_1 = \text{Diag}\ (\alpha_m)$ et $\Lambda_2 = \text{Diag}\ (\beta_m)$ telles que :

$$H_1 = U_1\,\Lambda_1\,U_1^*\,,\ H_2 = U_2\,\Lambda_2\,U_2^* \tag{2.127}$$

On suppose que les réels $\{\alpha_m\}$ et $\{\beta_m\}$ sont ordonnés par valeurs croissantes :

$$\alpha_1 = a \leq \alpha_2 \leq ... \leq \alpha_N = b\,,\ \beta_1 = c \leq \beta_2 \leq ... \leq \beta_N = d \tag{2.128}$$

Soit $\lambda = \alpha + i\,\beta\ \in\ \sigma(A)$ et $z = x + i\,y$ un vecteur propre associé,

$$A\,z = \lambda\,z \tag{2.129}$$

que l'on peut supposer orthonormé :

$$\|z\|_2^2 = z^*\,z = 1 \tag{2.130}$$

de sorte que :

$$\begin{aligned} \lambda &= z^*\,A\,z \\ &= \left(x^T - i\,y^T\right)\,(H_1 + i\,H_2)\,(x + i\,y) \end{aligned} \tag{2.131}$$

ce qui conduit à :

$$\alpha = \Re(\lambda) = x^T\,H_1\,x + y^T\,H_1\,y \tag{2.132}$$

et :

$$\beta = \Im(\lambda) = x^T\,H_2\,x + y^T\,H_2\,y \tag{2.133}$$

où le caractère hermitien des matrices H_1 et H_2 a été utilisé. Examinons maintenant les variations du « quotient de Rayleigh » suivant :

$$R(\xi) = \frac{\xi^*\,H_1\,\xi}{\xi^*\,\xi} = \frac{\xi^*\,U_1\,\Lambda_1\,U_1^*\,\xi}{\xi^*\,\xi} \tag{2.134}$$

où ξ est un vecteur non nul quelconque. On définit le vecteur :

$$\chi = U_1^* \, \xi \tag{2.135}$$

de sorte que :

$$\|\chi\|_2^2 = \chi^* \, \chi = \xi^* \, U_1 \, U_1^* \, \xi = \xi^* \, \xi = \|\xi\|_2^2 \tag{2.136}$$

car U_1^*, opérateur unitaire, conserve la norme ; de plus :

$$\xi = U_1 \, \chi \tag{2.137}$$

ce qui permet d'expliciter l'expression de $R(\xi)$ comme suit :

$$R(\xi) = \frac{\chi^* \, \Lambda_1 \, \chi}{\chi^* \, \chi} = \alpha_1 \, \frac{|\chi_1|^2}{\|\chi\|_2^2} + \alpha_2 \, \frac{|\chi_2|^2}{\|\chi\|_2^2} + ... + \alpha_N \, \frac{|\chi_N|^2}{\|\chi\|_2^2} \tag{2.138}$$

où $\|\chi\|_2^2 = \sum_j |\chi_j|^2$. Cette expression fait apparaître que le quotient de Rayleigh est à valeurs dans $\mathbb{R}$ et que ses bornes sont respectivement la plus petite et la plus grande valeurs propres de la matrice H_1 :

$$\min_{\xi \, , \, \xi \neq 0} \, R(\xi) = \alpha_1 = a \, , \; \max_{\xi \, , \, \xi \neq 0} \, R(\xi) = \alpha_N = b \tag{2.139}$$

Le vecteur x étant réel, il en résulte que :

$$a \left(x^T \, x \right) \leq x^T \, H_1 \, x \leq b \left(x^T \, x \right) \tag{2.140}$$

et de même pour y ; finalement :

$$a \leq \alpha \leq b \tag{2.141}$$

car $x^T \, x + y^T \, y = \|z\|_2^2 = 1$. En remplaçant H_1 par H_2 on obtient un encadrement analogue de la partie imaginaire :

$$c \leq \beta \leq d \tag{2.142}$$

ce qui complète la démonstration. □

Dans le cas d'une matrice réelle $A = \bar{A}$, les matrices H_1 et iH_2 sont les parties symétrique et antisymétrique de A :

$$H_1 = \frac{A + A^T}{2} \stackrel{\text{déf}}{=} S = S^T \, , \quad iH_2 = \frac{A - A^T}{2} \stackrel{\text{déf}}{=} \Sigma = -\Sigma^T \tag{2.143}$$

Le rectangle $[a, b] \times [c, d]$ est alors symétrique par rapport à l'axe des imaginaires, comme le schématise la figure 2.7.

EXEMPLE D'APPLICATION :

A titre d'illustration, on s'intéresse au spectre de la matrice A_h associée à l'opérateur de différence centrée dans le cas non périodique défini au paragraphe 2.2.2 par l'équation (2.86). Ces valeurs propres, précédemment indiquées à la figure 2.1 dans le cas où la dimension $N = M = 15$, ont des parties imaginaires dominantes et comprises entre -1 et 1, et des parties réelles strictement positives. On peut montrer par l'étude d'une équation transcendentale que, lorsque $h = \frac{1}{M} \to 0$, les parties imaginaires forment un ensemble dense dans [-1,1], et les parties réelles tendent vers 0. En ce sens, le spectre diffère peu du cas périodique. Essayons alternativement de retrouver cette localisation par application des théorèmes précédents.

L'application du théorème de Gershgorin à la matrice A_h fournit la localisation indiquée à la figure 2.8. On en conclut notamment que les parties imaginaires sont comprises entre -1 et 1, et les parties réelles entre -1 et 2.

L'application du théorème de Gerhgorin à la matrice transposée

$$A_h^T = \begin{pmatrix} 0 & -\frac{1}{2} & & & \\ \frac{1}{2} & 0 & -\frac{1}{2} & & \\ & \ddots & \ddots & -\frac{1}{2} & \\ & & \frac{1}{2} & 0 & -1 \\ & & & \frac{1}{2} & 1 \end{pmatrix} \tag{2.144}$$

permet d'affiner légèrement le résultat (voir figure 2.9). Désormais on sait que les parties réelles sont comprises entre -1 et $\frac{3}{2}$.

Dans le but d'appliquer le théorème de Bendixson à la matrice A_h qui est réelle, on calcule ses parties symétrique,

$$S = \frac{1}{2}\left(A_h + A_h^T\right) = \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & -\frac{1}{4} \\ & & & -\frac{1}{4} & 1 \end{pmatrix} \tag{2.145}$$

et antisymétrique :

$$\Sigma = \frac{1}{2}\left(A_h - A_h^T\right) = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{2} & 0 & \frac{3}{4} \\ & & & -\frac{3}{4} & 0 \end{pmatrix} \tag{2.146}$$

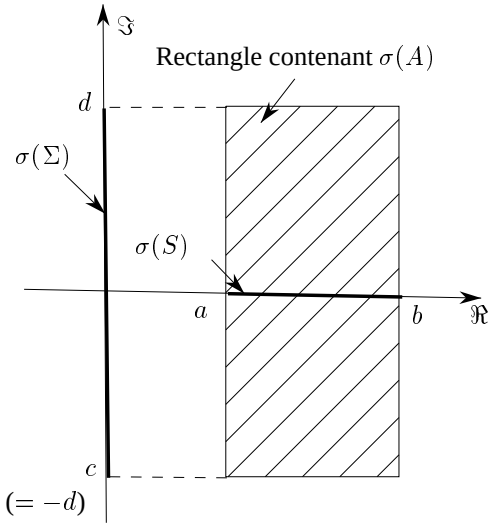


Figure 2.7. Localisation du spectre d'une matrice réelle par le théorème de Bendixson.

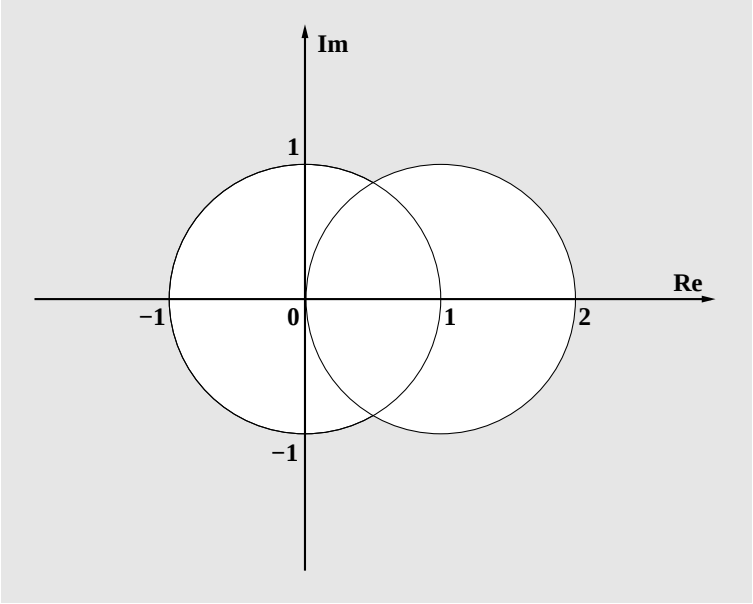


Figure 2.8. Localisation du spectre par application du théorème de Gershgorin à la matrice A_h.

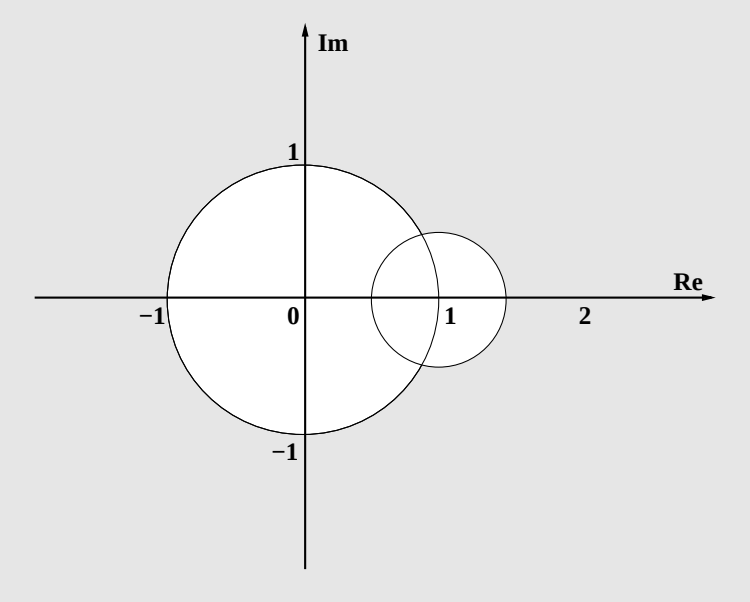


Figure 2.9. Localisation du spectre par application du théorème de Gershgorin à la matrice A_h^T.

Les valeurs propres (réelles) de la partie symétrique S sont $\lambda = 0$ (de multiplicité $N-2$), $\lambda = \left[1 - \sqrt{17}/4\right]/2 = a$ et $\lambda = \left[1 + \sqrt{17}/4\right]/2 = b$. Les valeurs propres (purement imaginaires) de la partie antisymétrique Σ ont des ordonnées comprises entre $c = -\frac{5}{4}$ et $d = \frac{5}{4}$ en vertu du théorème de Gershgorin. Ces considérations permettent d'exclure l'extérieur du rectangle $[a,b] \times [c,d]$ comme indiqué à la figure 2.10.

Malheureusement, bien que très proche de 0, le réel a est strictement négatif. On ne peut donc à ce stade exclure la possibilité de valeurs propres à parties réelles négatives. Pour atteindre ce but, il convient au préalable d'équilibrer la matrice A_h. Pour cela, on introduit la matrice diagonale suivante :

$$D = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & \sqrt{2} \end{pmatrix} \tag{2.147}$$

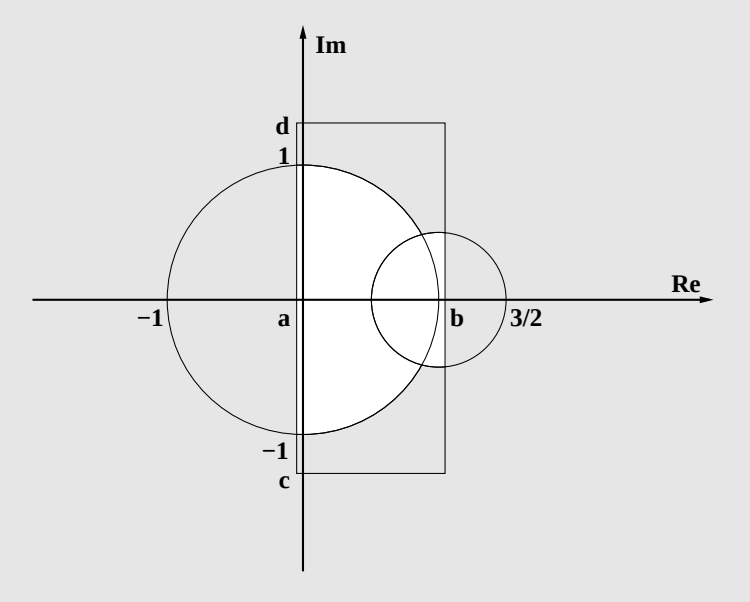


Figure 2.10. Localisation du spectre par application du théorème de Bendixson à la matrice A_h

et on remplace la matrice A_h par la matrice semblable suivante :

$$\tilde{A} = D^{-1}\, A_h\, D = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ -\frac{1}{2} & 0 & \ddots & & \\ & \ddots & \ddots & \frac{1}{2} & \\ & & -\frac{1}{2} & 0 & \frac{\sqrt{2}}{2} \\ & & & -\frac{\sqrt{2}}{2} & 1 \end{pmatrix} \tag{2.148}$$

L'application du théorème de Gershgorin aux matrices $\tilde{A}$ et $\tilde{A}^T$ ne conduit pas à une localisation plus fine que celle de la figure 2.9 que l'on conserve en vue. La partie symétrique de la matrice $\tilde{A}$,

$$\tilde{S} = \tfrac{1}{2}\left(\tilde{A} + \tilde{A}^T\right) = \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & 1 \end{pmatrix} \tag{2.149}$$

admet les valeurs propres $\lambda = a' = 0$ (de multiplicité $N-1$) et $\lambda = b' = 1$, et la partie antisymétrique,

$$\tilde{\Sigma} = \frac{1}{2}\left(\tilde{A} - \tilde{A}^T\right) = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ -\frac{1}{2} & 0 & \ddots & & \\ & \ddots & \ddots & \frac{1}{2} & \\ & & -\frac{1}{2} & 0 & \frac{\sqrt{2}}{2} \\ & & & -\frac{\sqrt{2}}{2} & 0 \end{pmatrix} \tag{2.150}$$

admet des valeurs propres (purement imaginaires) dont les ordonnées sont comprises entre $c' = -(1+\sqrt{2})/2$ et $d' = (1+\sqrt{2})/2$ (en vertu du théorème de Gershgorin). On peut donc exclure l'extérieur du rectangle $[a', b'] \times [c', d']$ et aboutir à la localisation (encore très approximative) de la figure 2.11 qui prouve néanmoins que la partie réelle d'aucune valeur propre n'est strictement négative, comme on sait.

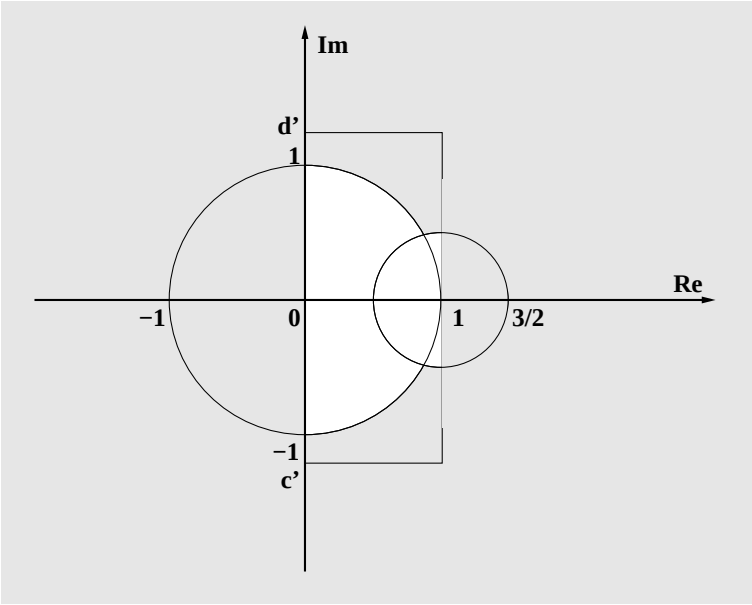


Figure 2.11. Localisation du spectre par application du théorème de Bendixson à la matrice $\tilde{A}$

2.4. Perturbations de matrices

Dans cette section, on s'intéresse encore aux situations où la diagonalisation formelle d'une matrice n'est pas connue, mais ici, on suppose que l'on sait diagonaliser une matrice « proche ». Si la différence entre ces deux matrices est suffisamment petite, peut-on garantir que la matrice d'origine est effectivement diagonalisable, et si

oui, peut-on approcher son spectre ? Ces questions relèvent de la très vaste théorie des équations algébriques dont on se contente ici de citer quelques résultats très classiques avant de les appliquer aux matrices associées aux modèles fondamentaux. On part du théorème suivant pour lequel on renvoie à [3] (chapitre 8) pour une démonstration :

Théorème 2.7 (Analyticité des zéros d'un polynôme)

Soit $N \geq 1$ un entier fixé, ε un « petit » paramètre, et $P_{N,\varepsilon}(\lambda)$ un polynôme en λ de degré N dont les coefficients sont des fonctions polynômiales de ε :

$$P_{N,\varepsilon}(\lambda) = a_N(\varepsilon)\,\lambda^N + a_{N-1}(\varepsilon)\,\lambda^{N-1} + \cdots + a_1(\varepsilon)\,\lambda + a_0(\varepsilon) \tag{2.151}$$

On suppose que $a_N(0) \neq 0$ et que le discriminant du polynôme $P_{N,0}(\lambda)$ n'est pas nul non plus, de sorte que les zéros de ce polynôme sont distincts. Alors, dans un voisinage de $\varepsilon = 0$, les zéros du polynôme $P_{N,\varepsilon}(\lambda)$ s'expriment par des séries entières de ε.

REMARQUE :

Dans le cas inverse où le discriminant est nul, le polynôme $P_{N,0}(\lambda)$ admet au moins un zéro de multiplicité $\alpha > 1$; il existe alors parmi les zéros du polynôme $P_{N,\varepsilon}(\lambda)$ des sous-groupes $\{G_\ell\}$ ($\ell = 1, 2, \ldots$; card $(G_\ell) = \alpha_\ell \geq 1$; $\sum_\ell \alpha_\ell = \alpha$) pour lesquels les zéros s'expriment par des séries entières de $\varepsilon^{1/\alpha_\ell}$ (cf. [96], chapitre 2). Par conséquent dans le cas le plus dégénéré, ces zéros s'expriment par des séries entières de $\varepsilon^{1/N}$.

Ce théorème implique en particulier que les zéros du polynôme $P_{N,\varepsilon}(\lambda)$ sont dans tous les cas des fonctions de ε continues en 0, prenant donc des valeurs distinctes pour ε suffisamment petit. Ce théorème admet le corollaire suivant :

Corollaire 2.1 (Analyticité des valeurs propres d'une matrice)

Soit $N \geq 1$ un entier fixé, ε un « petit » paramètre, et $A(\varepsilon)$ une matrice de dimension $N \times N$ dont les coefficients sont des fonctions affines de ε :

$$A(\varepsilon) = A_0 + \varepsilon\, A_0' \tag{2.152}$$

Si pour $\varepsilon = 0$ la matrice $A_0 = A(0)$ admet la diagonalisation suivante :

$$A_0 = X_0\, \Lambda_0\, X_0^{-1} \tag{2.153}$$

dans laquelle les matrices X_0 et Λ_0 sont de dimension $N \times N$, X_0 est inversible et Λ_0 diagonale, et si les valeurs propres de la matrice Λ_0 (c'est-à-dire de A_0) sont distinctes, alors il existe des séries entières de ε de la forme :

$$\left\{ \begin{array}{l} \Lambda(\varepsilon) = \Lambda_0 + \varepsilon\, \Lambda_0' + \ldots \\ X(\varepsilon) = X_0 + \varepsilon\, X_0' + \ldots \end{array} \right. \tag{2.154}$$

telles que pour tout ε suffisamment petit, $\Lambda(\varepsilon)$ est diagonale, $X(\varepsilon)$ inversible, et

$$A(\varepsilon) = X(\varepsilon)\,\Lambda(\varepsilon)X(\varepsilon)^{-1} \tag{2.155}$$

DÉMONSTRATION : il convient de considérer séparément les cas des valeurs et des vecteurs propres. Notons d'abord que les coefficients du polynôme caractéristique de la matrice $A(\varepsilon)$,

$$P_{N,\varepsilon}(\lambda) = \det\Big(A(\varepsilon) - \lambda I\Big) \tag{2.156}$$

sont des polynômes (homogènes) des coefficients de la matrice $A(\varepsilon)$, donc des polynômes de ε. Les zéros du polynôme $P_{N,0}(\lambda)$ qui sont les valeurs propres de la matrice A_0, sont distincts par hypothèse. Le polynôme $P_{N,\varepsilon}(\lambda)$ satisfait donc les hypothèses du théorème précédent, ce qui garantit, pour ε suffisamment petit, l'existence de valeurs propres distinctes pour la matrice $A(\varepsilon)$ développables en séries entières, d'où la matrice diagonale :

$$\Lambda(\varepsilon) = \Lambda_0 + \varepsilon\,\Lambda_0' + \dots \tag{2.157}$$

Les valeurs propres étant distinctes dans un voisinage de $\varepsilon = 0$, on peut les « suivre » par continuité à partir de l'ordre (arbitraire) qu'elles occupent dans la diagonale de la matrice Λ_0. On adopte bien évidemment le même ordre pour les vecteurs propres associés. La question de la normalisation est plus délicate. Il s'agit de prouver que pour tout ε, on peut associer à chaque valeur propre $\lambda(\varepsilon)$ un unique vecteur propre $\mathsf{x}(\varepsilon)$ de telle sorte que la fonction $\mathsf{x}(\varepsilon)$ soit analytique. Le vecteur $\mathsf{x}(\varepsilon)$ est (incomplètement) défini par la condition suivante :

$$\Big(A(\varepsilon) - \lambda(\varepsilon)\,I\Big)\,\mathsf{x}(\varepsilon) = 0 \tag{2.158}$$

qui constitue pour les composantes du vecteur inconnu $\mathsf{x}(\varepsilon)$ un système d'équations linéaires homogène dont les coefficients sont des fonctions analytiques de ε, et dont on sait qu'il est compatible et indéterminé, et de rang $N-1$, car par hypothèse, toutes les valeurs propres sont simples. Les solutions de ce système forment un espace vectoriel de dimension 1.

Pour $\varepsilon = 0$, le système à résoudre s'écrit :

$$\Big(A_0 - \lambda_0\,I\Big)\,\mathsf{x}(0) = 0 \quad (\lambda_0 = \lambda(0)) \tag{2.159}$$

Puisque le rang est égal à $N-1$, il existe au moins un couple d'indices (j,k) tel que le bloc B_0 de dimension $(N-1)\times(N-1)$ obtenu en éliminant la ligne j et la colonne k de la matrice $A_0 - \lambda_0\,I$ est inversible,

$$\det B_0 \neq 0 \tag{2.160}$$

alors que $\det\Big(A_0 - \lambda_0\, I\Big) = 0$. Dans ce cas, dans le système (2.159), on peut éliminer la j-ième équation, qui est redondante, et faire apparaître la k-ième composante du vecteur $\mathsf{x}(0)$ comme un paramètre :

$$B_0\, \overline{\mathsf{x}}(0) = -C_0^{j,k}\, \mathsf{x}_k(0) \tag{2.161}$$

où $\overline{\mathsf{x}}(0)$ et $C_0^{j,k}$ sont des vecteurs de dimension $N-1$; $\overline{\mathsf{x}}(0)$ est obtenu en éliminant du vecteur $\mathsf{x}(0)$ sa k-ème composante ; $C_0^{j,k}$ est obtenu en éliminant du k-ème vecteur colonne de la matrice $A_0 - \lambda_0\, I$ sa j-ème composante. D'où la solution formelle :

$$\overline{\mathsf{x}}(0) = -\Big(B_0\Big)^{-1}\, C_0^{j,k}\, \mathsf{x}_k(0)\,, \quad \mathsf{x}_k(0) \text{ arbitraire} \tag{2.162}$$

On peut choisir $\mathsf{x}_k(0)$ de manière que le vecteur qui en résulte soit identique au vecteur colonne approprié de la matrice X_0.

L'identification d'indices j et k rendant ces manipulations possibles s'appuie sur le fait que pour $\varepsilon = 0$, un certain déterminant n'est pas nul. Cette propriété reste vraie pour $\varepsilon \neq 0$ mais suffisamment petit, en raison de la continuité des fonctions analytiques. Par conséquent, en définissant maintenant la matrice B_ε comme le bloc de dimension $(N-1)\times(N-1)$ obtenu à partir de la matrice $A(\varepsilon) - \lambda(\varepsilon)\, I$ en éliminant la ligne j et la colonne k, il reste vrai dans un voisinage de $\varepsilon = 0$ que

$$\det B_\varepsilon \neq 0\,. \tag{2.163}$$

Par conséquent, le système (2.158) équivaut à :

$$B_\varepsilon\, \overline{\mathsf{x}}(\varepsilon) = -C_\varepsilon^{j,k}\, \mathsf{x}_k(\varepsilon) \tag{2.164}$$

où $\overline{\mathsf{x}}(\varepsilon)$ et $C_\varepsilon^{j,k}$ sont des vecteurs de dimension $N-1$; $\overline{\mathsf{x}}(\varepsilon)$ est obtenu en éliminant du vecteur $\mathsf{x}(\varepsilon)$ sa k-ième composante ; $C_\varepsilon^{j,k}$ est obtenu en éliminant du k-ième vecteur colonne de la matrice $A(\varepsilon) - \lambda(\varepsilon)\, I$ sa j-ième composante. D'où la solution formelle :

$$\overline{\mathsf{x}}(\varepsilon) = -\Big(B_\varepsilon\Big)^{-1}\, C_\varepsilon^{j,k}\, \mathsf{x}_k(\varepsilon)\,, \quad \mathsf{x}_k(\varepsilon) \text{ arbitraire} \tag{2.165}$$

En choisissant par exemple la normalisation suivante :

$$\mathsf{x}_k(\varepsilon) = \mathsf{x}_k(0) \quad (\varepsilon \neq 0) \tag{2.166}$$

on complète une définition unique possible du vecteur $\mathsf{x}(\varepsilon)$. De plus, l'expression ci-dessus, (2.165), ne fait alors intervenir que des fractions rationnelles de fonctions analytiques en ε, qui sont elles-mêmes analytiques dans tout ouvert qui ne contient pas de pôle, ce qui est bien le cas d'un voisinage de $\varepsilon = 0$; ce résultat fournit donc une définition analytique des vecteurs propres. En rassemblant ces vecteurs dans l'ordre convenable, on aboutit à la fonction analytique suivante à valeur matricielle :

$$X(\varepsilon) = X_0 + \varepsilon\, X_0' + \ldots \tag{2.167}$$

qui satisfait

$$A(\varepsilon)\, X(\varepsilon) = X(\varepsilon)\, \Lambda(\varepsilon) \tag{2.168}$$

par définition des vecteurs propres, ce qui implique, vu que les valeurs propres sont distinctes, que la matrice $X(\varepsilon)$ est inversible, et donc que :

$$A(\varepsilon) = X(\varepsilon)\, \Lambda(\varepsilon)\, X(\varepsilon)^{-1}\,,\ \forall \varepsilon. \tag{2.169}$$

□

Ce corollaire a confirmé la possibilité de diagonaliser la matrice $A(\varepsilon)$ dans un voisinage de $\varepsilon = 0$, sans préciser de résultat quantitatif. Le théorème suivant, à l'inverse, fournit l'expression de la *première variation* des valeurs propres. Afin d'en simplifier l'expression, on introduit la notation suivante :

Définition 2.13 (Notation d'opérateur partie-diagonale, Dg)

On convient de désigner par « opérateur partie-diagonale », noté Dg , l'opérateur qui transforme une matrice carrée quelconque M en la matrice diagonale $D = \mathrm{Dg}\,(M)$ ayant les mêmes éléments diagonaux ($D_{j,k} = \delta_{j,k}\, M_{j,k}\,,\ \forall\, j, k$).

REMARQUE :

Cet opérateur est linéaire et n'a aucun effet sur une matrice diagonale.

Théorème 2.8 (Premières variations des valeurs propres)

Les hypothèses du corollaire 2.1 étant faites, on a :

$$\boxed{\Lambda_0' = \mathrm{Dg}\left(X_0^{-1}\, A_0'\, X_0\right)} \tag{2.170}$$

Si l'on introduit la notation suivante pour les « premières variations »,

$$\begin{cases} \delta A \overset{\text{déf}}{=} \varepsilon\, A_0' \\ \delta \Lambda \overset{\text{déf}}{=} \varepsilon\, \Lambda_0' \end{cases} \tag{2.171}$$

le résultat de ce théorème peut s'exprimer de manière équivalente comme suit :

$$\boxed{\delta\Lambda = \mathrm{Dg}\left(X_0^{-1}\, \delta A\, X_0\right)} \tag{2.172}$$

DÉMONSTRATION : on tire de l'équation (2.155) l'expression suivante de la matrice des valeurs propres

$$\Lambda(\varepsilon) = X(\varepsilon)^{-1}\, A(\varepsilon)\, X(\varepsilon) \tag{2.173}$$

que l'on dérive par rapport à ε (chaque facteur séparément) ; puis on fait $\varepsilon = 0$:

$$\begin{aligned} \Lambda_0' &= X_0^{-1}\, A_0'\, X_0 + \left(X^{-1}\right)_0'\, A_0\, X_0 + X_0^{-1}\, A_0\, X_0' \\ &= X_0^{-1}\, A_0'\, X_0 + \left(X^{-1}\right)_0'\, X_0\, \Lambda_0 + \Lambda_0\, X_0^{-1}\, X_0' \end{aligned} \tag{2.174}$$

En outre, en dérivant la matrice constante $X(\varepsilon)^{-1}\, X(\varepsilon) = I$, il vient :

$$\left(X^{-1}\right)_0'\, X_0 + X_0^{-1}\, X_0' = 0 \tag{2.175}$$

de sorte qu'en posant

$$B \stackrel{\text{déf}}{=} \left(X^{-1}\right)_0'\, X_0 \tag{2.176}$$

on a :

$$\Lambda_0' = X_0^{-1}\, A_0'\, X_0 + B\, \Lambda_0 - \Lambda_0\, B \tag{2.177}$$

Enfin, on remarque que pour tout indice j :

$$\begin{aligned} \left(B\, \Lambda_0 - \Lambda_0\, B\right)_{j,j} &= \textstyle\sum_m B_{j,m} \left(\Lambda_0\right)_{m,j} - \sum_\ell \left(\Lambda_0\right)_{j,\ell} B_{\ell,j} \\ &= B_{j,j}\, \lambda_{0,j} - \lambda_{0,j}\, B_{j,j} \\ &= 0 \end{aligned} \tag{2.178}$$

($\lambda_{0,j}$: valeur pour $\varepsilon = 0$ de la j-ième valeur propre). Par conséquent

$$\mathrm{Dg}\left(B\, \Lambda_0 - \Lambda_0\, B\right) = 0 \tag{2.179}$$

et en appliquant l'opérateur Dg qui est linéaire et sans effet sur une matrice diagonale à (2.177), il vient :

$$\Lambda_0' = \mathrm{Dg}\left(\Lambda_0'\right) = \mathrm{Dg}\left(X_0^{-1}\, A_0'\, X_0\right) \tag{2.180}$$

□

Pour illustrer ce résultat, on considère le cas d'une équation d'advection-diffusion,

$$c\, u_x = \mu\, u_{xx} \quad (0 \leq x \leq 1) \quad (c, \mu > 0) \tag{2.181}$$

discrétisée par différences finies centrées sur un maillage uniforme $\{x_j\}$ ($j = 0, 1, ..., M+1\,;\ x_j = jh\,;\ h = 1/(M+1)$) :

$$\frac{c}{2h}\,(u_{j+1} - u_{j-1}) = \frac{\mu}{h^2}\,(u_{j+1} - 2u_j + u_{j-1}) \quad (j = 1, 2, ..., M) \tag{2.182}$$

et soumise aux conditions de Dirichlet suivantes :

$$u_0 = u(0) = \alpha\,,\ u_{M+1} = u(1) = \beta \tag{2.183}$$

En rassemblant ces équations, on obtient le système linéaire suivant :

$$A_h\,u_h = f_h \tag{2.184}$$

où, ici :

$$\boxed{A_h = \text{Trid}_{DD}\,\left(-\tfrac{1}{2}, 0, \tfrac{1}{2}\right) + \varepsilon\,\text{Trid}_{DD}\,(-1, 2, -1)} \tag{2.185}$$

est une matrice tridiagonale de dimension $M \times M$ dans laquelle intervient le paramètre ε qui est égal à l'inverse du « nombre de Reynolds de maille » (ou « nombre de Péclet de maille ») :

$$\boxed{\varepsilon = \frac{1}{\text{Re}_h} = \frac{\mu}{c\,h}} \tag{2.186}$$

On s'intéresse au spectre de la matrice d'approximation A_h particulièrement dans deux cas limites suivants :

(a) advection dominante : $\varepsilon \ll 1$;

(b) diffusion dominante : $\varepsilon \gg 1$.

En fait, la simplicité de cet exemple fait que l'on peut facilement diagonaliser la matrice A_h de manière formelle et complète et vérifier ainsi le bien-fondé de l'approximation (2.180). A cette fin, on considère d'abord la structure tridiagonale générale d'un opérateur de différences finies à 3 points lorsqu'il est soumis à des conditions de Dirichlet, à savoir :

$$A = \text{Trid}_{DD}\,(a, b, c) \overset{\text{déf}}{=} \begin{pmatrix} b & c & & & \\ a & b & c & & \\ & \ddots & \ddots & \ddots & \\ & & a & b & c \\ & & & a & b \end{pmatrix} \tag{2.187}$$

($a, b, c \in \mathbb{R}$). Pour identifier les éléments spectraux, on distingue deux cas :

1er cas : $a\,c > 0$ (signe(a) = signe(c)).

On considère alors les M vecteurs $\{X_m\}$ ($m = 1, 2, .., M$) dont les composantes sont les suivantes :

$$\boxed{X_{j,m} = \sqrt{\frac{2}{M+1}} \left(\sqrt{\frac{a}{c}}\right)^j \sin\, j\theta_m \quad (j = 1, 2, ..., M)} \tag{2.188}$$

où le paramètre de fréquence a la définition habituelle :

$$\boxed{\theta_m = \frac{m\,\pi}{M+1}} \tag{2.189}$$

il vient :

$$\begin{aligned} (A\,X_m)_\ell &= \sum_{j=1}^{M} A_{\ell,j}\,X_{j,m} \\ &= a\,X_{\ell-1,m} + b\,X_{\ell,m} + c\,X_{\ell+1,m} \\ &= \left[b + \sqrt{a\,c}\,\frac{\text{signe}(a)\,\sin(\ell-1)\theta_m + \text{signe}(c)\,\sin(\ell+1)\theta_m}{\sin\,\ell\theta_m}\right] X_{\ell,m} \\ &= \lambda_m\,X_{\ell,m} \end{aligned} \tag{2.190}$$

où après simplification :

$$\boxed{\lambda_m = b + 2\,\text{signe}(c)\,\sqrt{a\,c}\,\cos\,\theta_m} \tag{2.191}$$

Puisque le nombre complexe λ_m ne dépend que de l'indice m, ces équations prouvent que le vecteur X_m est un vecteur propre de la matrice A associé à la valeur propre λ_m. Ces valeurs propres étant distinctes et en nombre M, on obtient donc ainsi la diagonalisation complète de la matrice A :

$$\boxed{A = X\,\Lambda\,X^{-1}} \tag{2.192}$$

Quand de plus, $a = c$, la matrice A, alors réelle-symétrique, est, à l'addition près d'une matrice scalaire, égale à un multiple de la matrice de discrétisation du modèle

fondamental $A_h = h^{-2}\,\mathrm{Trid}_{D\,D}\,(-1,2,-1)$. La matrice $X = (X_{j,m})$ des vecteurs propres s'identifie alors à la matrice S_h (orthogonale et symétrique) des modes de Fourier discrets associés à ce modèle (voir théorème 1.1).

2e cas : $a\,c < 0$ (signe$(a) = -$signe(c)).

On pose alors:

$$\boxed{X_{j,m} = \sqrt{\frac{2}{M+1}}\,\left(i\sqrt{\left|\frac{a}{c}\right|}\right)^{j}\,\sin\,j\theta_m \quad (j = 1,2,\ldots,M)} \tag{2.193}$$

où i est le symbole des imaginaires. On obtient ici :

$$\begin{aligned}\frac{(A\,X_m)_\ell}{X_{\ell,m}} &= b + \sqrt{|a\,c|}\,\frac{\dfrac{\text{signe}(a)}{i}\,\sin\,(\ell-1)\theta_m + i\,\text{signe}(c)\,\sin\,(\ell+1)\theta_m}{\sin\,\ell\theta_m}\\ &= \lambda_m\end{aligned} \tag{2.194}$$

indépendamment de ℓ, où, après simplification :

$$\boxed{\lambda_m = b + 2\,\text{signe}(c)\,i\,\sqrt{|a\,c|}\,\cos\,\theta_m} \tag{2.195}$$

Puisque le nombre complexe λ_m ne dépend que de l'indice m, ces équations prouvent que le vecteur X_m est encore un vecteur propre de la matrice A associé à la valeur propre λ_m. Ces valeurs propres étant distinctes et en nombre M, on obtient donc ainsi la diagonalisation complète de la matrice A, conformément à (2.192).

On revient maintenant à l'étude du spectre de la matrice A_h associée au modèle d'advection-diffusion, (2.185), pour laquelle les paramètres $a,\ b,\ c$ de la structure tridiagonale ont les valeurs suivantes :

$$a = -\tfrac{1}{2} - \varepsilon\,,\ b = 2\,\varepsilon\,,\ c = \tfrac{1}{2} - \varepsilon \tag{2.196}$$

de sorte que $a < 0$, $b > 0$ et c est positif ou négatif suivant que ε est inférieur ou supérieur à $\frac{1}{2}$.

Dans le cas de l'advection dominante, $\varepsilon \ll 1$, on peut développer (2.195) suivant les puissances croissantes de ε :

$$\lambda_{h\,m} = 2\,\varepsilon + 2\,i\,\sqrt{\frac{1}{4} - \varepsilon^2}\,\cos\,\theta_m = i\,\cos\,\theta_m + 2\,\varepsilon + O\left(\varepsilon^2\right) \tag{2.197}$$

Alternativement, on peut identifier des matrices A_0 et A'_0 telles que

$$A_h = A_0 + \varepsilon\, A'_0 \tag{2.198}$$

à savoir :

$$A_0 = \text{Trid}_{DD}\ \left(-\tfrac{1}{2}, 0, \tfrac{1}{2}\right) \tag{2.199}$$

matrice de discrétisation centrée de la dérivée première, et :

$$A'_0 = \text{Trid}_{DD}\,(-1, 2, -1) = 2\,I - \left(\mathcal{B} + \mathcal{B}^T\right) \tag{2.200}$$

matrice de discrétisation centrée de la dérivée seconde, où $\mathcal{B} = \text{Trid}_{DD}\,(0, 0, 1)$. Pour l'advection pure, l'expression des vecteurs propres $X_0 = \{X_{0_{j,m}}\}$, se simplifie comme suit :

$$X_{0_{j,m}} = \sqrt{\frac{2}{M+1}}\, i^j\ \sin\ j\theta_m \quad (j = 1, 2, ..., M) \tag{2.201}$$

où l'indice 0 réfère aux valeurs prises pour $\varepsilon = 0$, et les valeurs propres correspondantes sont les suivantes :

$$\lambda_{h_{0_m}} = i\ \cos\ \theta_m \tag{2.202}$$

ce qui est bien la partie principale de (2.197). La perturbation d'ordre ε est, d'après (2.180), égale à la quantité suivante :

$$\begin{aligned} \delta\lambda_{h\,m} &= \delta\Lambda_{m,m} \\ &= \varepsilon \left(X_0^*\ \left(2\,I - \mathcal{B} - \mathcal{B}^T\right)\ X_0\right)_{m,m} \\ &= \varepsilon\,(2 - z_m - \overline{z_m}) \\ &= 2\,\varepsilon\ (1 - \Re(z_m)) \end{aligned} \tag{2.203}$$

où le nombre complexe z_m s'exprime comme suit :

$$\begin{aligned} z_m &= (X_0^*\,\mathcal{B}\,X_0)_{m,m} \\ &= \sum_{j,\ell} X^*_{0\,m,j}\,\mathcal{B}_{j,\ell}\,X_{0\,\ell,m} \\ &= \sum_{j} X^*_{0\,m,j}\,X_{0\,j+1,m} \\ &= \sum_{j} \overline{X_{0\,j,m}}\,X_{0\,j+1,m} \\ &= \sum_{j} \frac{2}{M+1}\,i\ \sin\ j\theta_m\ \sin(j+1)\theta_m \end{aligned} \tag{2.204}$$

de sorte que :

$$\Re(z_m) = 0 \tag{2.205}$$

et finalement :

$$\delta\lambda_{h_m} = 2\,\varepsilon \tag{2.206}$$

comme prévu. □

Examinons maintenant le cas inverse où la diffusion domine : $\varepsilon \gg 1$. Dans ce cas, on développe (2.191) suivant les puissances croissantes du petit paramètre ε^{-1} :

$$\lambda_{h_m} = 2\,\varepsilon - 2\,\sqrt{\varepsilon^2 - \frac{1}{4}}\,\cos\,\theta_m = 2\,\varepsilon - 2\,\varepsilon\,\cos\,\theta_m + O\left(\frac{1}{\varepsilon}\right) \tag{2.207}$$

de sorte que

$$\frac{\lambda_{h_m}}{\varepsilon} = 2 - 2\,\cos\,\theta_m + O\left(\frac{1}{\varepsilon^2}\right) \tag{2.208}$$

Autrement dit, la perturbation causée par l'advection sur les valeurs propres (normalisées) est du second ordre. Vérifions-le à partir de (2.180). Pour cela, on part de

$$\frac{A_h}{\varepsilon} = A_0' + \frac{1}{\varepsilon}\,A_0 \tag{2.209}$$

Le rôle des matrices A_0 et A_0' est ici inversé par rapport au cas précédent. La partie principale A_0' est diagonalisée par la matrice bien connue des modes de Fourier discrets du modèle discret fondamental, S_h (cf. Théorème 1.1), de sorte qu'ici (2.180) s'applique en faisant les changements de notation suivants :

$$\varepsilon \to \frac{1}{\varepsilon}\,,\ A_0 \to A_0'\,,\ A_0' \to A_0\,,\ X_0 \to S_h \tag{2.210}$$

ce qui donne

$$\delta\Lambda = \frac{1}{\varepsilon}\,\mathrm{Dg}\,\left(S_h^{-1}\,A_0\,S_h\right) \tag{2.211}$$

Or la matrice S_h est orthogonale ($S_h^{-1} = S_h^T$) et la matrice A_0 antisymétrique, de sorte que la matrice $S_h^{-1}\,A_0\,S_h$ est antisymétrique et ses éléments diagonaux sont tous nuls. En conséquence :

$$\delta\Lambda = 0 \tag{2.212}$$

ce qui confirme le résultat. □

Chapitre 3

Relaxation et lissage

3.1. Introduction

Le but de ce chapitre est d'une part de rappeler quelques notions essentielles sur les techniques de résolution itératives des systèmes linéaires, dites techniques de relaxation, pour lesquelles le lecteur pourra se référer notamment à [92] pour une présentation approfondie, et d'autre part d'introduire le concept d'opérateur de lissage, qui résulte d'une utilisation particulière de ces techniques de relaxation aux systèmes linéaires issus de la discrétisation d'une EDP linéaire (souvent elliptique).

3.2. Définitions classiques des itérations de Jacobi et Gauss-Seidel

Dans cette section, on s'intéresse à la résolution itérative d'un système linéaire régulier d'abord général :

$$\boxed{A\,u = f} \tag{3.1}$$

dans lequel A est une matrice inversible de dimension $N \times N$ (det $A \neq 0$). Il est coutumier de distinguer dans la matrice A la diagonale D, la partie triangulaire inférieure $-B$ et la partie triangulaire supérieure $-C$:

$$\boxed{A = D - (B + C)} \tag{3.2}$$

Le système à résoudre peut donc aussi bien s'écrire :

$$D\,u = (B + C)u + f \tag{3.3}$$

Dans le cas particulier, mais fondamental, où la matrice A est réelle-symétrique définie positive, son inversibilité exige celle de la matrice diagonale D. En effet, si un élément diagonal de la matrice D était nul, disons $D_{j,j} = 0$, notant e_j le j-ième vecteur de la base canonique (dont la k-ième composante est égale à $\delta_{j,k}$), on aurait $e_j^T\, A\, e_j = 0$, ce qui serait en contradiction avec le fait supposé vrai que la forme quadratique $q(x) = x^T\, A\, x$ est non dégénérée.

Revenant au cas général d'une matrice A non nécessairement réelle-symétrique, on suppose néanmoins dans tout le chapitre que l'on a :

$$\det D \neq 0 \tag{3.4}$$

En pratique, la satisfaction de cette condition est souvent le résultat d'un réarrangement du système par permutation des équations et choix judicieux des « pivots ». Une méthode itérative vient alors naturellement à l'esprit pour résoudre (3.1) ou (3.3) :

Définition 3.1 (Itération de Jacobi)

A partir de l'itéré $u^n = \{u_j^n\}$, on construit l'itéré suivant, $u^{n+1} = \{u_j^{n+1}\}$, en faisant une boucle sur les équations, et en résolvant chacune par rapport à l'inconnue de même indice, soit :

Pour $j = 1, 2, ..., N$, on calcule :

$$u_j^{n+1} = \frac{1}{D_{j,j}} \Big(\sum_{k<j} B_{j,k}\, u_k^n + \sum_{k>j} C_{j,k}\, u_k^n + f \Big) . \tag{3.5}$$

Avant de s'interroger sur la convergence de cette itération, notons qu'elle équivaut, en notation matricielle, à poser :

$$\boxed{u^{n+1} = G_J\, u^n + D^{-1}\, f} \tag{3.6}$$

où l'on a posé :

$$\boxed{G_J = D^{-1}\,(B + C) = D^{-1}\,(D - A) = I - D^{-1}\, A} \tag{3.7}$$

Définition 3.2 (Diagonale dominance)

On dit que la matrice carrée A est à diagonale dominante (strictement), ssi :

$$\boxed{\forall\, j\ ,\ |A_{j,j}| > \sum_{k\ ,\ k \neq j} |A_{j,k}|} \tag{3.8}$$

On a alors le théorème très important suivant :

Théorème 3.1

Si la matrice A est à diagonale dominante (strictement), elle est inversible et l'itération de Jacobi converge.

DÉMONSTRATION : l'hypothèse implique que l'on a :

$$\| G_J \|_\infty < 1 \tag{3.9}$$

(norme infinie induite, voir annexe A). En conséquence, le rayon spectral associé à l'itération de Jacobi, satisfait les relations suivantes :

$$\boxed{\rho_J \stackrel{\text{déf}}{=} \rho\,(G_J) \leq \| G_J \|_\infty < 1} \tag{3.10}$$

ce qui est d'ailleurs évident d'après le théorème de Gershgorin : en effet, les disques de Gershgorin associés à la matrice G_J sont concentriques à l'origine et lorsque la matrice A est à diagonale dominante (strictement), ils sont tous de rayon inférieur à 1. Par conséquent, aucune valeur propre de la matrice G_J n'est égale à 1 ; donc aucune valeur propre de la matrice :

$$A = D\,(I - G_J) \tag{3.11}$$

n'est nulle, ce qui établit l'inversibilité. On note alors :

$$e^\infty \stackrel{\text{déf}}{=} A^{-1}\, f \tag{3.12}$$

la solution du système linéaire, et :

$$e^n \stackrel{\text{déf}}{=} u^n - u^\infty \tag{3.13}$$

le « vecteur erreur (itérative) » à l'itération courante n.

Dans ce chapitre consacré à la convergence itérative, on omet l'indice h de discrétisation spatiale sauf en cas de notation ambigüe.

On a alors la récurrence suivante :

$$e^{n+1} = G_J\, e^n \tag{3.14}$$

qui implique que :

$$e^n = G_J^n\, e^0 \tag{3.15}$$

dont on tire la majoration suivante :

$$\| e^n \|_\infty \leq \| G_J^n \|_\infty \| e^0 \|_\infty \leq (\| G_J \|_\infty)^n \| e^0 \|_\infty \tag{3.16}$$

En conséquence :

$$\lim_{n \to \infty} \| e^n \|_\infty = 0 \tag{3.17}$$

et

$$\lim_{n \to \infty} u^n = u^\infty \tag{3.18}$$

ce qui établit la convergence. □

Le théorème 3.1 donne une condition *suffisante* de convergence de l'itération de Jacobi. Il ne s'agit pas d'une condition nécessaire. En particulier, on dit que la matrice A est à diagonale dominante, en sous-entendant « au sens large », lorsque (3.8) est vraie pour au moins une ligne j et qu'il y a égalité pour les autres ; dans ce cas, certains résultats, tels que l'inversibilité du système, peuvent encore être démontrés à condition de supposer en outre que la matrice A est « irréductible ». Le lecteur souhaitant approfondir cette question pourra notamment consulter [92].

Un contre-exemple notoire est en effet fourni par le cas du modèle discret unidimensionnel pour lequel la matrice

$$A = A_h = \frac{1}{h^2}\, \mathrm{Trid}_{D\,D}\,(-1, 2, -1) \tag{3.19}$$

(de dimension $M \times M$, $M = N$) est associée aux matrices B, D et C suivantes :

$$\begin{cases} B = h^{-2}\, \mathrm{Trid}_{D\,D}\,(1, 0, 0) \\ D = h^{-2}\, \mathrm{Diag}\,(2) \\ C = h^{-2}\, \mathrm{Trid}_{D\,D}\,(0, 0, 1) \end{cases} \tag{3.20}$$

et n'est visiblement pas à diagonale dominante strictement, mais seulement au sens large. On reprend ici la preuve de convergence de l'itération de Jacobi à ce cas particulier par le calcul direct du rayon spectral.

Le nombre complexe g est une valeur propre de la matrice G_J ssi :

$$G_J\, u = g\, u \quad (u \neq 0) \tag{3.21}$$

ce qui équivaut à

$$(g\, D - B - C)\, u = 0 \quad (u \neq 0) \tag{3.22}$$

c'est-à-dire spécifiquement :

$$\text{Trid}_{D\,D}\,(-1, 2g, -1)\, u = 0 \quad (u \neq 0) \tag{3.23}$$

Cette condition équivaut à l'existence d'une valeur propre nulle pour la matrice $\text{Trid}_{DD}\,(-1, 2g, -1)$ dont le spectre $\{\,\lambda_m\,\}$ ($m = 1, 2, ..., M$) est donné par la formule générale (2.191) qui se réduit ici à :

$$\lambda_m = 2g - 2\cos\,\theta_m \tag{3.24}$$

où les θ_m sont les paramètres de fréquence habituels. Par conséquent, la condition

$$\lambda_m = 0 \tag{3.25}$$

fournit l'ensemble suivant des valeurs propres de la matrice G_J :

$$g_m = \cos\,\theta_m \quad (m = 1, 2, ..., M) \tag{3.26}$$

Le rayon spectral en résulte :

$$\rho_J = \cos\,\frac{\pi}{M+1} = \cos\,\pi h = 1 - \frac{\pi^2}{2}\,h^2 + ... \tag{3.27}$$

(Noter que cette expression est conforme au résultat de l'annexe D.)

Revenant au cas d'une matrice A générale, on remarque que dans l'application de (3.5) on n'utilise pas les estimations les plus récentes des inconnues. En effet, dans le cas où le balayage s'effectue par valeurs croissantes de l'indice j, les valeurs $\{\,u_k^{n+1}\,\}$ ($k < j$) sont déjà connues lors du calcul de $\{\,u_j^{n+1}\,\}$. Cette remarque nous amène à considérer l'algorithme suivant comme alternative :

Définition 3.3 (Itération de Gauss-Seidel)

A partir de l'itéré $u^n = \{u_j^n\}$, on construit l'itéré suivant, $u^{n+1} = \{u_j^{n+1}\}$, en faisant une boucle sur les équations, et en résolvant chacune par rapport à l'inconnue de même

indice, ceci en utilisant toujours les estimations les plus récentes, soit :

Pour $j = 1, 2, ..., N$, on calcule :

$$u_j^{n+1} = \frac{1}{D_{j,j}} \Big(\sum_{k<j} B_{j,k}\, u_k^{n+1} + \sum_{k>j} C_{j,k}\, u_k^n + f \Big) . \tag{3.28}$$

Cette définition équivaut à l'équation matricielle suivante :

$$D\, u^{n+1} = B\, u^{n+1} + C\, u^n + f \tag{3.29}$$

qui elle-même donne :

$$\boxed{u^{n+1} = G_{GS}\, u^n + (D - B)^{-1}\, f} \tag{3.30}$$

où l'on a posé :

$$\boxed{G_{GS} = (D - B)^{-1}\, C = (D - B)^{-1}\, (D - B - A) = I - (D - B)^{-1}\, A} \tag{3.31}$$

Intuitivement, par rapport à l'itération de Jacobi, dans l'itération de Gauss-Seidel on inverse « exactement » une plus grande partie de la matrice A, à savoir la partie triangulaire inférieure $D - B$ au lieu de la partie diagonale D uniquement. On en escompte donc une réduction du rayon spectral qui est effective dans de nombreux cas. Vérifions-le dans le cas particulier du modèle discret unidimensionnel (1.10) en identifiant le spectre de la matrice G_{GS}.

Le nombre complexe g' est une valeur propre de la matrice G_{GS} ssi :

$$G_{GS}\, u' = g'\, u' \quad (u' \neq 0) \tag{3.32}$$

ce qui équivaut à :

$$\Big(g'\, (D - B) - C \Big)\, u' = 0 \quad (u' \neq 0) \tag{3.33}$$

c'est-à-dire spécifiquement :

$$\text{Trid}_{DD}\, (-g', 2g', -1)\, u' = 0 \quad (u' \neq 0) \tag{3.34}$$

Cette condition équivaut à l'existence d'une valeur propre nulle pour la matrice $\text{Trid}_{DD}\, (-g', 2g', -1)$ dont le spectre $\{\, \lambda'_m \,\}$ ($m = 1, 2, ..., M$) est donné par la formule générale (2.191) qui se réduit ici à :

$$\lambda'_m = 2g' - 2\sqrt{g'}\, \cos\, \theta_m \tag{3.35}$$

où les θ_m sont les paramètres de fréquence habituels. Par conséquent, la condition

$$\lambda'_m = 0 \tag{3.36}$$

fournit l'ensemble suivant des valeurs propres de la matrice G_{GS} :

$$g'_m = \cos^2 \theta_m \quad (m = 1, 2, ..., M) \tag{3.37}$$

Le rayon spectral en résulte :

$$\rho_{GS} = \cos^2 \frac{\pi}{M+1} = \cos^2 \pi h = \rho_J^2 \tag{3.38}$$

En conséquence, dans le cas particulier du modèle fondamental, l'itération de Gauss-Seidel converge 2 fois plus vite que celle de Jacobi. Ce résultat s'étend facilement au cas du laplacien discrétisé sur un maillage tensoriel en plusieurs dimensions d'espace pour lequel il est commode d'utiliser les notations de « somme directe » du paragraphe 3.5.2.

Il existe une autre situation générale pour laquelle l'algorithme de Gauss-Seidel est deux fois plus rapide que celui de Jacobi. Cette situation correspond au cas où ces algorithmes sont appliqués « par blocs » après avoir partitionné les inconnues en deux sous-ensembles, disons :

$$u = \begin{pmatrix} u_I \\ u_{II} \end{pmatrix} \quad u_I \in \mathbb{R}^p \,,\; u_{II} \in \mathbb{R}^q \,,\; p+q = N \tag{3.39}$$

Autrement dit :

$$u_I = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix} \quad u_{II} = \begin{pmatrix} u_{p+1} \\ u_{p+2} \\ \vdots \\ u_N \end{pmatrix} \tag{3.40}$$

En conséquence, on partitionne la matrice A et le second membre f comme suit :

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad f = \begin{pmatrix} f_I \\ f_{II} \end{pmatrix} \tag{3.41}$$

où les blocs diagonaux A_{11} et A_{22} sont carrés, inversibles et de dimensions respectives $p \times p$ et $q \times q$, alors que les blocs extra-diagonaux sont rectangulaires et de dimensions respectives $p \times q$ et $q \times p$. On introduit alors la définition suivante :

Définition 3.4 (Itérations de Jacobi ($\nu = 0$) et de Gauss-Seidel ($\nu = 1$) par blocs)

On appelle ainsi les itérations dont le schéma fonctionnel :

$$u^n = \begin{pmatrix} u_I \\ u_{II} \end{pmatrix}^n \quad \longrightarrow \quad u^{n+1} = \begin{pmatrix} u_I \\ u_{II} \end{pmatrix}^{n+1} \tag{3.42}$$

est défini comme suit :

$$\left\{ \begin{array}{l} u_I^{n+1} = A_{11}^{-1} \, (f_I - A_{12} \, u_{II}^n) \\ u_{II}^{n+1} = A_{22}^{-1} \, \left(f_{II} - A_{21} \, u_I^{n+\nu} \right) \end{array} \right. \tag{3.43}$$

Bien évidemment, dans le cas particulier d'un système de 2 équations à 2 inconnues ($p = q = 1$), cette définition redonne les définitions classiques des algorithmes.

Il résulte de cette définition que la matrice G_J de l'itération de Jacobi ($\nu = 0$) s'exprime comme suit :

$$G_J = - \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}^{-1} \begin{pmatrix} 0 & A_{12} \\ A_{21} & 0 \end{pmatrix} \tag{3.44}$$

alors que celle associée à l'itération de Gauss-Seidel ($\nu = 1$) est donnée par :

$$G_{GS} = - \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}^{-1} \begin{pmatrix} 0 & A_{12} \\ 0 & 0 \end{pmatrix} \tag{3.45}$$

Soit alors μ une valeur propre non nulle quelconque de la matrice G_{GS} et

$$v = \begin{pmatrix} v_I \\ v_{II} \end{pmatrix} \tag{3.46}$$

un vecteur propre associé non nul. Du fait que $\mu \neq 0$, on a :

$$\left\{ \begin{array}{l} \mu \, A_{11} \, v_I + A_{12} \, v_{II} = 0 \\ A_{21} \, v_I + A_{22} \, v_{II} = 0 \end{array} \right. \tag{3.47}$$

Soit alors λ l'une des racines carrées de μ :

$$\lambda = \pm \sqrt{\mu} \tag{3.48}$$

et u le vecteur non nul défini par :

$$u_I = v_I \quad u_{II} = \frac{v_{II}}{\lambda} \tag{3.49}$$

En reportant ces relations dans (3.47), il vient après simplification par le nombre non nul λ :

$$\left\{ \begin{array}{l} \lambda \, A_{11} \, u_I + A_{12} \, u_{II} = 0 \\ A_{21} \, u_I + \lambda \, A_{22} \, u_{II} = 0 \end{array} \right. \tag{3.50}$$

Enfin, puisque le vecteur u n'est pas nul, ce système prouve que λ est une valeur propre de la matrice G_J. En conclusion, à toute valeur propre $\mu \neq 0$ de l'itération de Gauss-Seidel correspond les 2 valeurs propres $\pm \sqrt{\mu}$ de l'itération de Jacobi. En

conséquence, les rayons spectraux de ces itérations satisfont dans le cas particulier considéré la relation suivante :

$$\rho_{GS} = \rho_J^2 \tag{3.51}$$

□

Nous venons d'examiner deux cas particuliers indépendants correspondant au laplacien discret d'une part et aux systèmes dont on a partionné les inconnues en deux sous-ensembles d'autre part. Dans ces deux cas, les vitesses de convergence respectives des itérations de Jacobi et de Gauss-Seidel sont dans un rapport simple, $\frac{1}{2}$. Par ailleurs, on ne connaît classiquement aucune relation quantitative de la sorte valable dans le cas général. Cependant, lorsque les coefficients de la matrice de Jacobi G_J sont des réels non négatifs, les itérations de Jacobi et de Gauss-Seidel sont toutes deux convergentes, ou toutes deux divergentes ; dans le cas où $0 < \rho_J < 1$, on a de plus :

$$0 < \rho_{GS} < \rho_J < 1 \tag{3.52}$$

ce qui implique la plus grande efficacité de l'itération de Gauss-Seidel (voir théorème 3.3 [92]).

3.3. Redéfinition et généralisation de l'itération de Jacobi

Lorsque la partie diagonale D de la matrice A est inversible, il n'est pas très restrictif de supposer que les équations sont préalablement normalisées par application de la matrice D^{-1}. De manière équivalente, on choisit plutôt de redéfinir la matrice A de telle sorte que ses éléments diagonaux soient égaux à une constante ($= 2$ dans le cas du modèle fondamental). Dans ce cas, l'écriture de l'itération de Jacobi se simplifie comme suit :

$$\boxed{u^{n+1} = u^n - \tau\ (A\,u^n - f)} \tag{3.53}$$

où τ est un paramètre réel. On généralise alors quelque peu la définition de l'itération en s'autorisant à régler ce paramètre en fonction de critère à définir.

3.4. Annihilation de modes propres et lissage

Cette section est pour l'essentiel tirée de la publication [37] à laquelle on renvoie pour des exemples d'application spécifiques.

3.4.1. Généralités

On considère une itération linéaire de $\mathbb{R}^M$ dans $\mathbb{R}^M$ définie par la récurrence suivante :

$$\boxed{u^{n+1} = \mathbf{g}(u^n) = G\,u^n + b} \tag{3.54}$$

Dans cette équation, u^n est le n-ème itéré (un vecteur de $\mathbb{R}^M$), b un vecteur donné (de $\mathbb{R}^M$), et G une matrice $M \times M$ diagonalisable :

$$G = T\,\Gamma\,T^{-1} \tag{3.55}$$

où T est une matrice inversible et Γ une matrice diagonale :

$$\Gamma = \begin{pmatrix} g_1 & & & \\ & g_2 & & \\ & & \ddots & \\ & & & g_M \end{pmatrix} \tag{3.56}$$

On fait l'hypothèse que les valeurs propres $\{g_m\}$ de l'itération qui constituent le spectre des matrices semblables G et Γ, sont telles que la condition classique de convergence [92] [85] portant sur le rayon spectral ρ est satisfaite :

$$\boxed{\rho \stackrel{\text{déf}}{=} \rho(G) = \rho(\Gamma) = \max_m\,(|g_m|) < 1} \tag{3.57}$$

Notons que cette condition peut être interprétée comme la particularisation au cas linéaire de la « condition de contraction » du théorème général du point fixe applicable à une itération non linéaire sur un espace de Banach (cf. e.g. [51], ou [77]).

Dans cette hypothèse, aucune des valeurs propres de G n'est égale à 1 et la matrice $I - G$ est inversible de sorte que (3.54) admet un point fixe unique :

$$u^\infty = (I - G)^{-1}\,b\,. \tag{3.58}$$

Il en résulte que le « vecteur erreur (itérative) » défini par :

$$e^n \stackrel{\text{déf}}{=} u^n - u^\infty\,, \tag{3.59}$$

vérifie l'équation récurrente linéaire homogène suivante :

$$e^{n+1} = G\,e^n\,. \tag{3.60}$$

En conséquence,

$$e^n = G^n\, e^0 = T\, \Gamma^n\, T^{-1}\, e^0\,, \tag{3.61}$$

et on pose :

$$\epsilon^n = T^{-1}\, e^n = \Gamma^n\, \epsilon^0\,. \tag{3.62}$$

Il vient :

$$\boxed{\begin{aligned} \epsilon^n &= \epsilon_1^0\, g_1^n\, \hat{I}_1 + \epsilon_2^0\, g_2^n\, \hat{I}_2 + \cdots + \epsilon_M^0\, g_M^n\, \hat{I}_M \\ e^n &= \epsilon_1^0\, g_1^n\, \hat{T}_1 + \epsilon_2^0\, g_2^n\, \hat{T}_2 + \cdots + \epsilon_M^0\, g_M^n\, \hat{T}_M \end{aligned}} \tag{3.63}$$

où $\hat{I}_1$, $\hat{I}_2$, ..., $\hat{I}_M$ sont les vecteurs colonnes de la matrice identité (c'est-à-dire la base canonique), et ϵ_1^0, ϵ_2^0, ..., ϵ_M^0 les composantes du vecteur ϵ^0 dans cette base, et $\hat{T}_1$, $\hat{T}_2$, ..., $\hat{T}_M$ sont les vecteurs colonnes de la matrice T, c'est-à-dire les vecteurs (ou modes) propres de la « matrice d'amplification » G. On voit donc qu'à un changement de base près, l'effet d'une itération est d'atténuer chaque composante de l'erreur d'un facteur égal au module de la valeur propre correspondante.

Ces relations permettent de donner un sens précis à la vitesse de convergence de l'algorithme itératif. Pour cela, on utilise des majorations par exemple en norme-p (voir annexe A pour les définitions de normes). La matrice Γ étant diagonale on a :

$$\|\Gamma^n\|_p = \rho^n\,, \tag{3.64}$$

où ρ désigne ici spécifiquement le rayon spectral de l'itération, c'est-à-dire celui des matrices G ou Γ. Alors la majoration suivante résulte directement de (3.61) :

$$\boxed{\|e^n\|_p \leq K \rho^n \|e^0\|_p} \tag{3.65}$$

où $K = \|T\|_p . \|T^{-1}\|_p$ est le nombre de conditionnement de la matrice des vecteurs propres. Supposons pour fixer les idées que :

$$\rho = |g_1| \geq |g_2| \geq \cdots \geq |g_M|\,, \tag{3.66}$$

et que $|\epsilon_1^0| \neq 0$. Alors, en vertu de (3.63), lorsque $n \to \infty$:

$$\|\epsilon^n\|_p \sim C_1\, \rho^n\,, \quad \|e^n\|_p \sim C_2\, \rho^n\,, \tag{3.67}$$

et les quantités $-\dfrac{\ln \|\epsilon^n\|_p}{n}$ et $-\dfrac{\ln \|e^n\|_p}{n}$ admettent une même limite finie,

$$\boxed{\upsilon = -\ln \rho} \tag{3.68}$$

appelée « vitesse de convergence asymptotique ». La quantité $1/\upsilon$ est alors une mesure du nombre moyen d'itérations nécessaires asymptotiquement à une réduction de la norme de l'erreur d'un facteur égal au nombre e. On dit que la convergence asymptotique de l'itération est *linéaire*. Il est d'usage de représenter la suite des valeurs de la quantité normalisée $\|e^n\|_p/\|e^0\|_p$ en échelle logarithmique en fonction de n ; dans cette représentation, la suite des points admet une droite de pente $-\upsilon$ comme asymptote.

Enfin, considérons un cas où la matrice G est proche d'une matrice défective, de sorte que la matrice des vecteurs propres est mal conditionnée, et $K \gg 1$. Alors, la majoration donnée en (3.65) suggère que la convergence est conforme au résultat asymptotique seulement pour n très grand. Un tel comportement peut se produire et on renvoie à [36] [38] pour une discussion détaillée du cas défectif et des illustrations numériques.

3.4.2. Analogie fondamentale, annihilation, sur-relaxation

On se place maintenant dans le cas où le spectre de G noté $\sigma(G)$, à savoir le nuage formé des valeurs $\{g_1,\ g_2,\ \ldots,\ g_M\}$ forme dans le plan complexe, un « agrégat » localisé et connu. Ceci signifie que l'on connaît un domaine du plan complexe, pas nécessairement convexe, qui contient ce nuage ; de plus, on suppose que ce nuage n'est pas diffus dans le disque de rayon unité ; à l'inverse, il occupe une situation particulière dans ce disque (voir figure 3.1).

On pose :

$$\boxed{A = I - G} \tag{3.69}$$

de sorte que l'itération (3.54) prend la forme suivante :

$$u^{n+1} = (I - A)\, u^n + b \tag{3.70}$$

que l'on identifie à un « pas en temps » de la méthode d'Euler appliquée au système d'équations différentielles ordinaires

$$\dot{u} = b - Au \tag{3.71}$$

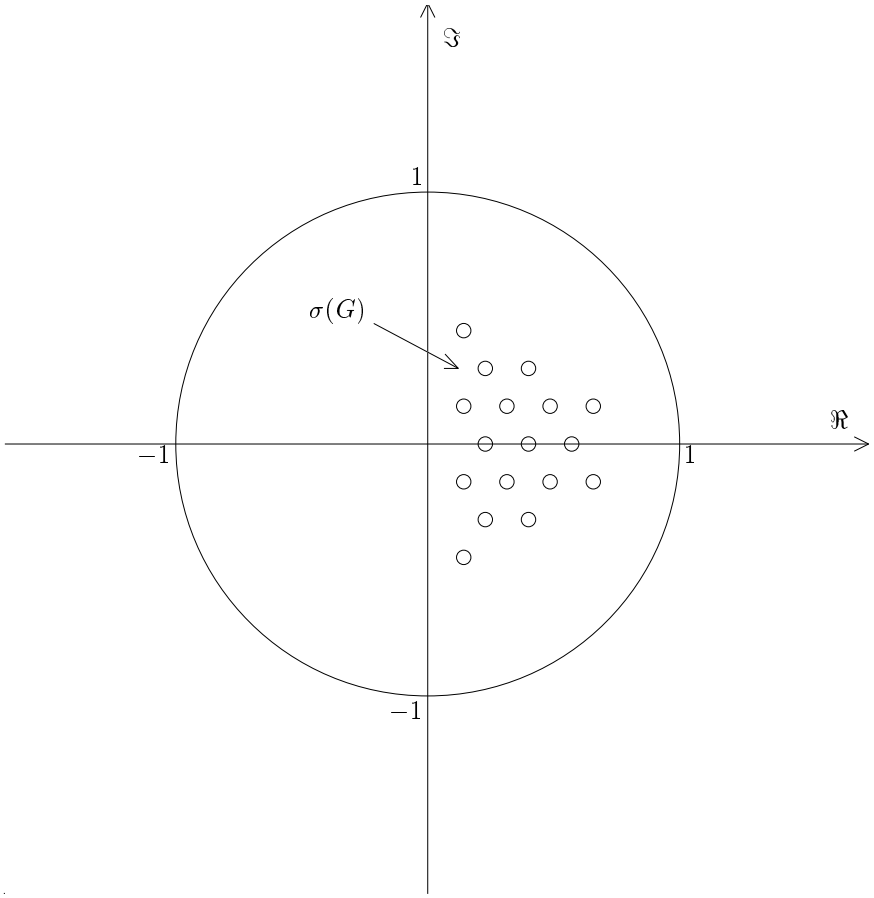


Figure 3.1. Spectre de G schématisé

avec :

$$\Delta t = 1\,. \tag{3.72}$$

Les valeurs propres de la matrice A sont les suivantes :

$$\lambda_m = 1 - g_m \tag{3.73}$$

et forment un spectre noté $\sigma(A)$, dont on sait par hypothèse qu'il constitue un agrégat contenu dans un domaine Ω à l'intérieur du disque centré en $z = 1$ et de rayon 1. Dans le cas où les matrices G et A sont réelles, ce spectre est de plus symétrique par rapport à l'axe des réels (voir figure 3.2).

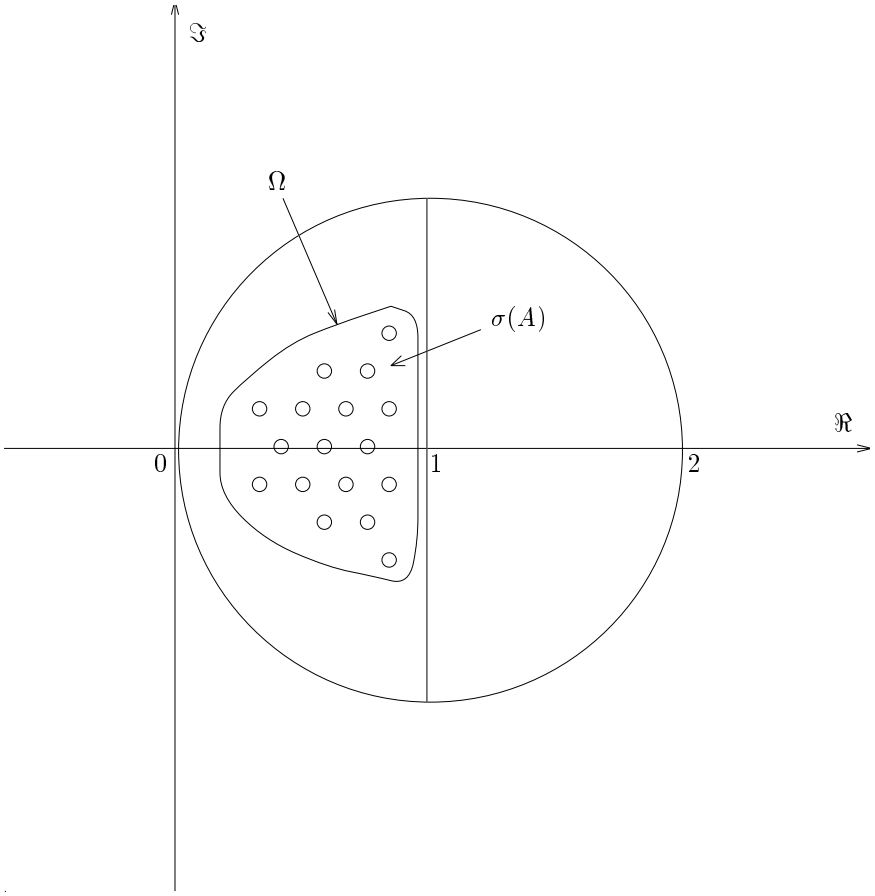


Figure 3.2. Spectre de A schématisé

Toutes ces valeurs propres sont donc à partie réelle strictement positive, et par conséquent aucune n'est nulle, ce qui implique l'existence d'un point fixe unique :

$$u^\infty = A^{-1}\, b = (I - G)^{-1}\, b \tag{3.74}$$

qui est le point fixe de l'itération de départ ; de plus, la solution exacte du système instationnaire

$$u(t) = u^\infty + (u(0) - u^\infty)\, \exp(-At) \tag{3.75}$$

tend vers u^∞ quand $t \to \infty$. Dans les applications aux EDP, A est la « matrice d'approximation » (préconditionnée), c'est-à-dire la matrice de représentation d'un opéra-

teur discret stationnaire.

On vient donc d'établir l'analogie entre une itération (linéaire, convergente) quelconque et l'intégration par la méthode d'Euler d'un système d'équations différentielles ordinaires associé, dont la solution admet une limite, quand $t \to \infty$, dite solution stationnaire, identique au point fixe de l'itération.

REMARQUE FONDAMENTALE :

Soit τ un nombre non nul, réel pour l'instant ; si $1/\tau \in \sigma(A)$, alors un seul pas d'intégration par la méthode d'Euler avec $\Delta t = \tau$ suffit à éliminer la composante du vecteur erreur dans sa décomposition suivant les vecteurs propres de la matrice A. On dit qu'il y a « annihilation » du mode propre correspondant [65]. Cette propriété résulte directement de l'expression du $(n+1)$-ième itéré de l'erreur (dans la base des vecteurs propres), ϵ^{n+1}, en fonction des composantes ϵ_m^n du précédent :

$$\epsilon^{n+1} = \sum_{m=1}^{M} (1 - \lambda_m \tau)\ \epsilon_m^n\ \hat{I}_m\,. \tag{3.76}$$

Inversement, on peut vouloir chercher à « annihiler » la composante de l'erreur dans la direction du vecteur propre associé à la valeur propre λ_m en réglant le pas de temps comme suit :

$$\Delta t = \tau = \frac{1}{\lambda_m}\,. \tag{3.77}$$

Lorsque $\lambda_m \in \mathbb{R}$ la réalisation de l'algorithme est évidente :

$$u^{n+1} = (I - \tau A)u^n + \tau b \tag{3.78}$$

ce qui s'interprète comme un schéma de sur(sous)-relaxation appliquée à l'itération de base :

$$\boxed{\begin{array}{c} v^{n+1} = (I - A)u^n + b = \mathrm{g}(u^n) \\ u^{n+1} = (1 - \omega)\ u^n + \omega\ v^{n+1} \end{array}} \tag{3.79}$$

où l'on a posé $\omega = \tau$, pour se conformer à une notation usuelle.

REMARQUE : cette interprétation permet d'étendre ces notions au cas non linéaire, en identifiant la matrice G au Jacobien de la fonction g (voir figure 3.3).

Revenons maintenant au cas général où λ_m est complexe dans (3.77). Si les matrices G et A sont elles-mêmes complexes, l'utilisation d'un pas de temps complexe

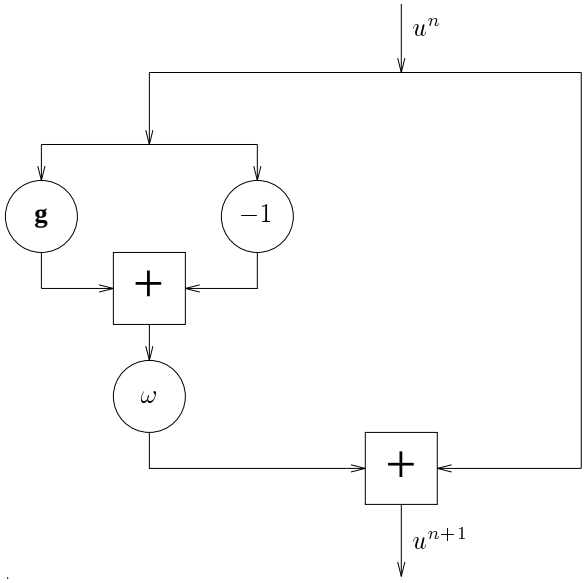


Figure 3.3. Schéma de sur(sous)-relaxation simple

ne pose aucune difficulté supplémentaire d'interprétation. Par contre, dans le cas de matrices réelles, les spectres sont symétriques par rapport à l'axe des réels et on effectue un cycle de deux pas de temps : le premier avec $\Delta t = 1/\lambda_m = \tau$, le second avec $\Delta t = \bar{\tau}$, ce qui donne :

$$\begin{aligned} u^{n+1} &= (I - \tau A)u^n + \tau b \\ u^{n+2} &= (I - \bar{\tau} A)u^{n+1} + \bar{\tau} b\,. \end{aligned} \tag{3.80}$$

Ce cycle est naturel car $\bar{\lambda}_m$ est aussi une valeur propre ; il équivaut à :

$$u^{n+2} = (I - \bar{\tau} A)(I - \tau A)u^n + b' \tag{3.81}$$

où le vecteur

$$\begin{aligned} b' &= (I - \bar{\tau} A)\tau b + \bar{\tau} b \\ &= 2\Re(\tau) b - |\tau|^2 A b \end{aligned} \tag{3.82}$$

est réel, et comme

$$\begin{aligned} (I - \bar{\tau} A)(I - \tau A) &= I - 2\Re(\tau) A + |\tau|^2 A^2 \\ &= I - 2\Re(\tau) A \left(I - \frac{|\tau|^2}{2\Re(\tau)} A \right)\,, \end{aligned} \tag{3.83}$$

le cycle se réalise en *arithmétique réelle* par la séquence « prédicteur-correcteur » suivante [49] :

Prédicteur :

$$v^{n+1} = u^n - \frac{|\tau|^2}{2\Re(\tau)} \left(A u^n - b \right) \ , \tag{3.84}$$

Correcteur :

$$u^{n+2} = u^n - 2\Re(\tau) \left(A v^{n+1} - b \right) \ . \tag{3.85}$$

Notons que la réalisation de ce cycle nécessite donc la mise en mémoire d'un vecteur supplémentaire, v^{n+1}. Enfin posons :

$$\begin{aligned} \omega_1 &= \frac{|\tau|^2}{2\Re(\tau)} \\ \omega_2 &= 2\Re(\tau) \end{aligned} \tag{3.86}$$

et remplaçons la matrice A par $I - G$, afin d'obtenir la nouvelle formulation du schéma prédicteur-correcteur suivante :

Prédicteur :

$$\boxed{\begin{aligned} v^{n+1} &= u^n + \omega_1 \left[G\, u^n + b - u^n \right] \\ &= u^n + \omega_1 \left[\mathbf{g}\left(u^n \right) - u^n \right] \end{aligned}} \tag{3.87}$$

Correcteur :

$$\boxed{\begin{aligned} u^{n+2} &= u^n + \omega_2 \left[G\, v^{n+1} + b - v^{n+1} \right] \\ &= u^n + \omega_2 \left[\mathbf{g}\left(v^{n+1} \right) - v^{n+1} \right] \end{aligned}} \tag{3.88}$$

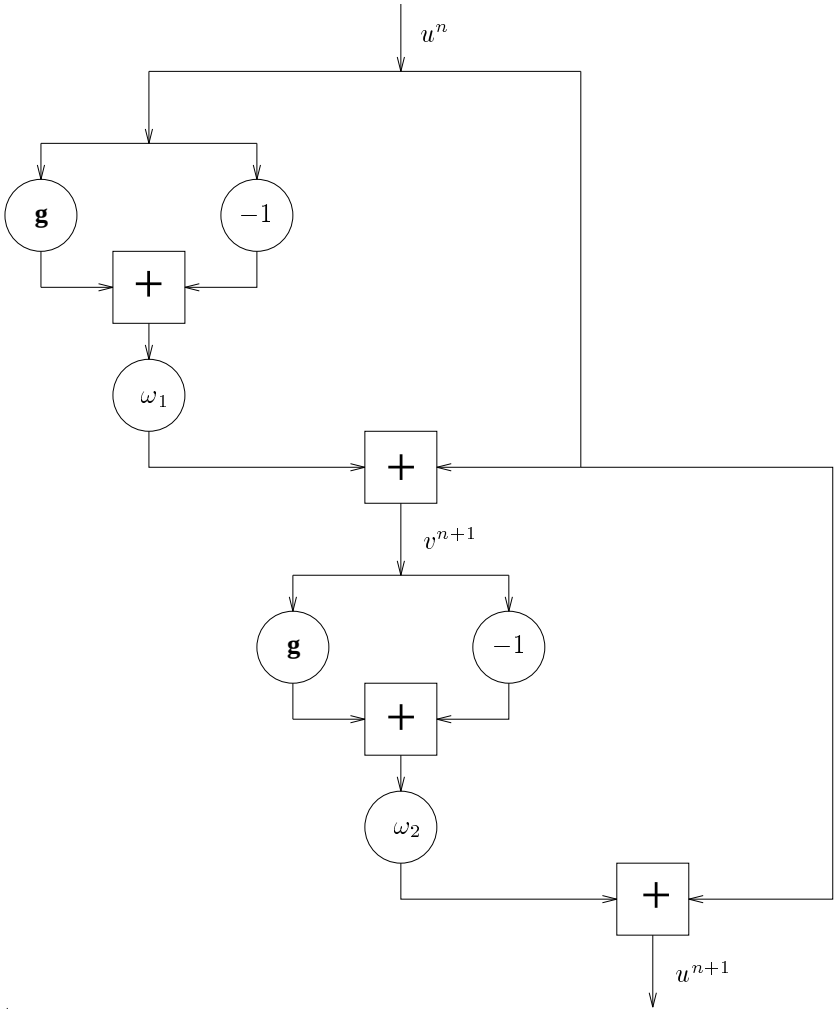


Figure 3.4. Schéma de sur(sous)-relaxation double

Sous cette forme, le cycle s'interprète comme deux étapes de sur(sous)-relaxation appliquées à un schéma prédicteur-correcteur, en boucle interne au prédicteur et externe au correcteur. A nouveau, l'extension au cas non linéaire ne pose aucune difficulté d'interprétation algorithmique (voir figure 3.4).

Notons, enfin que si $\tau = \tau_r + i\,\tau_i$, on a :

$$\left\{ \begin{array}{l} \omega_1 = \dfrac{\tau_r^2 + \tau_i^2}{2\tau_r}\,, \\ \\ \omega_2 = 2\tau_r\,, \end{array} \right. \tag{3.89}$$

ou inversement,

$$\left\{ \begin{array}{l} \tau_r = \dfrac{\omega_2}{2}\,, \\ \\ \tau_i = \pm\sqrt{\omega_2\left(\omega_1 - \dfrac{\omega_2}{4}\right)}\,. \end{array} \right. \tag{3.90}$$

Ces équations montrent que lorsque $(\tau_r, \tau_i) \in \mathbb{R}_+^* \times \mathbb{R}$, le couple (ω_1, ω_2) appartient au secteur angulaire de $\mathbb{R}_+^* \times \mathbb{R}_+^*$ correspondant à $\omega_2 \leq 4\,\omega_1$. Dans cette zone, ω_1 et ω_2 peuvent être supérieurs ou inférieurs à 1 indépendamment, donnant des cas de sur ou sous-relaxation respectivement (voir figure 3.5).

3.4.3. Le problème classique du min-max

Plus généralement, lorsqu'on effectue un cycle de pas de temps complexes $\Delta t = \tau_1,\ \tau_2,\ \ldots,\ \tau_k$, l'optimisation de ces paramètres est réalisée lorsque le rayon spectral du cycle est minimisé. Ceci conduit à poser le problème suivant :

$$\min_{\tau_1,\ \tau_2,\ \ldots\ \tau_k}\ \max_{\lambda\,\in\,\sigma(A)} \left| \prod_{j=1}^{k} (1 - \lambda\tau_j) \right|^{1/k}. \tag{3.91}$$

(L'exposant $1/k$ ne joue aucun rôle dans la détermination de l'optimum ; il permet seulement d'exprimer la valeur du min-max en termes de rayon spectral équivalent par évaluation de la fonction g.) Bien évidemment, la solution de ce problème est triviale lorsque $k \geq M$, puisqu'alors il suffit d'annihiler chaque mode l'un après l'autre pour atteindre un minimum absolu égal à zéro. Dans tout ce qui suit, on sous-entend que l'on se place dans le cas inverse, $k < M$, et même, en pratique, $k \ll M$. De plus, afin de travailler en continu, on peut être amené à remplacer le spectre discret $\sigma(A)$ par le domaine Ω qui l'englobe et formuler le problème comme suit :

$$\min_{\tau_1,\ \tau_2,\ \ldots\ \tau_k}\ \max_{\lambda\,\in\,\Omega} \left| \prod_{j=1}^{k} (1 - \lambda\tau_j) \right|^{1/k}. \tag{3.92}$$

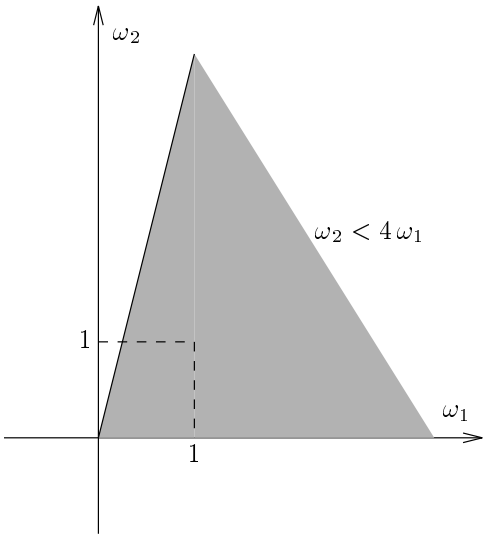


Figure 3.5. Domaine possible pour le couple (ω_1, ω_2)

Ce problème admet, suivant la forme du domaine Ω, quelques solutions exactes, et d'autres approchées, que nous allons maintenant examiner.

3.4.4. Solutions exactes connues et solutions approchées

Cas d'un spectre réel

Plaçons-nous dans le cas où $\Omega = [a, b]$, a et b étant deux réels positifs ($0 \leq a < b$). Il est évident que les valeurs optimales des paramètres τ_j sont alors réelles et strictement positives.

Considérons le polynôme en λ suivant :

$$P(\lambda) = \prod_{j=1}^{k} (1 - \lambda\tau_j) \ . \tag{3.93}$$

Ce polynôme admet les propriétés suivantes : ses coefficients et ses zéros sont réels, son degré est égal à k exactement, sa valeur en $\lambda = 0$ est égale à 1. Plus généralement, soit $\mathcal{P}$ la classe de tous les polynômes en λ ayant précisément ces propriétés, de sorte que $P \in \mathcal{P}$. Réciproquement, tout polynôme de $\mathcal{P}$ peut se mettre sous la forme (3.93), de sorte que le problème de min-max de (3.92) est une optimisation dans $\mathcal{P}$ en entier ; clairement, il s'agit de la minimisation de la norme infinie sur $[a, b]$.

Le problème (3.92) équivaut donc à trouver dans $\mathcal{P}$ l'élément de plus petite norme :

$$\min_{P \in \mathcal{P}} \|P\|_{\infty/[a,b]} . \tag{3.94}$$

Soit $T_k(x)$ le k-ème polynôme de Tchebychev, à savoir le polynôme défini pour $x \in [-1,1]$ par

$$T_k(x) = \cos\,(k \,\mathrm{Arccos}\, x)\ . \tag{3.95}$$

Ce polynôme est à coefficients réels, de degré k exactement, et ses zéros sont les réels suivants :

$$\xi_j = \cos\left(\frac{(2j-1)\pi}{2k}\right) \quad (j = 1, 2, ..., k)\,, \tag{3.96}$$

dont on voit qu'ils appartiennent tous à l'intervalle ouvert]-1,1[. En conséquence, le nombre $(b+a)/(b-a)$ qui est supérieur ou égal à 1, n'appartient pas à cet intervalle et n'est donc pas un zéro de T_k. On pose :

$$\boxed{c = \frac{b+a}{b-a} = \cosh\sigma} \tag{3.97}$$

et

$$A_k = T_k(c)\,, \tag{3.98}$$

de sorte que $A_k \neq 0$. Il résulte directement de leur définition, que les polynômes de Tchebychev vérifient la relation de récurrence bien connue suivante :

$$\forall\, x \in \mathbb{R},\ \forall\, k \in \mathbb{N}^* \ :\ T_{k+1}(x) + T_{k-1}(x) = 2x\, T_k(x)\,. \tag{3.99}$$

En conséquence :

$$\forall\, k \in \mathbb{N}^* \ :\ A_{k+1} + A_{k-1} = 2c\, A_k\,, \tag{3.100}$$

et comme de plus,

$$A_0 = 1\,,\ A_1 = c\,, \tag{3.101}$$

on obtient facilement que :

$$\boxed{A_k = \cosh(k\,\sigma) = \cosh(k\ \cosh^{-1} c)} \tag{3.102}$$

Soit alors le polynôme :

$$\boxed{P^*(\lambda) = \frac{1}{A_k}\, T_k\left(\frac{b+a-2\lambda}{b-a}\right) = \frac{(-1)^k}{A_k}\, T_k\left(\frac{2\lambda-(b+a)}{b-a}\right)} \qquad (3.103)$$

où l'on a utilisé le fait que le polynôme $T_k(x)$ a la même parité que l'entier k. Le polynôme $P^*(\lambda)$ est à coefficients réels, de degré k exactement, ses k zéros sont les réels

$$\mu_j = \frac{b+a}{2} + \frac{b-a}{2}\,\xi_j \quad (j = 1, 2, ..., k)\,, \qquad (3.104)$$

et sa valeur à l'origine est égale à 1. P^* appartient donc à $\mathcal{P}$. Il reste à prouver que ce polynôme réalise l'optimum. A cette fin, on raisonne par l'absurde.

Supposons qu'il existe dans $\mathcal{P}$, un élément Q ayant une norme strictement inférieure à celle de P^*. Le polynôme T_k est de norme infinie sur [-1,1] égale à 1 et il atteint alternativement les valeurs extrêmes 1 et -1 aux points :

$$\eta_j = \cos\left(\frac{j\pi}{k}\right) \quad (j = 0, 1, 2, ..., k)\,, \qquad (3.105)$$

de sorte que :

$$-1 = \eta_k < \eta_{k-1} < ... < \eta_1 < \eta_0 = 1\,. \qquad (3.106)$$

En conséquence, quand λ croît de a à b, la variable

$$\frac{b+a-2\lambda}{b-a} \qquad (3.107)$$

décroît de $\eta_0 = 1$ à $\eta_k = -1$, et la valeur du polynôme $P^*(\lambda)$ passe de son maximum, $1/A_k$, à son minimum, $-1/A_k$, aux points

$$\nu_j = \frac{b+a}{2} + \frac{b-a}{2}\,\eta_j \quad (j = 0, 1, 2, ..., k)\,, \qquad (3.108)$$

avec (exactement) k alternances de signe. Soit alors le polynôme :

$$R(\lambda) = P^*(\lambda) - Q(\lambda)\,. \qquad (3.109)$$

On a, quel que soit $j = 0, 1, 2, ..., k$:

$$|P^*(\nu_j)| = \frac{1}{A_k} = \|P^*\|_{\infty/[a,b]}\,, \qquad (3.110)$$

et

$$|Q(\nu_j)| \leq \|Q\|_{\infty/[a,b]} < \|P^*\|_{\infty/[a,b]}\,, \tag{3.111}$$

la deuxième inégalité étant vraie par hypothèse. Il en résulte que

$$\text{signe }(R(\nu_j)) = \text{signe }(P^*(\nu_j)) \quad (j = 0, 1, 2, ..., k)\,. \tag{3.112}$$

Le polynôme R admet donc sur $]a, b[$, k alternances de signe et donc k zéros réels *strictement* positifs. En outre, 0 est aussi un zéro de ce polynôme car les polynômes P^* et Q ont la même valeur 1 à l'origine, par définition de l'ensemble $\mathcal{P}$. On en conclut à l'existence de $k+1$ zéros distincts. Cette conclusion est en contradiction avec le fait que le polynôme R est de degré au plus égal à k. La contradiction se lève en rejetant l'hypothèse selon laquelle il existe un polynôme Q de $\mathcal{P}$ dont la norme infinie sur $[a, b]$ est strictement inférieure à celle de P^*. □

Les paramètres optimaux $\{\tau_j\}$ solution du problème du min-max, (3.92), sont donc les inverses des zéros du polynôme optimal P^* :

$$\boxed{\tau_j = \frac{1}{\mu_j}} \tag{3.113}$$

Le cycle itératif qui en résulte est connu sous le nom de « Méthode itérative de Richardson » ou « accélération de Tchebychev » [92].

Enfin, on peut mesurer la vitesse de convergence de l'itération par la quantité :

$$\boxed{\upsilon = -\frac{1}{k}\ln\left(\|P^*\|_{\infty/[a,b]}\right) = \frac{1}{k}\ln\left(A_k\right) = \frac{1}{k}\ln\left[\cosh(k\,\sigma)\right]} \tag{3.114}$$

La quantité $1/\upsilon$ représente le nombre moyen d'itérations (c'est-à-dire d'évaluations de la fonction g) qu'il faut pour réduire le mode le plus lent à converger d'un facteur égal à e. D'ailleurs, on a l'équivalence :

$$A_k \sim \frac{\exp\left(k\,\sigma\right)}{2} \quad (k \to \infty)\,, \tag{3.115}$$

de sorte que :

$$\begin{aligned} \lim_{k\to\infty}\ \upsilon &= \sigma \\ &= \cosh^{-1} c \\ &= \ln\left(c + \sqrt{c^2-1}\right)\,. \end{aligned} \tag{3.116}$$

Par conséquent, lorsque que le nombre de pas de temps du cycle tend vers l'infini, l'itération est asymptotiquement équivalente à une itération linéaire dont le rayon spectral ρ^* serait le suivant :

$$\rho^* = \lim_{k \to \infty} (A_k)^{-\frac{1}{k}} . \tag{3.117}$$

On obtient donc l'expression suivante du « rayon spectral équivalent » de l'itération accélérée :

$$\boxed{\rho^* = \frac{1}{c + \sqrt{c^2 - 1}}} \tag{3.118}$$

Dans ce qui suit, on examine en détail l'application de cette méthode au modèle discret unidimensionnel, (1.10), pour lequel les valeurs propres sont bien connues :

$$\lambda_m = 2 - 2\cos\theta_m \quad (m = 1, 2, ..., M) , \tag{3.119}$$

où le paramètre de fréquence θ_m est donné par :

$$\theta_m = \frac{m\pi}{M+1} \quad (m = 1, 2, ..., M) . \tag{3.120}$$

Par conséquent,

$$\boxed{\forall\, m = 1, 2, ..., M \ : \ 0 \leq \lambda_m \leq 4} \tag{3.121}$$

En résumé, un choix spécifique des paramètres a et b a pour effet d'optimiser la vitesse à laquelle le sous-ensemble des modes propres associés aux valeurs propres $\lambda \in [a, b]$ sont atténués. Plusieurs choix de l'intervalle $[a, b] \subset [0, 4]$ sont donc possibles. Pour chacun, on calcule les nombres c par (3.97), A_k par (3.102), et υ par (3.114).

<u>1er essai :</u> $a = 0, b = 4$.

Alors quel que soit k, $c = 1 = \eta_0$, $A_k = 1$ et

$$\boxed{\upsilon = 0} \tag{3.122}$$

Ceci n'a rien d'étonnant puisqu'on a englobé le spectre de A dans un domaine qui contient la valeur limite $a = 0$, qui, si elle était vraiment une valeur propre, serait associée à un mode stationnaire, non réductible. Donc, quel que soit le cycle de

pas de temps que l'on pourrait choisir, le rayon spectral serait égal à 1 et la vitesse de convergence nulle. On aboutit à une conclusion bien connue : dans le cas discret, l'itération converge parce que la plus petite valeur propre n'est pas tout à fait nulle, et la convergence est d'autant plus lente que cette valeur est proche de 0. On doit donc refaire le calcul en utilisant les limites exactes du spectre discret.

2e essai : $a = \lambda_1 = 4\sin^2\left(\frac{\pi}{2(M+1)}\right) = O(\Delta x^2)\,,\ b = \lambda_M = 4\cos^2\left(\frac{\pi}{2(M+1)}\right) \approx 4\,.$

On a alors :

$$c = \frac{\kappa+1}{\kappa-1} = 1 + \frac{2}{\kappa} + O\left(\frac{1}{\kappa^2}\right) \tag{3.123}$$

où :

$$\kappa = \frac{\lambda_M}{\lambda_1} = \tan^{-2}\left(\frac{\pi}{2(M+1)}\right) = O(M^2) \gg 1\,, \tag{3.124}$$

est le nombre de conditionnement du système discret. Quelques calculs de développements limités permettent de tirer de (3.102) l'expression suivante :

$$A_k = 1 + \frac{2\,k^2}{\kappa} + O\left(\frac{1}{\kappa^2}\right)\,, \tag{3.125}$$

ce qui donne finalement d'après (3.114) :

$$\boxed{v = \frac{2\,k}{\kappa} + O\left(\frac{1}{\kappa^2}\right)} \tag{3.126}$$

Donc, en moyenne un cycle fini optimal de k pas de temps converge k fois plus vite que l'itération avec un seul pas de temps optimisé. Cependant, l'itération reste relativement peu efficace puisque sa vitesse de convergence est inversement proportionnelle au nombre de conditionnement κ.

3e essai : lissage.

La méthode de Richardson peut également être utilisée comme « lisseur » idéal dans le contexte d'une stratégie multigrille en elliptique. Dans ce cas, on se limite à viser la partie « haute fréquence » du spectre, en choisissant ici

$$a = 2\,,\ b = 4\,, \tag{3.127}$$

ce qui donne dans (3.97) :

$$c = 3\,. \tag{3.128}$$

D'où,

$$\sigma = \cosh^{-1}(3) = \ln\left(3 + \sqrt{8}\right), \tag{3.129}$$

$$A_k = \frac{(3+\sqrt{8})^k + (3-\sqrt{8})^k}{2}, \tag{3.130}$$

et enfin,

$$\boxed{\upsilon = \frac{1}{k}\,\ln\left(\frac{(3+\sqrt{8})^k + (3-\sqrt{8})^k}{2}\right)} \tag{3.131}$$

de sorte que

$$\lim_{k\to\infty}\ \upsilon = \ln\left(3+\sqrt{8}\right) \approx 1.256\,. \tag{3.132}$$

L'efficacité de l'itération pour atténuer les modes de haute fréquence est donc caractérisée par une vitesse de convergence υ qui ne dépend pas du nombre de conditionnement κ.

A titre d'illustration, on a consigné dans le tableau 3.1, l'expression du polynôme $T_k(x)$, et les valeurs de A_k et υ pour $k = 1$, 2, 3, 6 et ∞.

On constate en particulier, qu'un cycle de 3 pas de temps optimaux améliore la vitesse de convergence d'environ 40 % seulement (par rapport à l'utilisation d'un seul pas de temps optimisé pour $\lambda \in [2, 4]$), et qu'en utilisant un plus grand nombre de pas de temps, la vitesse de convergence ne peut être augmentée de plus de 25 % environ. Plutôt que prolonger le cycle, il est plus efficace de « transférer » le problème, comme on va maintenant l'expliquer sommairement.

La conclusion principale de ce qui précède est que la partie haute fréquence du spectre est uniformément atténuée par le cycle, d'un facteur proche de 0.01 (1/99 pour être précis) lorsque $k = 3$, indépendamment du conditionnement du problème initial. La solution peut alors être transférée sur une grille deux fois plus grossière, où la même technique peut être à nouveau utilisée pour atténuer au dessous du centième la partie haute fréquence du spectre associé à cette nouvelle grille, et ainsi de suite jusqu'au niveau le plus grossier où l'on résout exactement un problème censé être trivial. Les transferts inverses sont des prolongements. Intuitivement, ils introduisent des erreurs de type « haute fréquence », car ce sont des composantes dans les directions des vecteurs propres qui n'ont pas de représentation dans la grille immédiatement plus grossière. Ce type d'erreur est à nouveau facilement éliminé par la méthode de Richardson. On aboutit au concept classique du V-cycle (voir chapitre 5).

k	$T_k(x)$	A_k	v
1	x	3	$\ln 3 \approx 1.09$
2	$2x^2 - 1$	17	$(\ln 17)/2 \approx 1.42$
3	$4x^3 - 3x$	99	$(\ln 99)/3 \approx 1.53$
6	$32x^6 - 48x^4 + 18x^2 - 1$	19601	$(\ln 19601)/6 \approx 1.65$
∞		∞	≈ 1.76

Tableau 3.1. Vitesse de convergence de la méthode de Richardson appliquée seulement à la partie « haute fréquence » du spectre

Cas d'un spectre complexe

On ne connaît pas la solution générale du problème du min-max, (3.92), lorsque le domaine Ω est quelconque.

Cependant, considérons toutes les ellipses dont l'équation est de la forme :

$$\frac{(d-x)^2}{a^2} + \frac{y^2}{a^2 - c^2} = 1 \,. \tag{3.133}$$

Ces ellipses ont des axes parallèles aux axes de coordonnées. Définissons le domaine Ω comme l'adhérence de l'intérieur de la plus petite d'entres elles pour laquelle ce domaine contient le spectre de A. Alors, si Ω est strictement contenu dans le demi-plan de droite, la solution du problème (3.92) est connue grâce à un résultat dû à Manteufel [67].

Malheureusement, on verra que, dans les applications qui nous concernent, cette dernière condition n'est pas réalisée, rendant le résultat de Manteufel inutilisable. Pour cette raison, on renvoie à [67], ainsi qu'à [82] (chapitre 6) pour un approfondissement de cette question.

3.4.5. Conclusion : notion de lisseur

Afin d'illustrer l'effet de lissage, on effectue ici une série d'expériences numériques portant sur le modèle discret fondamental en fixant le nombre de degrés de liberté à 20. Sur toutes les figures de cette section, on représente le vecteur d'erreur

itérative, ce qui, dans le cas d'un problème linéaire constitue une information équivalente à celle de la solution elle-même.

A la figure 3.6, on a représenté une condition initiale correspondant à des valeurs nodales obtenues par tirage aléatoire. Cette condition initiale est représentée à gauche par ses composantes naturelles nodales et à droite par ses composantes fréquentielles dans la base des modes de Fourier discrets.

Dans une première expérience, on applique la méthode de Richardson constituée d'un cycle de 3 itérations de Jacobi optimisé sur le spectre complet. Le résultat est indiqué à la figure 3.7. On constate une légère atténuation de toutes les composantes, mais aucune discrimination particulière entre les basses et les hautes fréquences.

Dans une deuxième expérience, figure 3.8, on applique le lisseur à la condition initiale de la figure 3.6, au lieu de la méthode itérative de base. La moitié des composantes fréquentielles, les hautes, subissent une réduction au dessous du $1/99^e$ de la valeur initiale et ont donc pratiquement disparu. En conséquence, le vecteur des composantes nodales, à l'inverse de la condition initiale aléatoire, devient une structure organisée, « lisse » : on y distingue le discrétisé d'une fonction régulière représentable sur une grille plus grossière ayant 2 fois moins de degrés de liberté.

A la figure 3.9, on a prolongé l'application du lisseur sur un très grand nombre d'itérations. Afin de pouvoir observer le phénomène, on a renormalisé l'erreur à chaque itération. Le lisseur agit comme un « filtre passe-bas ». Asymptotiquement, quelle que soit la condition initiale (ici une répartition aléatoire), on obtient uniquement le mode de plus basse fréquence.

3.5. Effet de la dimension d'espace sur la construction du lisseur

Nous venons de voir que pour le laplacien discret en une dimension d'espace, il était possible d'optimiser les 3 (pseudo-)pas de temps d'un cycle de Richardson afin que le taux d'atténuation des modes de hautes fréquences soit égal à $\rho_{HF} = 1/99$. On se pose ici la question suivante : de combien de (pseudo-)pas de temps optimisés faut-il composer le cycle de Richardson pour maintenir une performance au moins égale dans le cas de 2 ou 3 dimensions d'espace ?

On se place donc dans le cadre spécifique au lissage des résidus associés aux discrétisations classiques d'un laplacien en une ou plusieurs dimensions d'espace.

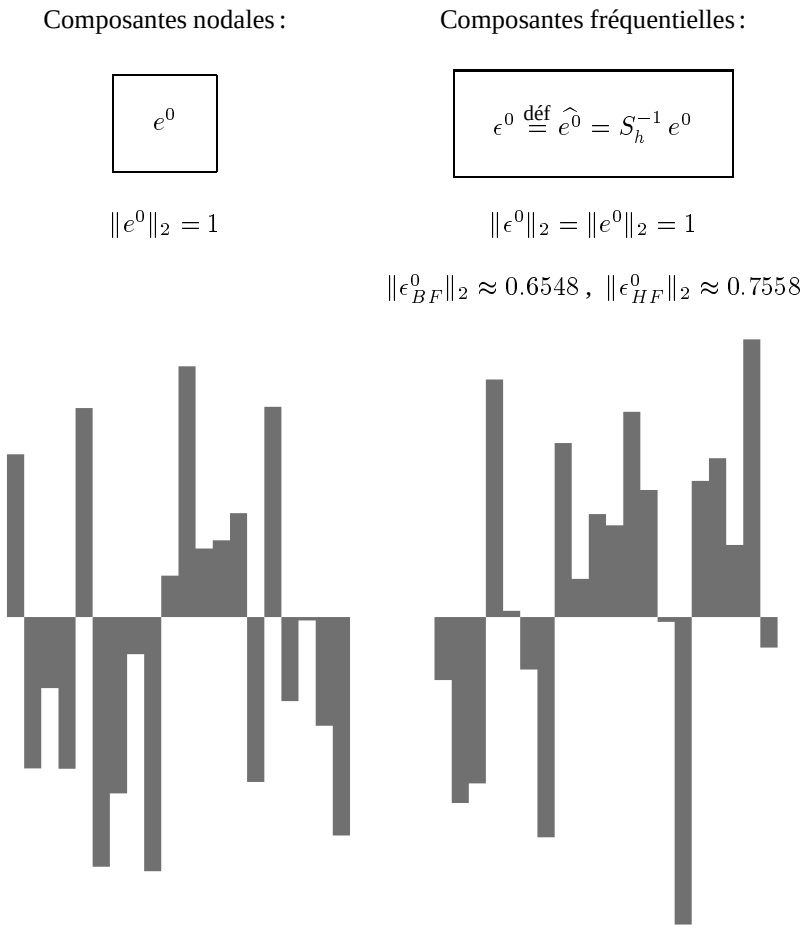


Figure 3.6. Condition initiale obtenue par tirage aléatoire des composantes nodales

Composantes nodales : Composantes fréquentielles :

e^3 $\epsilon^3 \stackrel{\text{déf}}{=} \widehat{e^3} = S_h^{-1}\, e^3$

$\|e^3\|_2 \approx 0.6945$ $\|\epsilon^3\|_2 = \|e^3\|_2 \approx 0.6945$

$\|\epsilon^3_{BF}\|_2 \approx 0.4391,\ \|\epsilon^3_{HF}\|_2 \approx 0.5381$

$$e^3 = \prod_{\ell=1}^{3} \left(Id - \tau_\ell\, h^2\, A_h\right) .\, e^0\,,\ \tau_\ell = 1/\lambda_\ell\ (\ell = 1,2,3)$$

$$\lambda_1 = 2 + 2(-\sqrt{3}/2)$$
$$\lambda_2 = 2 + 2(0)$$
$$\lambda_3 = 2 + 2(\sqrt{3}/2)$$

Figure 3.7. Méthode itérative de base : algorithme de Richardson optimisé au spectre complet

Composantes nodales :

$e^{3'}$

$\|e^{3'}\|_2 \approx 0.39917$

Composantes fréquentielles :

$\epsilon^{3'} \stackrel{\text{déf}}{=} \widehat{e^{3'}} = S_h^{-1}\, e^{3'}$

$\|\epsilon^{3'}\|_2 = \|e^{3'}\|_2 \approx 0.39917$

$\|{\epsilon^{3'}}_{BF}\|_2 \approx 0.39913,\ \|{\epsilon^{3'}}_{HF}\|_2 \approx 5 \times 10^{-3}$

$$e^{3'} = \prod_{\ell=1}^{3} \left(Id - \tau'_\ell\, h^2\, A_h\right)\, .\, e^0\,,\ \tau'_\ell = 1/\lambda'_\ell\ (\ell = 1,2,3)$$

$$\lambda'_1 = 3 + 1(-\sqrt{3}/2)$$
$$\lambda'_2 = 3 + 1(0)$$
$$\lambda'_3 = 3 + 1(\sqrt{3}/2)$$

Figure 3.8. Application du lisseur : algorithme de Richardson optimisé aux HF seulement

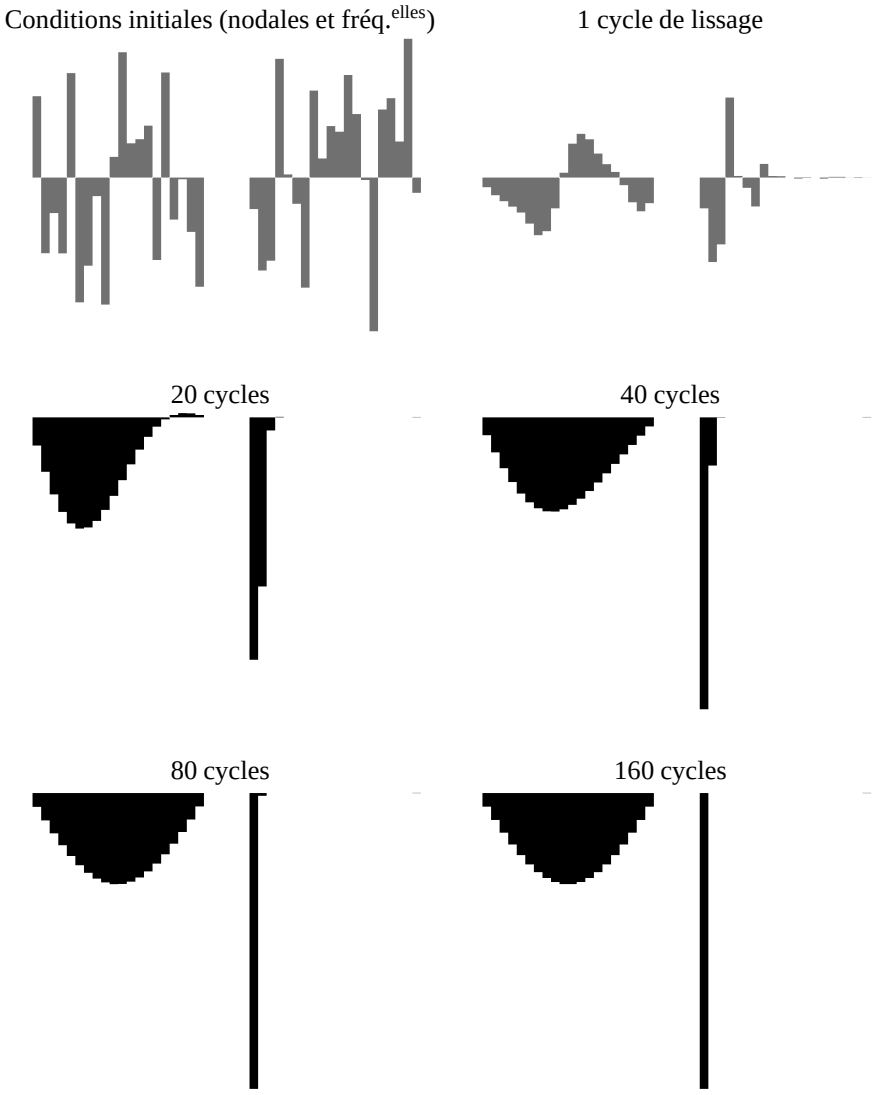


Figure 3.9. Le lisseur agissant comme un filtre passe-bas ; le vecteur erreur est renormalisé à chaque cycle à partir du second

Soit k le nombre d'itérations de Jacobi du cycle de Richardson. Le facteur d'atténuation d'un mode propre associé à la valeur propre λ est donné par (3.103) :

$$g(\lambda) = P^*(\lambda) = \frac{T_k\left(\dfrac{\dfrac{b+a}{2} - \lambda}{\dfrac{b-a}{2}}\right)}{T_k\left(\dfrac{b+a}{b-a}\right)} \tag{3.134}$$

Suivant la dimension d'espace, la définition des intervalles suivants varie :

– σ : intervalle englobant le spectre complet

– $\sigma_{HF} = [a, b] \subset \sigma$: intervalle englobant la partie du spectre identifiée comme correspondant aux modes de hautes fréquences.

On optimise le lisseur pour que l'atténuation des modes pour lesquels $\lambda \in [a, b]$ soit la plus efficace possible. Ces modes englobent les « hautes fréquences » mais peut-être aussi des basses. Une fois cette optimisation faite, on trace la courbe représentative de $g(\lambda)$, λ décrivant la totalité du spectre σ.

3.5.1. Cas d'une dimension d'espace

Au chapitre 2 (figures 2.4-2.5), on a introduit une séparation un peu arbitraire entre les modes de basses et de hautes fréquences. Anticipant sur la notion de « transfert de grille à grille » du chapitre 4, ici on souhaite d'abord redéfinir cette séparation fréquentielle en tenant compte du type de transferts que l'on souhaite appliquer aux fonctions discrètes.

Redéfinition des basses/hautes fréquences – Aliasing

On dispose de deux grilles : une fine, une grossière. On suppose que la grille grossière est obtenue en retenant un nœud sur deux de la grille fine. Toute autre hypothèse sur la relation entre ces grilles conduirait à une séparation différente entre les basses et les hautes fréquences. A une fonction discrète définie sur la grille fine, appliquons un « opérateur d'aller-retour », à savoir : on injecte cette fonction sur la grille grossière en éliminant une valeur sur deux, et on prolonge le résultat sur la grille fine en interpolant avec une grande précision. Idéalement, cette interpolation pourraît être réalisée par synthèse d'une série tronquée de Fourier basée sur les modes de Fourier discrets associés à la grille grossière. La structure très particulière des modes de Fourier du modèle fondamental définis par le théorème 1.1, fait que cet opérateur d'aller-retour laisse invariants les modes de basses fréquences de la figure 2.4 pour lesquels le para-

mètre de fréquence vérifie la condition :

$$\theta_m < \frac{\pi}{2} \tag{3.135}$$

A l'inverse, les modes de hautes fréquences de la figure 2.5, pour lesquels

$$\theta_m \geq \frac{\pi}{2} \tag{3.136}$$

sont corrompus par cet opérateur puisqu'ils sont transformés en basses fréquences, un phénomène connu sous le nom d'*aliasing*.

D'une manière générale, on définit les modes de basses fréquences comme ceux pour lesquels la perte d'information est faible par un double transfert de la grille fine à la grille grossière et retour, et les modes de hautes fréquences comme ceux qui à l'inverse subissent un phénomène d'*aliasing*.

Cette définition intuitive met en évidence que la séparation entre les modes de basses et de hautes fréquences dépend des choix que l'on fait pour construire les différents niveaux de grille ainsi que des opérateurs de transfert.

Lissage

Pour le modèle fondamental 1D, le spectre complet de la matrice d'approximation (normalisée) $h^2\, A_h$ occupe (dans la limite $h \to 0$) l'intervalle

$$\sigma^{1\,D} = [0, 4] \tag{3.137}$$

alors que les hautes fréquences « HF » en occupent seulement le sous-intervalle

$$\sigma_{H\,F}^{1\,D} = [2, 4] \stackrel{\text{déf}}{=} [a, b] \tag{3.138}$$

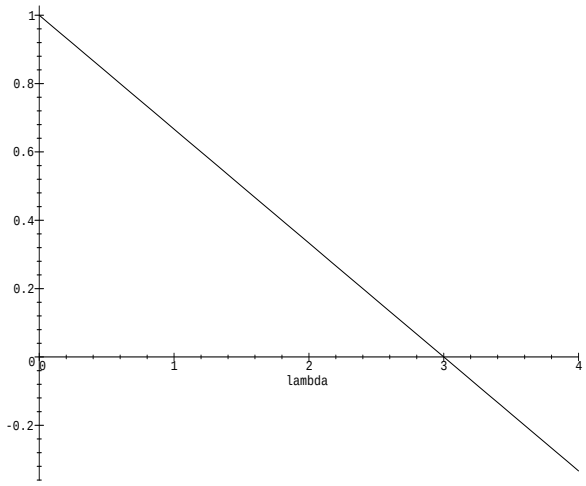
Cette identification de $[a, b]$ permet de compléter la définition d'un lisseur pour lequel la courbe représentative du facteur d'atténuation $g(\lambda)$ défini par (3.134) est fournie par la figure 3.10 pour $k = 1, 2, 3$.

3.5.2. Cas de 2 dimensions d'espace

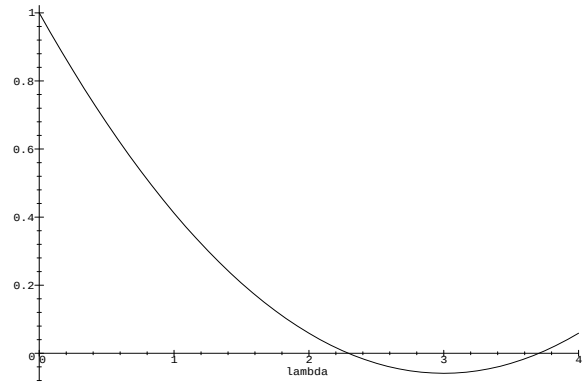
Dans ce cas, si le domaine est un rectangle uniformément discrétisé, la matrice d'approximation est une somme directe dont l'expression est la suivante en supposant un stockage « par colonne » :

$$\begin{aligned} A_h &= \text{« Penta »} \left(..., -\frac{1}{h_x^2}, ..., -\frac{1}{h_y^2}, \frac{2}{h_x^2} + \frac{2}{h_y^2}, -\frac{1}{h_y^2}, ..., -\frac{1}{h_x^2}, ... \right) \\ &= A_{h_x} \otimes I_L + I_M \otimes A_{h_y} \\ &= A_{h_x} \oplus A_{h_y} \end{aligned} \tag{3.139}$$

1 seul τ :
$\rho'_{HF} = \frac{1}{3} \approx 0.33$



2 τ's :
$\rho'_{HF} = \frac{1}{17} \approx 0.06$



3 τ's :
$\rho'_{HF} = \frac{1}{99} \approx 0.01$

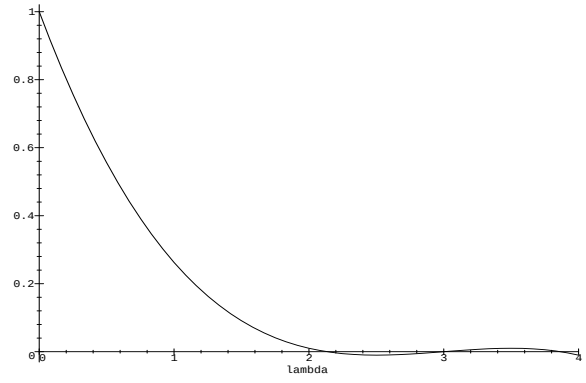


Figure 3.10. Facteur d'amplification $g(\lambda)$; cas d'une dimension d'espace ; on rappelle qu'en 1D le spectre complet correspond à $\lambda \in [0, 4]$, et l'enveloppe des HF à $\lambda \in [2, 4]$

où

$$A_{h_x} = \frac{1}{h_x^2}\,\text{Trid}_{DD\,M\times M}\,(-1,2,-1) \tag{3.140}$$

$$A_{h_y} = \frac{1}{h_y^2}\,\text{Trid}_{DD\,L\times L}\,(-1,2,-1) \tag{3.141}$$

(voir annexe B). En conséquence les modes propres ont la structure tensorielle suivante, analogue discret de la « séparation des variables » :

$$\begin{aligned}(s_{x\,y})_{j,k}^{(m,\ell)} &= (\psi_{x\,y})^{(m,\ell)}\,(x_j,y_k)\\ &= (\psi_x)^{(m)}\,(x_j)\times(\psi_y)^{(\ell)}\,(y_k)\\ &= \sqrt{\frac{2}{M+1}}\,\sin\,m\pi x_j\times\sqrt{\frac{2}{L+1}}\,\sin\,\ell\pi y_k\\ &= \frac{2}{\sqrt{(M+1)\,(L+1)}}\,\sin\,j\theta_{x\,m}\,\sin\,k\theta_{y_\ell}\end{aligned} \tag{3.142}$$

où l'on a supposé des conditions aux limites de Dirichlet pour lesquelles les paramètres de fréquence, au nombre de 2 en 2D, admettent les expressions suivantes :

$$\theta_{x\,m} = \frac{m\pi}{M+1} = m\pi h_x\,,\ \theta_{y_\ell} = \frac{\ell\pi}{L+1} = m\pi h_y \tag{3.143}$$

Les valeurs propres associées sont les nombres :

$$\begin{aligned}\underline{\lambda}_{m,\ell}^{h} &= \underline{\lambda}_m^{h_x} + \underline{\lambda}_\ell^{h_y}\\ &= \frac{2-2\cos\,\theta_{x\,m}}{h_x^2} + \frac{2-2\cos\,\theta_{y_\ell}}{h_y^2}\quad (m=1,2,...,M\,;\ \ell=1,2,...,L)\end{aligned} \tag{3.144}$$

Dans le cas d'un maillage cartésien tel que :

$$h_x = h_y = h \tag{3.145}$$

on peut alléger l'écriture en posant :

$$\lambda_{m,\ell}^{h} = h^2\,\underline{\lambda}_{m,\ell}^{h} \tag{3.146}$$

$$\lambda_m^{h_x} = h_x^2\,\underline{\lambda}_m^{h_x} = 2-2\cos\,\theta_{x\,m} \tag{3.147}$$

$$\lambda_\ell^{h_y} = h_y^2\,\underline{\lambda}_\ell^{h_y} = 2-2\cos\,\theta_{y_\ell} \tag{3.148}$$

de sorte que ces modes propres peuvent être identifiés dans un carré, produit cartésien des intervalles de variation des variables $\lambda_m^{h_x}$ et $\lambda_\ell^{h_y}$ (cf. figure 3.11).

A chaque point symbolique de ce carré correspond une valeur propre du système discret 2D égale à la somme suivante de contributions 1D,

$$\lambda^{h}_{m,\ell} = \lambda^{h_x}_{m} + \lambda^{h_y}_{\ell} \tag{3.149}$$

que l'on a portée sur un demi-axe de réels positifs à la figure 3.12.

Il résulte de ces considérations que l'intervalle englobant la totalité du spectre est

$$\sigma^{2D} = [0, 8] \tag{3.150}$$

La dimension d'espace affecte donc la répartition basses/hautes fréquences sur l'axe des λ de la figure 3.12. Lorsque la grille grossière est obtenue à partir de la grille fine en ne retenant qu'un point sur 2 dans chaque direction de coordonnées, le phénomène d'*aliasing* est évité par les modes qui sont le produit tensoriel de modes discrets 1D de basses fréquences suivant chaque direction. Avec ces hypothèses, on est donc amené à définir les modes de basses fréquences 2D comme ceux pour lesquels les 2 paramètres de fréquence remplissent les conditions suivantes :

$$\theta_{x\,m} < \frac{\pi}{2} \text{ et } \theta_{\mathrm{y}\,\ell} < \frac{\pi}{2} \tag{3.151}$$

A l'inverse, les modes de hautes fréquences sont tous ceux qui subissent un *aliasing* en x, ou en y, ou en x et en y simultanément, c'est-à-dire ceux pour lesquels

$$\theta_{x\,m} \geq \frac{\pi}{2} \text{ ou } \theta_{\mathrm{y}\,\ell} \geq \frac{\pi}{2} \tag{3.152}$$

ce qui ici équivaut à

$$\lambda^{h_x}_{m} \geq 2 \text{ ou } \lambda^{h_y}_{\ell} \geq 2 \tag{3.153}$$

En nombre, ces modes recouvrent asympotiquement ($h_x \to 0\,,\ h_y \to 0$) les 3/4 du carré de la figure 3.11. Les images de ces points sur l'axe de la figure 3.12 occupent le sous-intervalle suivant de σ^{2D} :

$$\sigma^{2D}_{HF} = [2, 8] \stackrel{\text{déf}}{=} [a, b] \tag{3.154}$$

On remarque que contrairement au cas d'une dimension d'espace, l'intervalle $[a, b]$ ici contient aussi des valeurs propres associées à des modes de basses fréquences. Cette identification de $[a, b]$ permet de compléter la définition d'un lisseur pour lequel la courbe représentative du facteur d'atténuation $g(\lambda)$ défini par (3.134) est fournie par la figure 3.13 pour $k = 3, 4, 5$.

En conclusion, on constate que l'augmentation de la dimension d'espace oblige à optimiser le lisseur sur une plus grande proportion du spectre complet (à savoir les 3/4 en 2D). La réalisation d'un lisseur dont le taux d'atténuation des hautes fréquences

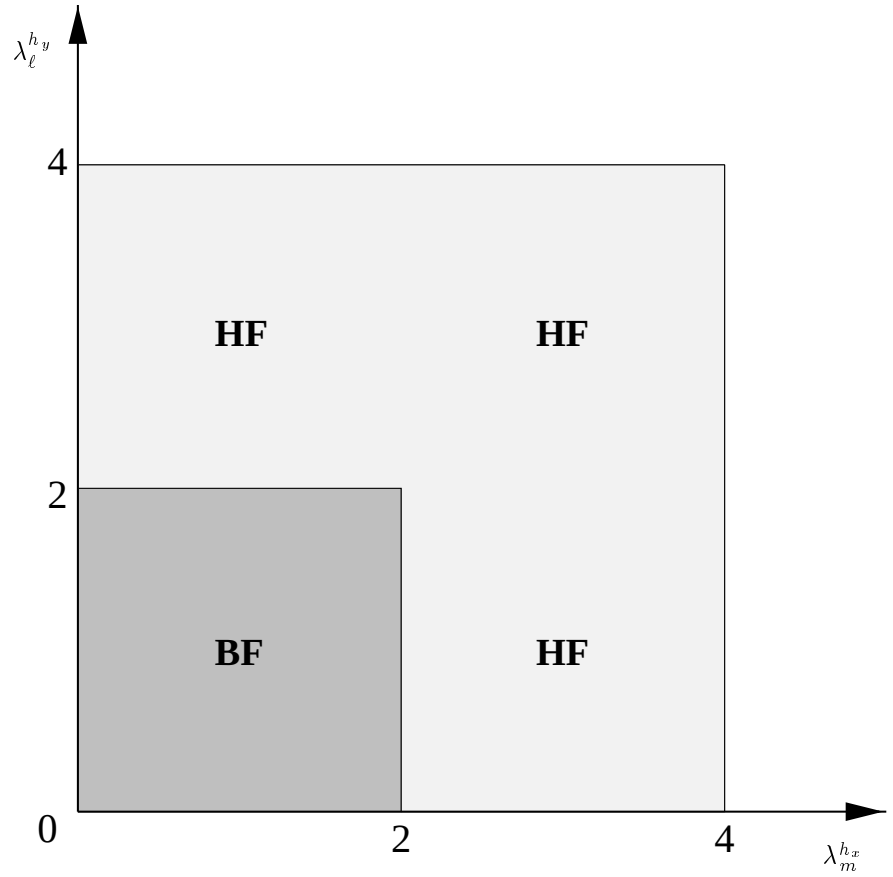


Figure 3.11. Domaine symbolisant les modes de Fourier 2D du laplacien discret dans le plan $\left(\lambda_m^{h_x}, \lambda_\ell^{h_y}\right)$

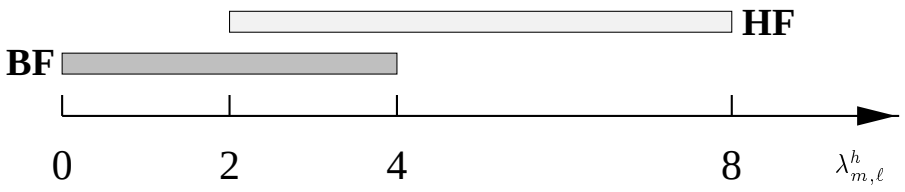
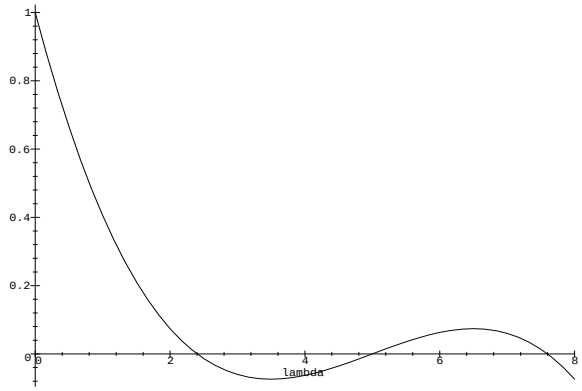
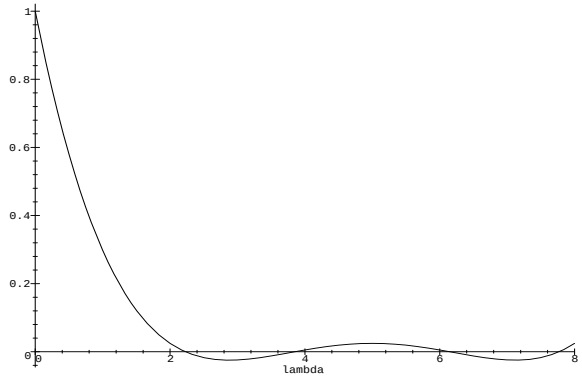


Figure 3.12. Spectre des valeurs propres du laplacien 2D discret $\lambda_{m,\ell}^h = \lambda_m^{h_x} + \lambda_\ell^{h_y}$

3 τ's :
$\rho'_{HF} = \frac{27}{365} \approx 0.07$



4 τ's :
$\rho'_{HF} = \frac{81}{3281} \approx 0.024$



5 τ's :
$\rho'_{HF} = \frac{243}{29525} \approx 0.008$

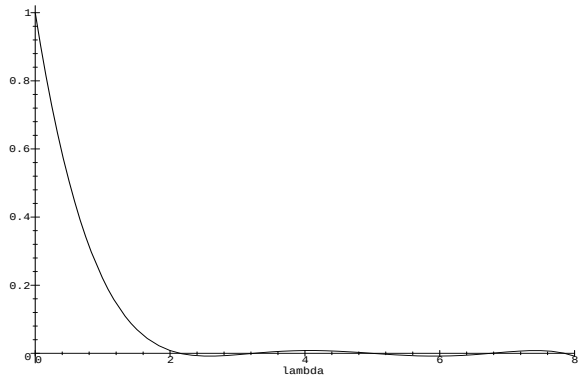


Figure 3.13. Facteur d'amplification $g(\lambda)$; cas de 2 dimensions d'espace (On rappelle qu'en 2D le spectre complet correspond à $\lambda \in [0, 8]$, et l'enveloppe des HF à $\lambda \in [2, 8]$)

est au plus égal à une constante donnée nécessite donc l'introduction d'un plus grand nombre de cycles de relaxation de Jacobi.

REMARQUE : effet d'anisotropie de maillage

Dans ce qui précède, on a traité le cas d'un laplacien discrétisé sur un maillage bidimensionnel tensoriel uniforme tel que $h_x = h_y$. On considère ici deux cas voisins qui en diffèrent par la modification d'un seul paramètre. Dans le premier cas, on suppose que le maillage est encore tensoriel et uniforme mais présente une forte anisotropie :

$$h_x \gg h_y \tag{3.155}$$

Une telle situation est typique du traitement d'une couche limite. Dans le deuxième, on garde le même maillage, mais on considère un opérateur elliptique autre que le laplacien dont les coefficients sont très différents l'un de l'autre :

$$\sigma_x \frac{\partial^2 u}{\partial x^2} + \sigma_y \frac{\partial^2 u}{\partial y^2} = f \ , \ \sigma_x \ll \sigma_y \tag{3.156}$$

Dans ces deux cas, les modes de Fourier, $s_{j,k}^{(m,\ell)}$, sont les mêmes que précédemment et l'identification des modes de hautes et de basses fréquences est inchangée. Par contre, l'expression des valeurs propres est ici la suivante :

$$\underline{\lambda}^h_{(m,\ell)} = \frac{\sigma_x}{h_x^2} \ (2 - 2\cos \ \alpha_m) + \frac{\sigma_y}{h_y^2} \ (2 - 2\cos \ \beta_\ell) \tag{3.157}$$

La proportion des valeurs propres associées aux modes de hautes fréquences par rapport au spectre complet est alors exprimée par le rapport :

$$\frac{\text{Enveloppe des HF}}{\text{Spectre complet}} = \frac{4 \left(\frac{\sigma_x}{h_x^2} + \frac{\sigma_y}{h_y^2} \right) - 2 \left(\frac{\sigma_x}{h_x^2} \right)}{4 \left(\frac{\sigma_x}{h_x^2} + \frac{\sigma_y}{h_y^2} \right)} = 1 - \frac{1}{2(1 + \mathcal{A})} \tag{3.158}$$

où :

$$\mathcal{A} = \frac{\sigma_y}{\sigma_x} \frac{h_x^2}{h_y^2} \gg 1 \tag{3.159}$$

est le « facteur de forme ». Par conséquent, dans le cas de mailles très étirées ($\mathcal{A} \gg 1$ ou $\ll 1$), il convient d'adapter l'optimisation du lisseur à la presque totalité du spectre complet.

Afin de mettre en évidence une manière d'adapter le lisseur à une situation de maillage étiré, examinons comment cette adaptation peut se traduire dans le cas du

lisseur constitué par la méthode de Jacobi avec un seul paramètre de relaxation τ (optimisé vis-à-vis de l'atténuation des hautes fréquences).

Dans le cas d'un maillage non étiré ($\mathcal{A} = 1$), d'après (3.158) les valeurs de λ correspondant aux hautes fréquences s'étalent sur les 3/4 du spectre complet, de sorte que la valeur optimale de τ qui d'une manière générale est égale à l'inverse de la moyenne arithmétique des bornes supérieure et inférieure de l'intervalle en λ que l'on souhaite atténuer optimalement, est ici égale à :

$$\tau^* = \left(\frac{\frac{\lambda_M}{4} + \lambda_M}{2} \right)^{-1} = \frac{8}{5\,\lambda_M} \tag{3.160}$$

alors que la valeur maximale de τ pour laquelle l'itération est stable est donnée par :

$$\tau_{\max} = \frac{2}{\lambda_M} \tag{3.161}$$

de sorte que

$$\frac{\tau^*}{\tau_{\max}} = \frac{4}{5} \tag{3.162}$$

Dans le cas inverse d'un maillage très étiré ($\mathcal{A} \gg 1$), les valeurs de λ correspondant aux hautes fréquences recouvrent la quasi-totalité du spectre, de sorte que

$$\tau^* \approx \left(\frac{0 + \lambda_M}{2} \right)^{-1} = \frac{2}{\lambda_M} \tag{3.163}$$

et

$$\frac{\tau^*}{\tau_{\max}} \approx 1 \tag{3.164}$$

Par conséquent, en présence de mailles étirées, on opérera l'algorithme pratiquement à la limite de stabilité, alors que dans le problème modèle isotrope, un meilleur lisseur est obtenu en réduisant le paramètre de relaxation τ (ici de 20 % par rapport à la limite de stabilité).

Une alternative intéressante d'adaptation du lisseur consiste à traiter « implicitement » la direction dans laquelle les mailles sont tassées. Dans ce cas, afin d'éviter une augmentation indue du coût d'application du lisseur, qui serait dommageable à la complexité d'un algorithme multigrille dont nous verrons que le lisseur est la base, une telle procédure d'implicitation pour être vraiment effective doit être utilisée seulement localement. Notons cependant que la conception d'une telle procédure implicite locale est délicate, particulièrement lorsqu'on utilise des maillages non structurés.

Enfin, anticipant sur le chapitre 5, notons qu'une alternative à l'adaptation du lisseur consiste à utiliser plusieurs niveaux de grille emboîtés de telle sorte que la procédure de « déraffinement » ou d'« appauvrissement » de maillage qui permet de construire un niveau de grille grossière, agisse localement de manière directionnelle, non isotrope (voir section 6.4.4.)

3.5.3. Cas de 3 dimensions d'espace

En développant un raisonnement analogue au cas 2D, on est amené à conclure qu'en 3D, si le maillage est cartésien, le spectre complet est englobé par l'intervalle

$$\sigma^{3D} = [0, 12] \tag{3.165}$$

alors que les valeurs propres associées aux modes de hautes fréquences sont englobées par le sous-intervalle

$$\sigma^{3D}_{HF} = [2, 12] \stackrel{\text{déf}}{=} [a, b] \tag{3.166}$$

Cette identification de $[a, b]$ permet de compléter la définition d'un lisseur pour lequel la courbe représentative du facteur d'atténuation $g(\lambda)$ défini par (3.134) est fournie par la figure 3.14 pour $k = 3, 5, 6$.

3.6. Illustration de convergences itératives

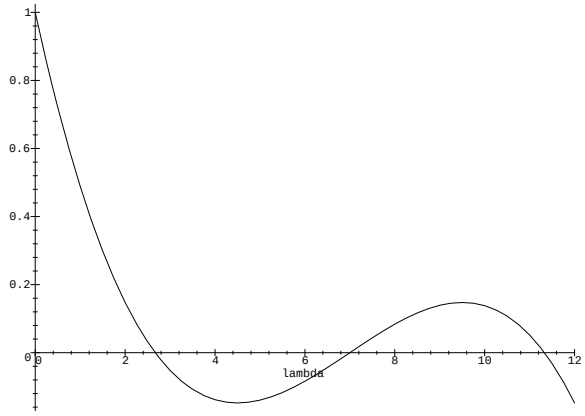
A la figure 3.15, on a représenté les « courbes » ou « historiques » de convergence de trois itérations démarrées à partir de la même condition initiale qui correspond aux données de la figure 3.6 dont on rappelle qu'elles ont été générées par tirage aléatoire des composantes nodales de l'erreur itérative suivi d'une normalisation du vecteur. On analyse la convergence itérative des trois algorithmes suivants :

a : Jacobi : 1 τ optimisé aux $BF \cup HF$
b : Richardson : 3 τ's optimisés aux $BF \cup HF$
c : Lissage : 3 τ's optimisés aux HF uniquement

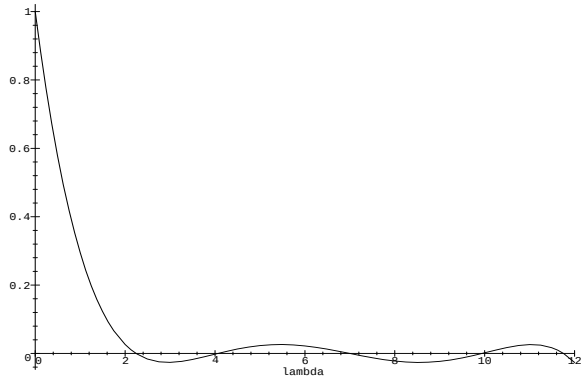
Dans chaque cas, on suit la décroissance de l'erreur itérative $\|e^n\|_2$ en fonction du nombre d'itérations n sur un diagramme en « échelle semi-logarithmique ». Pour accroître la lisibilité de la figure, les points d'une même suite sont reliés pour former une courbe. Sur un tel diagramme, une suite géométrique $\{r^n\}$ ($r \in \mathbb{R}+;\ n = 0, 1, ...$) serait représentée par des points alignés. L'examen de la figure permet de faire de nombreuses observations.

La première évidence est que les itérations a, b, et c sont toutes trois très rapides au démarrage (observer les zooms en particulier) ; l'itération c (le lisseur) est la plus performante des trois dans ces toutes premières itérations, puis la convergence ralentit dans les trois cas, et enfin chaque courbe approche une asymptote rectiligne propre ; dans cette phase asymptotique, la tendance initiale s'inverse de sorte que l'itération c

$3\ \tau$'s :
$\rho'_{HF} = \frac{125}{847} \approx 0.15$



$5\ \tau$'s :
$\rho'_{HF} = \frac{3125}{119287} \approx 0.026$



$6\ \tau$'s :
$\rho'_{HF} = \frac{15625}{1419190} \approx 0.011$

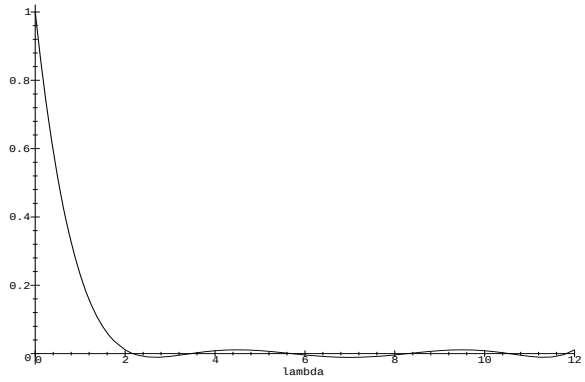
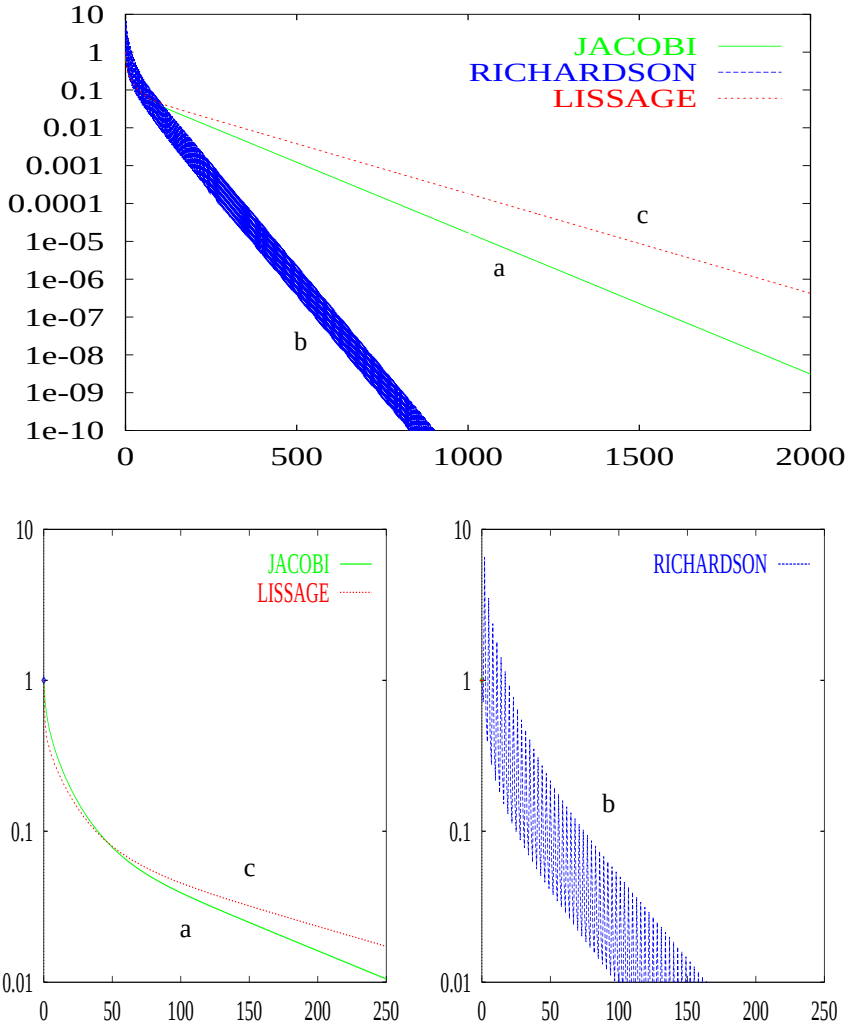


Figure 3.14. Facteur d'amplification $g(\lambda)$; cas de 3 dimensions d'espace ; on rappelle qu'en 3D le spectre complet correspond à $\lambda \in [0, 12]$, et l'enveloppe des HF à $\lambda \in [2, 12]$

$$\|e^n\|_2 = \|e_h^n\|_2 = fonct.(n) \quad (M = 20)$$



a : Jacobi : 1 τ optimisé aux $BF \cup HF$
b : Richardson : 3 τ's optimisés aux $BF \cup HF$
c : Lissage : 3 τ's optimisés aux HF uniquement

Figure 3.15. Convergences itératives de plusieurs méthodes appliquées au modèle discret fondamental 1.10

devient la moins performante. Pour expliquer ces observations, on note d'abord que le critère représenté,

$$\|e^n\|_2 = \sqrt{\sum_j \left(e_j^n\right)^2} = \|\epsilon^n\|_2 = \sqrt{\sum_m \left(\epsilon_m^n\right)^2} \tag{3.167}$$

(car la base des modes de Fourier, ici discrets, est orthonormale) est une combinaison de l'ensemble des composantes (nodales ou modales) et ne priviligie aucun mode fréquentiel particulier. Dans la condition initiale (voir figure 3.6) les composantes de hautes fréquences (HF) sont à peu près de même importance en ordre de grandeur que celles de basses fréquences (BF). Or, on sait que les modes de HF sont atténués beaucoup plus rapidement que les modes de BF (par l'une quelconque des trois itérations) et il en résulte dans les trois cas une réduction initialement très rapide de l'erreur. Le phénomène est le plus marqué dans le cas de l'itération c dont on a optimisé la performance vis-à-vis de l'atténuation des HF. Mais, après quelques itérations seulement (qui agissent comme un filtre « passe-bas ») les composantes de HF de l'erreur ont quasiment disparu et la convergence itérative désormais dictée par les composantes de BF, seules persistantes, ralentit notablement. A mesure que ce filtre agit, la courbe de convergence approche de plus en plus l'asymptote rectiligne qui correspond à l'évolution géométrique de la composante la plus lente à converger, c'est-à-dire celle associée à la fréquence la plus basse. Dans cette phase asymptotique, le lisseur (itération c) est la méthode itérative la moins bien adaptée, donc la moins efficace.

La deuxième évidence concerne les vitesses de convergence asymptotiques des itérations a et b. L'algorithme a qui résulte de l'optimisation d'un seul paramètre de relaxation a une convergence monotone. La convergence de l'algorithme b qui combine en un cycle 3 paramètres optimisés assez différents, n'est visiblement pas monotone. Bien que la courbe de convergence présente une oscillation, elle maintient en moyenne une certaine pente décroissante qui reflète la vitesse de convergence asymptotique moyenne de l'algorithme. Autrement dit, la courbe de convergence par cycle (de 3 itérations) est monotone. On observe sur la figure, que le niveau de l'erreur atteint après 500 itérations de l'algorithme b est à peu près équivalent à celui atteint après 1500 itérations de l'algorithme a. La vitesse de convergence asymptotique moyenne de l'algorithme b est donc 3 fois supérieure à celle de l'algorithme a, ce qui confirme l'estimation théorique (3.126).

Exercice 3.1 (Propriétés de lissage de l'itération de Gauss-Seidel en périodique)

On considère la résolution itérative du modèle discret unidimensionnel,

$$-u_{j-1} + 2\,u_j - u_{j+1} = h^2\,f_j \tag{3.168}$$

dans le cas périodique où l'analyse de Fourier s'applique.

(1) Déterminer l'expression du facteur d'atténuation de l'erreur $g_1(\theta)$ (et de son module) en fonction du paramètre de fréquence $\theta \in [-\pi, \pi]$.

(2) Peut-on améliorer les propriétés de lissage de l'itération par sur ou sous-relaxation ? (On notera $g_\omega(\theta)$ le facteur d'amplification.)

Exercice 3.2 (Itération de Gauss-Seidel pour le problème de Dirichlet)

Il s'agit d'étudier les performances de l'itération de Gauss-Seidel avec sur ou sous-relaxation lorsqu'on l'applique au modèle discret unidimensionnel fondamental (1.10) soumis à des conditions aux limites de Dirichlet. L'algorithme s'écrit comme suit :

$$\text{Pour } j = 1, 2, ..., M \text{ (dans cet ordre) :}$$
$$v_j^{n+1} = \tfrac{1}{2} v_{j-1}^{n+1} + \tfrac{1}{2} u_{j+1}^n + \tfrac{1}{2} h^2 f_j \tag{3.169}$$
$$u_j^{n+1} = u_j^n + \omega \left(v_j^{n+1} - u_j^n \right) \tag{3.170}$$

Identifier la valeur optimale ω^* du paramètre de relaxation ω et le rayon spectral correspondant ρ^*, et tracer la courbe du « facteur d'amplification » (ici d'atténuation) $g(\theta)$ en fonction du paramètre de fréquence $\theta \in [0, \pi]$ dans les deux cas suivants :

(a) atténuation de tous les modes fréquentiels,
(b) atténuation des modes de hautes fréquences uniquement (« lissage »).

Estimer le gain en vitesse de convergence réalisé par la méthode optimale ($\omega = \omega^*$) par rapport à l'itération de base ($\omega = 1$).

On posera

$$B = \text{Trid}_{DD}\left(\tfrac{1}{2}, 0, 0\right) \tag{3.171}$$
$$C = \text{Trid}_{DD}\left(0, 0, \tfrac{1}{2}\right) = B^T \tag{3.172}$$

ainsi que :

$$v^{n+1} = \begin{pmatrix} v_1^{n+1} \\ v_2^{n+1} \\ \vdots \\ v_M^{n+1} \end{pmatrix}, \quad u^n = \begin{pmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_M^n \end{pmatrix} \tag{3.173}$$

ce qui permettra de formuler l'itération comme suit :

$$v^{n+1} = B\, v^{n+1} + C\, u^n \tag{3.174}$$
$$u^{n+1} = u^n + \omega \left(v^{n+1} - u^n \right) \tag{3.175}$$

et on cherchera les modes propres,

$$\begin{pmatrix} u^n \\ v^n \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} \qquad \begin{pmatrix} u^{n+1} \\ v^{n+1} \end{pmatrix} = g \begin{pmatrix} u \\ v \end{pmatrix} \tag{3.176}$$

en se ramenant à une équation du type :

$$\text{Trid}_{D\,D}\left(a(g,\omega), b(g,\omega), c(g,\omega)\right) u = 0 \quad (u \neq 0) \tag{3.177}$$

Chapitre 4

Technique d'enrichissement progressif de maillage

4.1. La théorie

4.1.1. Maillages emboîtés, opérateurs de transfert

On considère deux discrétisations uniformes de l'intervalle d'étude, $\mathcal{M}_h$, maillage fin, et $\mathcal{M}_{2h}$, maillage grossier. On suppose que ces maillages sont « emboîtés » c'est-à-dire que les nœuds du maillage grossier sont aussi des nœuds du maillage fin. Pour fixer les idées, et bien que ceci ne constitue pas une condition nécessaire, on suppose de plus que le maillage grossier est deux fois moins dense que le maillage fin.

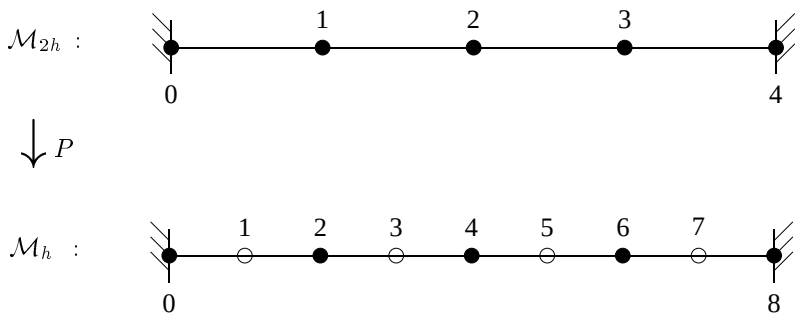
Définition 4.1 (Opérateur de prolongement)

On appelle prolongement tout opérateur d'interpolation sur le maillage fin de fonctions connues par leurs discrétisations sur le maillage grossier.

Nota Bene : En employant le terme d'opérateur d'interpolation on suppose implicitement qu'il satisfait une condition de consistance et même de précision, que l'on s'attachera à vérifier dans les cas particuliers.

Exemple de prolongement :

$$P = I_{2h}^{h} \ : \ \mathcal{M}_{2h} \longrightarrow \mathcal{M}_h \tag{4.1}$$



« Interpolation linéaire »

$$u_{2h} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \longrightarrow u_h = \begin{pmatrix} u'_1 = (u_0 + u_1)/2 \\ u'_2 = u_1 \\ u'_3 = (u_1 + u_2)/2 \\ u'_4 = u_2 \\ u'_5 = (u_2 + u_3)/2 \\ u'_6 = u_3 \\ u'_7 = (u_3 + u_4)/2 \end{pmatrix} \tag{4.2}$$

où les valeurs aux limites, dans l'exemple u_0 et u_4, ne sont pas des degrés de liberté et n'interviennent pas dans l'identification de l'opérateur de prolongement :

$$P = \begin{pmatrix} 1/2 & 0 & 0 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 0 & 0 & 1/2 \end{pmatrix} \tag{4.3}$$

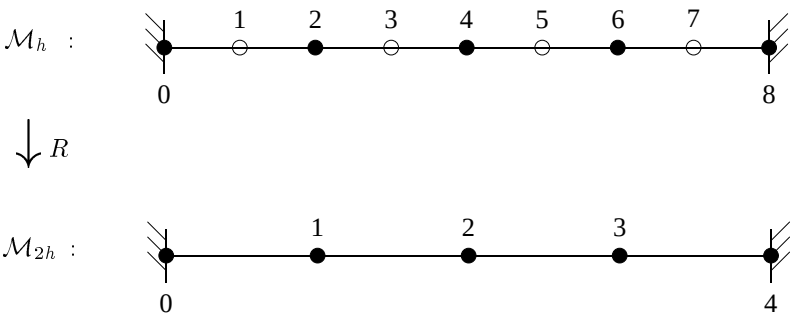
Définition 4.2 (Opérateur de restriction)

On appelle restriction tout opérateur d'interpolation sur le maillage grossier de fonctions connues par leurs discrétisations sur le maillage fin.

Nota Bene : A nouveau, les notions de consistance et de précision sont implicites.

Exemples de restrictions :

$$R = I_h^{2h} \ : \ \mathcal{M}_h \longrightarrow \mathcal{M}_{2h} \tag{4.4}$$



1. « Injection »

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{4.5}$$

   Dans le cas du modèle discret fondamental, on constate qu'en « injectant » des vecteurs propres associés à la discrétisation fine (cf. Théorème 1.1), on obtient les vecteurs propres associés à la discrétisation grossière. En ce sens, l'« injection respecte les modes propres ».

2. L'opérateur de restriction suivant se révèle préférable :

$$R = \frac{1}{2}\, P^T = \begin{pmatrix} 1/4 & 1/2 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 1/2 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1/2 & 1/4 \end{pmatrix} \tag{4.6}$$

Exercice 4.1 (Approximation de grille grossière)

Vérifier que lorsque le choix est fait de ces opérateurs de prolongement et de restriction, on a le résultat important suivant :

$$\boxed{R\, A_h\, P = (2h)^{-2}\ \text{Trid}\,(-1, 2, -1) = A_{2h}} \tag{4.7}$$

4.1.2. Algorithme d'enrichissement de maillage

Cet algorithme est connu dans la littérature anglophone sous le nom de « Nested Iteration » [55] [19]. Le processus consiste à résoudre le même problème d'EDP d'abord sur une grille grossière, puis à prolonger le résultat sur une grille plus fine, à itérer à nouveau, puis à prolonger, et ainsi de suite jusqu'à obtention de la solution convergée sur la grille fine. Intuitivement, on anticipe un gain en efficacité, la réduction du coût de l'itération étant due au début du processus, au faible nombre de degrés

de liberté, et à la fin, au fait qu'on dispose d'une très bonne initialisation par interpolation d'une solution convergée sur un maillage presqu'assez fin. Dans le but d'estimer le gain en efficacité, rappelons les éléments dont on dispose :

– une suite de maillages (emboîtés) de finesse croissante :

$$\begin{gathered}\mathcal{M}_{h_k} \subset \mathcal{M}_{h_{k-1}} \subset \ldots \subset \mathcal{M}_{h_1} \subset \mathcal{M}_{h_0} \text{ (fin)} \\ h_\ell = 2^\ell h\end{gathered} \tag{4.8}$$

– sur chaque maillage, une approximation de l'EDP de type donné, de sorte que l'« erreur d'approximation » a la même dépendance de la taille h_ℓ de la maille sur chaque niveau de grille ℓ :

$$\|e_{h_\ell}\| \sim C_a \ (h_l)^\alpha \quad \text{(e.g. } \alpha = 2) \tag{4.9}$$

– sur chaque niveau de grille, une méthode itérative dont le rayon spectral s'exprime comme suit en fonction de la maille :

$$\rho_{h_\ell} = 1 - C_b \ (h_\ell)^\beta + \ldots \quad \Longrightarrow -\ln \rho_{h_\ell} \sim C_b \, h_\ell^\beta \tag{4.10}$$

permettant de résoudre le système discret correspondant à convergence partielle. Le paramètre β de cette équation n'est autre que l'ordre de l'EDP étudiée ($\beta = 2$ pour un laplacien ; voir annexe D).

Plus précisément, on procède comme suit.

Algorithme de convergence par enrichissement

1. Maillage grossier ($\ell = k$)

   Itérer jusqu'à satisfaction du critère d'arrêt relatif à $\mathcal{M}_{h_k}$. Il en résulte une approximation discrète de la solution notée

   $$u_{h_k}^{\nu_k} \tag{4.11}$$

   qui, en général, ne satisfait ni le critère de convergence itérative complète (ν_k fini), ni le critère de convergence de l'approximation (h_k trop grand par rapport à la précision souhaitée seulement atteinte par la discrétisation de grille fine).

2. Maillages fins ($\ell = k-1, k-2, \ldots, 0$) : sur chaque nouvelle grille

   (a) Prolonger : $u_{h_\ell}^0 = I_{h_{\ell+1}}^{h_\ell} \, u_{h_{\ell+1}}^{\nu_{\ell+1}}$
   (b) Itérer jusqu'à satisfaction du critère d'arrêt relatif à $\mathcal{M}_{h_\ell}$ pour obtenir :

   $$u_{h_\ell}^{\nu_\ell} \tag{4.12}$$

4.1.3. Gain théorique

On souhaite « résoudre » le problème sur la grille fine, $\mathcal{M}_h = \mathcal{M}_{h_0}$. On dispose de deux méthodes possibles : la méthode itérative de base que l'on a caractérisée par son rayon spectral, et l'algorithme d'enrichissement progressif, qui utilise aussi cette méthode itérative de base, mais en construisant simultanément des discrétisations de plus en plus fines jusqu'à atteindre la même grille fine. Faisant référence à notre discussion du chapitre 1 sur les critères d'arrêt, on rappelle que « résoudre » signifie itérer jusqu'à satisfaction du critère d'arrêt, c'est-à-dire jusqu'à ce que l'erreur itérative soit de l'ordre de l'erreur d'approximation. Pour évaluer le gain, nous allons estimer le coût de résolution correspondant à chacune des deux méthodes.

Coût de la méthode d'origine

Le nombre d'itérations μ_h nécessaires à la satisfaction du critère d'arrêt est donné par :

$$\frac{\|e_h^{\mu_h}\|}{\|e_h^0\|} \sim \rho_h^{\mu_h} \sim C_a\, h^{\alpha} \tag{4.13}$$

d'où l'estimation

$$\mu_h \sim \frac{\alpha\, \ln 1/h}{\ln 1/\rho_h} \sim \frac{-\alpha\, \ln h}{C_b\, h^{\beta}} \tag{4.14}$$

Sachant que par itération, le nombre d'opérations effectuées est proportionnel par hypothèse au nombre de degrés de liberté, on obtient l'estimation suivante du coût :

$$\text{coût} \sim \frac{-\alpha c\, \ln h}{C_b\, h^{\beta+1}} \quad (M = \frac{1}{h}) \tag{4.15}$$

(où c est le coût par degré de liberté et par itération).

Coût de l'algorithme d'enrichissement

Le coût de l'amorçage sur la grille grossière $\mathcal{M}_{h_k}$ satisfait la même loi à condition de remplacer le paramètre h par $h_k = 2^k\, h$:

$$\text{coût}_k \sim \frac{-\alpha c\, \ln h_k}{C_b\, h_k^{\beta+1}} < \frac{1}{2^{k(\beta+1)}}\, \frac{-\alpha c\, \ln h}{C_b\, h^{\beta+1}} \tag{4.16}$$

Cette contribution est négligée.

Pour ce qui est de la résolution partielle effectuée sur une grille intermédiaire, $\mathcal{M}_{h_\ell}$ ($\ell < k$), le nombre suffisant d'itérations ν_ℓ est caractérisé par la condition suivante :

$$\rho_{h_\ell}^{\nu_\ell} = \frac{\|e_{h_\ell}\|}{\|e_{h_{\ell+1}}\|} \sim 2^{-\alpha} \quad (\forall\, \ell) \tag{4.17}$$

d'où l'estimation :

$$\nu_\ell \sim \frac{\alpha \ln 2}{-\ln \rho_{h_\ell}} \sim \frac{\alpha \ln 2}{C_b\, h_\ell^\beta} \tag{4.18}$$

de sorte que le coût des opérations effectuées sur cette grille est donné par :

$$\text{coût}_\ell \sim \frac{\alpha c \ln 2}{C_b\, h_\ell^{\beta+1}} \quad (M_\ell = \frac{1}{h_\ell}) \tag{4.19}$$

Une estimation du coût global de l'algorithme s'obtient en sommant seulement ces contributions ($\ell = 0, 1, ..., k-1$) ; il vient :

$$\begin{aligned} \text{coût global} &\sim \frac{\alpha c \ln 2}{C_b} \frac{1}{h^{\beta+1}} \left(1 + \frac{1}{2^{\beta+1}} + \frac{1}{4^{\beta+1}} + \dots\right) \\ &\sim \frac{\alpha c \ln 2}{K C_b} \frac{1}{h^{\beta+1}} \end{aligned} \tag{4.20}$$

où la constante K a l'expression suivante :

$$K = \frac{1 - (\frac{1}{2})^{\beta+1}}{1 - (\frac{1}{2})^{k(\beta+1)}} \tag{4.21}$$

Cette constante est inférieure à, mais proche de 1.

Gain

En comparant les estimations précédentes du coût, (4.15)-(4.20), on constate que l'enrichissement progressif de maillage permet de réaliser une économie de calcul d'un facteur de :

$$\boxed{\frac{-K \ln h}{\ln 2} = K \log_2 M = \frac{K}{d} \log_2 N} \tag{4.22}$$

Dans cette équation on a fait apparaître le nombre :

$$N = M^d \tag{4.23}$$

qui serait représentatif du nombre de degrés de liberté dans le cas d'un problème à d dimensions d'espace.

Le gain théorique est proportionnel au logarithme du nombre de degrés de liberté, ici en base 2, parce que l'on a supposé le rapport des densités de points de deux grilles successives égal à 2.

En conclusion, on constate que l'enrichissement progressif du maillage permet de réaliser un gain effectif en efficacité, même si celui-ci est modeste, puisqu'il est de l'ordre du nombre maximum de niveaux de grille que l'on peut emboîter dans la grille fine utilisée.

4.2. Travaux pratiques

Exercice 4.2 (Expérimentation numérique d'enrichissement progressif de maillage)

On considère à nouveau le problème modèle fondamental (1.8), ici dans le cas spécifique où le terme source f est la fonction :

$$f \equiv \pi^2 \sin \pi x \tag{4.24}$$

et sa discrétisation habituelle sur un maillage uniforme $\{x_j = jh\}$ ($j = 0, 1, ..., M+1\ ;\ h = 1/(M+1)$), explicitée en (1.10).

On note $\{r_j\}$ les composantes du résidu :

$$r_j = \frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} - f_j \quad (j = 1, 2, ..., M) \tag{4.25}$$

1. On rappelle (cf. exercice 1.1) que pour ce problème, l'« erreur d'approximation », $\| u_h - u \|_\infty$, satisfait la majoration suivante :

$$\| u_h - u \|_\infty \leq \frac{\mu h^2}{96} \tag{4.26}$$

où le paramètre

$$\mu = \max_{x \in [0,1]} |f''(x)| \tag{4.27}$$

sera évalué. Pour diverses itérations, on souhaite définir un critère d'arrêt qui garantisse que l'erreur itérative e_h^n satisfasse la même majoration :

$$\| e_h^n \|_\infty \leq \frac{\mu h^2}{96} \tag{4.28}$$

Montrer qu'une condition suffisante pour que ceci soit vrai est que le résidu itératif respecte la condition suivante :

$$\| r_h^n \|_\infty = \max_j \left| r_j^n \right| \leq \frac{\mu h^2}{12} \tag{4.29}$$

2. Résoudre numériquement ce problème dans le cas où $M + 1 = 32$ par la méthode itérative de Jacobi. On initialisera la solution par la fonction nulle, et on interrompra l'itération à satisfaction du critère d'arrêt (4.29). On notera dans un tableau le nombre

d'intervalles de discrétisation, $M+1$, les valeurs initiale et finale de la norme du résidu, le nombre d'itérations effectuées et le nombre d'« unités de travail, UT » dépensées (nombre d'itérations $\times$ nombre de points intérieurs au maillage). Observer, justifier et commenter.

3. Refaire l'expérience avec une suite de maillages emboîtés de densité croissante ($M+1 = 4,\ 8,\ 16,\ 32$). Sur chaque nouveau niveau ($M+1 = 8,\ 16,\ 32$), on utilisera comme condition initiale, la solution convergée sur le niveau précédent ($M+1 = 4,\ 8,\ 16$) prolongée par interpolation. L'itération de Jacobi sera interrompue à satisfaction du critère d'arrêt (4.29) particularisé à la valeur du pas d'espace ($h = \Delta x = \frac{1}{4},\ \frac{1}{8},\ \frac{1}{16},\ \frac{1}{32}$) correspondant au niveau de grille.

Cette expérience sera effectuée pour deux types d'interpolation :

• l'interpolation linéaire définie comme suit :
Soit $\mathcal{M}_{2h}$ la grille grossière sur laquelle une approximation de la solution est connue : $\{u_j\}$ $(j = 0, 1, 2, ..., (M+1)/2)$ satisfaisant les conditions aux bords $u_0 = u_{(M+1)/2} = 0$. On note $\{u'_j\}$ $(j = 0, 1, 2, ..., M+1)$ les valeurs initiales extrapolées sur la grille fine (2 fois plus dense) $\mathcal{M}_h$, et on pose pour $j = 0, 1, 2, ..., (M+1)/2$:

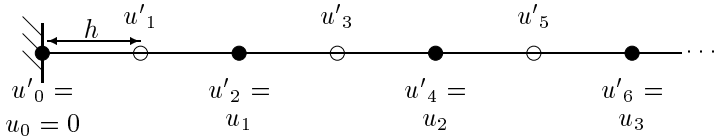
$$\begin{cases} u'_{2j} = u_j \\ u'_{2j-1} = \dfrac{u'_{2j-2} + u'_{2j}}{2} = \dfrac{u_{j-1} + u_j}{2} \quad (j \geq 1) \end{cases} \tag{4.30}$$

• l'interpolation suivante (dont on établira la précision) définie avec les mêmes notations comme suit :

– points communs aux deux grilles, pour $j = 0, 1, 2, ..., (M+1)/2$:

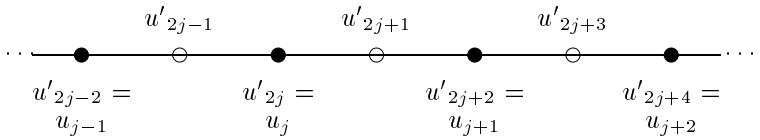
$$u'_{2j} = u_j \tag{4.31}$$

– point intérieur gauche de la grille fine :



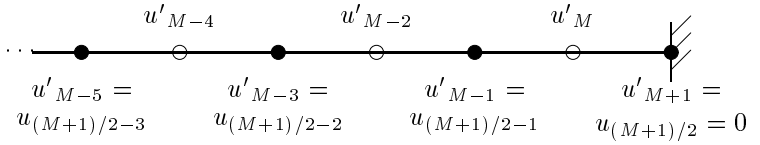
$$\begin{aligned} u'_1 &= \tfrac{5}{16}u'_0 + \tfrac{15}{16}u'_2 - \tfrac{5}{16}u'_4 + \tfrac{1}{16}u'_6 \\ &= \tfrac{5}{16}u_0 + \tfrac{15}{16}u_1 - \tfrac{5}{16}u_2 + \tfrac{1}{16}u_3 \end{aligned} \tag{4.32}$$

– points (vraiment) intérieurs n'appartenant pas à la grille grossière, $j = 1, 2, ..., (M+1)/2 - 2$:



$$\begin{aligned} u'_{2j+1} &= -\tfrac{1}{16}u'_{2j-2} + \tfrac{9}{16}u'_{2j} + \tfrac{9}{16}u'_{2j+2} - \tfrac{1}{16}u'_{2j+4} \\ &= -\tfrac{1}{16}u_{j-1} + \tfrac{9}{16}u_j + \tfrac{9}{16}u_{j+1} - \tfrac{1}{16}u_{j+2} \end{aligned} \tag{4.33}$$

– point intérieur droit de la grille fine :



$$\begin{aligned} u'_M &= \tfrac{1}{16}u'_{M-5} - \tfrac{5}{16}u'_{M-3} + \tfrac{15}{16}u'_{M-1} + \tfrac{5}{16}u'_{M+1} \\ &= \tfrac{1}{16}u_{(M+1)/2-3} - \tfrac{5}{16}u_{(M+1)/2-2} + \tfrac{15}{16}u_{(M+1)/2-1} + \tfrac{5}{16}u_{(M+1)/2} \end{aligned} \tag{4.34}$$

Constituer un tableau analogue au précédent pour chaque type d'interpolation (une ligne de résultats par niveau de grille). Observer, justifier et commenter.

4. On reprend l'expérimentation de l'enrichissement progressif du maillage avec interpolation linéaire de la solution sur chaque nouveau niveau de grille, mais ici, après chaque prolongement, on effectue une phase de lissage de la solution préalablement au démarrage de l'itération de Jacobi. Comme lisseur, on pourra considérer 2 applications d'un cycle de Richardson constitué de 3 pseudo-pas de temps ajustés pour atténuer optimalement les modes de hautes fréquences (ce qui équivaut en coût à 6 itérations de Jacobi). Observer, justifier et commenter.

Chapitre 5

La méthode multigrille en elliptique

Il s'agit de résoudre le grand système (d'abord supposé linéaire)

$$\boxed{A_h\, u_h = f_h} \tag{5.1}$$

qui représente la discrétisation d'une EDP elliptique sur une certaine grille $\mathcal{M}_h$ dont la finesse, réglée par l'exigence de précision, est donnée. Au chapitre précédent, on a vu qu'un certain gain en efficacité peut effectivement être réalisé en construisant d'autres maillages, moins fins, et en appliquant pour résoudre une stratégie d'enrichissement progressif. Cependant, ce gain est modeste, de l'ordre du nombre de niveaux de grille utilisés pour accéder à la grille fine dans la dernière phase du calcul. Nous allons voir maintenant que la clé du succès de la méthode multigrille qui se base aussi sur une hiérarchie de niveaux de grille, repose sur les deux éléments essentiels suivants :

- la construction d'un cycle utilisant la grille fine dans la phase initiale,
- le réglage de l'itération en opérateur de lissage.

Pour préciser ces notions, on se place d'abord dans le cas de deux grilles.

5.1. Méthode bigrille idéale

On considère deux grilles emboîtées : $\mathcal{M}_h$ (grille fine, M_h degrés de liberté) et $\mathcal{M}_{2h}$ (grille grossière, M_{2h} degrés de liberté). On suppose pour simplifier l'exposé que la grille grossière est deux fois moins dense. Dans le cas de maillages unidimensionnels uniformes, ceci correspond à

$$M_h + 1 = 2\ (M_{2h} + 1) \tag{5.2}$$

Cette hypothèse a un impact sur le réglage du lisseur, mais toute hypothèse du même type serait tout aussi viable.

5.1.1. Cycle symétrique

On souhaite définir un cycle de résolution utilisant l'itération de grille fine, non seulement dans la phase finale, mais également dans la phase initiale (là est la nouveauté). Ce cycle comporte trois phases dans sa version (symétrique) la plus simple (figure 5.1).

$$\mathcal{M}_h \searrow \mathcal{M}_{2h} \nearrow \mathcal{M}_h \tag{5.3}$$

Figure 5.1. Schéma du cycle bigrille symétrique le plus simple

Phase 1 : Lissage sur la grille fine $\mathcal{M}_h$
En notant u_h^0 *et* u'_h *l'approximation (grille fine) initiale et le résultat de cette phase, on a dans le cas linéaire :*

$$\boxed{u'_h = G_h\, u_h^0 + b_h} \tag{5.4}$$

Par exemple, dans le cas du laplacien discret 1D, on a vu au chapitre 3 qu'un algorithme de lissage efficace était réalisé par la méthode de Richardson avec 3 paramètres optimisés aux hautes fréquences. Dans ce cas, la matrice d'amplification (ou plutôt d'atténuation) G_h a la forme suivante :

$$G_h = \left(I - \tau_3\, h^2 A_h\right)\left(I - \tau_2\, h^2 A_h\right)\left(I - \tau_1\, h^2 A_h\right) \tag{5.5}$$

Le vecteur b_h est de la forme :

$$b_h = B_h\, f_h \tag{5.6}$$

où la matrice B_h est aussi un polynôme de A_h.

A l'issue de cette phase, on évalue le résidu :

$$\boxed{r'_h = A_h\, u'_h - f_h} \tag{5.7}$$

La définition de la deuxième phase se fonde sur les éléments intuitifs suivants :

– en linéaire, les définitions du résidu (ci-dessus) et de l'erreur

$$e'_h = u'_h - u_h \tag{5.8}$$

impliquent l'équivalence :

$$A_h\, u_h = f_h \Longleftrightarrow A_h\, e'_h = r'_h \tag{5.9}$$

ainsi que l'équation suivante,

$$u_h = u'_h - e'_h \tag{5.10}$$

qui donnerait la solution exacte du problème discret, u_h, si on connaissait l'erreur, e'_h, associée à l'approximation la plus récente, u'_h ;

– les composantes « hautes fréquences » (HF) du vecteur erreur e'_h (et non de la solution courante u'_h!) sont négligeables en raison du lissage effectué dans la première phase. Par exemple, pour le modèle discret fondamental, dans le cas de la méthode de Richardson avec 3 pseudo-pas de temps optimisés aux HF, les composantes HF ont été atténuées par le facteur $\frac{1}{99}$ dans le pire des cas. Une bonne estimation du vecteur e'_h peut donc se construire par approximation seulement de ses composantes « basses fréquences » (BF) ;

– les composantes BF du vecteur e'_h sont suffisamment bien représentées par une discrétisation sur la grille grossière $\mathcal{M}_{2h}$.

A partir de ces remarques, on cherche à calculer une approximation du vecteur e'_h proche du prolongement d'une fonction discrétisée sur la grille grossière $\mathcal{M}_{2h}$:

$$e'_h \approx P\, e'_{2h} \tag{5.11}$$

En conséquence, on a l'équation approchée :

$$A_h\, P\, e'_{2h} \approx r'_h \tag{5.12}$$

qui est redondante puisqu'on a réduit de moitié le nombre d'inconnues, mais vraie de manière approchée par hypothèse. On se ramène à un système carré par application de l'opérateur de restriction R ce qui donne :

$$R\, A_h\, P\, e'_{2h} \approx R\, r'_h \tag{5.13}$$

Mais on a vu au chapitre précédent (cf. (4.7)) que pour certains types de discrétisations et certains choix des opérateurs de transfert P et R on peut avoir :

$$R\, A_h\, P = A_{2h} \tag{5.14}$$

c'est-à-dire la matrice d'approximation de la même EDP sur la grille grossière. En général, cette équation est satisfaite seulement de manière approchée. Ici, on supposera qu'elle est satisfaite de manière exacte. On est donc amené à définir la deuxième phase de l'algorithme comme suit :

Phase 2 : Correction de grille grossière
Dans la méthode bigrille « idéale » on résout de manière « exacte » le problème suivant sur $\mathcal{M}_{2h}$:

$$\boxed{A_{2h}\, e'_{2h} = R\, r'_h} \tag{5.15}$$

puis on prolonge l'estimation de l'erreur :

$$\boxed{e'_h = P\, e'_{2h}} \tag{5.16}$$

et on corrige l'approximation de grille fine précédente :

$$\boxed{u''_h = u'_h - e'_h} \tag{5.17}$$

Notre intuition nous laisse penser que l'on a ainsi bien résolu les composantes « basses fréquences » du problème (celles pour lesquelles les transferts de grille à grille sont précis), mais que le prolongement de l'estimation de l'erreur, (5.16), a peut-être réintroduit des erreurs de hautes fréquences. Ces erreurs s'éliminent aisément par une nouvelle phase de lissage qui constitue la phase ultime de l'algorithme bigrille symétrique le plus simple.

Phase 3 : Lissage sur la grille fine $\mathcal{M}_h$

$$\boxed{u'''_h = G_h\, u''_h + b_h} \tag{5.18}$$

Pour analyser rigoureusement cet algorithme, on met de côté les remarques intuitives, et on rassemble les équations intervenant dans la définition des trois phases :

$$\left\{\begin{array}{l} u'_h = G_h\, u^0_h + b_h \\ r'_h = A_h\, u'_h - f_h \\ (R\, A_h\, P)\; e'_{2h} = R\, r'_h \\ e'_h = P\, e'_{2h} \\ u''_h = u'_h - e'_h \\ u'''_h = G_h\, u''_h + b_h \end{array}\right. \tag{5.19}$$

Par conséquent, le cycle complet équivaut à l'application affine suivante :

$$u_h''' = G\, u_h^0 + b_h' \tag{5.20}$$

où la matrice d'amplification G a la structure suivante,

$$G = G_h \left\{ I - P\,(R\,A_h\,P)^{-1}\; R\,A_h \right\} G_h \tag{5.21}$$

qui reflète les trois phases de la résolution : lissage grille fine/correction grille grossière/lissage grille fine.

L'efficacité du cycle bigrille est mise en évidence par le théorème suivant dont la démonstration est donnée dans l'annexe E.

Théorème 5.1 (Symétrisation du cycle bigrille idéal)

On se place dans le cadre du problème modèle 1D, et on suppose que la matrice G_h associée aux phases de lissage s'exprime comme le polynôme suivant de la matrice d'approximation A_h :

$$G_h = \mathcal{P}\left(h^2\, A_h\right) \tag{5.22}$$

et on introduit la fonction :

$$d(\lambda) = \sqrt{\lambda}\,\mathcal{P}(\lambda) \tag{5.23}$$

En particulier, dans le cas d'un lissage par la méthode de Richardson avec 3 pseudo-pas de temps optimisés :

$$\mathcal{P}(\lambda) = \prod_{\ell=1}^{3} (1 - \tau_\ell\,\lambda) = \frac{T_3(3-\lambda)}{T_3(3)} \tag{5.24}$$

(1) La matrice d'amplification G est semblable à la matrice symétrique suivante :

$$G' = D\,\Sigma\,D \tag{5.25}$$

où l'on a posé :

$$\boxed{\begin{array}{c} D = d\,(\Lambda_h) \quad \textit{(diagonale)} \\ \Sigma = \Lambda_h^{-1} - 8\,\sigma^T\,\Lambda_{2h}^{-1}\,\sigma \quad \textit{(symétrique)} \\ \sigma = S_{2h}\,R\,S_h \end{array}} \tag{5.26}$$

où les matrices (S_h, Λ_h) et (S_{2h}, Λ_{2h}) interviennent dans la diagonalisation des matrices d'approximation associées aux deux grilles :

$$A_h = h^{-2}\,S_h\,\Lambda_h\,S_h\,,\ A_{2h} = (2h)^{-2}\,S_{2h}\,\Lambda_{2h}\,S_{2h} \tag{5.27}$$

(2) La matrice diagonale $D = \mathrm{Diag}\,(d_m)$ associée aux phases de lissage, admet les éléments diagonaux suivants :

$$\boxed{d_m = D_{m,m} = d\left(\lambda_m^h\right)} \tag{5.28}$$

où $\left\{\lambda_m^h\right\}$ $(m = 1, 2, ..., M_h)$ sont les valeurs propres de la matrice $h^2\,A_h$ (différences non divisées).

(3) La « matrice de correction de grille grossière » (en base fréquentielle) Σ a la structure suivante :

$$\Sigma = \begin{pmatrix} \frac{1}{4} & & & & & & & & \frac{1}{4} \\ & \frac{1}{4} & & & & & & \frac{1}{4} & \\ & & \ddots & & & & \iddots & & \\ & & & \frac{1}{4} & & \frac{1}{4} & & & \\ & & & & \frac{1}{2} & & & & \\ & & & \frac{1}{4} & & \frac{1}{4} & & & \\ & & \iddots & & & & \ddots & & \\ & \frac{1}{4} & & & & & & \frac{1}{4} & \\ \frac{1}{4} & & & & & & & & \frac{1}{4} \end{pmatrix} \tag{5.29}$$

REMARQUE : la forme « en croix » de la matrice Σ met bien en évidence le phénomène d'« *aliasing* » introduit par les transferts de grille à grille. En effet, chaque mode fréquentiel se transforme en $\frac{1}{4}$ de lui-même auquel s'ajoute $\frac{1}{4}$ du mode de fréquence complémentaire.

En conséquence, on a le résultat suivant concernant la vitesse de convergence :

Corollaire 5.1 (rayon spectral du cycle bigrille idéal)

Le rayon spectral du cycle bigrille idéal s'exprime comme suit

$$\boxed{\rho = \rho\left(G\right) = \max_{m=1,\ldots,M_{2h}+1} \left(\frac{d_m^2 + d_{M_h+1-m}^2}{4} \right)} \tag{5.30}$$

et admet la borne supérieure suivante

$$\boxed{B_0 = \max_{\lambda \in [0,2]} \left[\frac{d(\lambda)^2 + d(4-\lambda)^2}{4} \right]} \tag{5.31}$$

indépendamment de h. En particulier, dans le cas du cycle de Richardson précédent :

$$\boxed{B_0 = \frac{5032 + 313\sqrt{313}}{264627} \approx 0.04} \tag{5.32}$$

Ce résultat confirme l'efficacité de la construction. En réexaminant (5.21), on comprend le rôle complémentaire des phases de lissage, performantes dans l'atténuation des composantes « hautes fréquences » de l'erreur, et de la phase de correction-grille grossière dont l'action est caractérisée par la matrice entre accolades $\{...\}$. Cette matrice serait formellement nulle si les opérateurs de prolongement P et de restriction R étaient des matrices carrées inversibles. En fait, ce n'est pas le cas puisque ce sont des matrices rectangulaires, car les transferts de grille à grille correspondent à des changements de dimension d'espace qui s'accompagnent de pertes d'information. Intuitivement, le résultat de l'application de la matrice $\{...\}$ au discrétisé d'une fonction est proche de 0 lorsque ces transferts sont précis, c'est-à-dire lorsqu'ils agissent sur le discrétisé d'une fonction qui ne contient que de « basses fréquences ». Chaque facteur dans (5.21) a donc sa fonction propre et le produit attaque le spectre complet.

REMARQUE : pour l'algorithme multigrille idéal, lorsque le niveau de convergence itérative est fixé par la précision, ce qui constitue une hypothèse un peu différente de l'« hypothèse d'école » de la section précédente, le nombre d'applications nécessaires du cycle est fini. Par conséquent, le rayon spectral, qui en toute rigueur ne caractérise que la vitesse de convergence *asymptotique* de l'algorithme, n'est pas le paramètre le

plus informatif sur la convergence. Une vraie norme serait préférable. Certaines majorations en norme sont fournies par les ouvrages théoriques, principalement [46] et [18], mais aussi [95].

CONTRE-REMARQUE : le fait d'avoir établi une majoration aussi favorable du rayon spectral suffit à justifier qu'un faible nombre de cycles multigrilles sont suffisant en pratique.

5.1.2. Cycles non symétriques en « dent de scie »

Certains auteurs préfèrent introduire une seule phase de lissage par cycle. Par exemple, si on lisse seulement avant le transfert sur la grille grossière, et dans l'hypothèse où les deux phases de lissage précédentes sont regroupées en une seule, la matrice d'amplification du cycle, en remplacement de (5.21) devient

$$G' = \left\{ I - P \left(R\, A_h\, P \right)^{-1}\, R\, A_h \right\} G_h^2 \tag{5.33}$$

A l'inverse, si on lisse seulement après la correction de grille grossière, on aura plutôt :

$$G" = G_h^2 \left\{ I - P \left(R\, A_h\, P \right)^{-1}\, R\, A_h \right\} \tag{5.34}$$

Dans la littérature, on attribue la dénomination de « cycles en dent de scie » (*saw-tooth*) aux algorithmes correspondants.

On rappelle que pour tout couple de matrices carrées A et B de même dimension (inversibles ou non), les matrices AB et BA sont semblables. Il est alors évident que les matrices G, G' et $G"$ sont semblables et ont donc le même rayon spectral. Cet argument ne s'étend pas au cas non linéaire pour lequel des considérations liées au stockage de la solution militent en faveur du lissage préalable au transfert sur la grille grossière.

5.2. Généralisations : cycle multigrille et méthode multigrille complète

Nous venons de voir, qu'on peut construire une méthode de résolution itérative du système discret dont la vitesse de convergence, dans l'hypothèse « idéale » où l'on résout à chaque itération, exactement et complètement un système discret du même type ayant deux fois moins de degrés de liberté, est indépendante de la densité du maillage fin (c'est-à-dire indépendante de h).

En remplaçant dans l'algorithme bigrille idéal la phase 2 de correction grille grossière par un algorithme bigrille idéal associé à des niveaux de grille $\mathcal{M}_{2h}$ et $\mathcal{M}_{4h}$, on obtient l'algorithme *trigrille idéal*. En généralisant récursivement ce concept au cas

de $k+1$ maillages emboîtés

$$\begin{gathered}\text{(fin) } \mathcal{M}_{h_0} \supseteq \mathcal{M}_{h_1} \supseteq \ldots \supseteq \mathcal{M}_{h_{k-1}} \supseteq \mathcal{M}_{h_k} \\ h_\ell = 2^\ell h\end{gathered} \tag{5.35}$$

on aboutit à l'algorithme *multigrille idéal* dont nous supposerons sans démonstration le rayon spectral borné indépendamment de la densité du maillage fin (c'est-à-dire indépendamment de $h = h_0$) :

$$\boxed{\rho_{MG} \leq B < 1\,,\ \forall\, h} \tag{5.36}$$

En conséquence, notons ν le nombre de cycles multigrilles (MG) suffisant à ce que l'erreur itérative atteigne le niveau de l'erreur d'approximation associée à la grille fine. Par définition,

$$\rho_{MG}^{\nu} = O\left(h^{\alpha}\right) \tag{5.37}$$

ce qui donne

$$\nu = O\left(\frac{\ln\, 1/h}{\ln\, 1/\rho_{MG}}\right) = O\left(-\ln h\right) \tag{5.38}$$

car

$$\frac{1}{\ln\, 1/\rho_{MG}} \leq \frac{1}{\ln\, 1/B} = \text{constante.} \tag{5.39}$$

En supposant que la discrétisation est locale, c'est-à-dire telle qu'un nombre borné de voisins interviennent dans l'équation discrète en chaque nœud (3 pour le laplacien 1D, 5 en 2D, etc.), le nombre d'opérations effectuées par itération de lissage est lui-même constant par nœud. Le coût d'une itération de lissage est donc proportionnel au nombre de degrés de liberté

$$\boxed{N = M_h^d} \tag{5.40}$$

(d : dimension d'espace). De même, le coût global d'un cycle complet multigrille est proportionnel à N.

En définitive, le coût global de la résolution du problème par l'algorithme multigrille est donné par :

$$\boxed{\text{coût } = O\left(-N\, \ln h\right) = O\left(N\, \ln N\right)} \tag{5.41}$$

On dit aussi que l'algorithme est de « complexité » $N \ln N$.

Parfois, lorsque le nombre de degrés de liberté associés à la grille la plus grossière est suffisamment faible, le système discret correspondant n'est pas résolu itérativement mais par une méthode directe telle que l'élimination de Gauss. Dans le cas inverse, l'itération est interrompue lorsque l'erreur itérative atteint le niveau de grandeur de l'erreur d'approximation associée à cette discrétisation. Dans ce cas, le cycle est appelé « Méthode Multigrille » sans le qualificatif « idéale ».

On aboutit enfin au concept le plus intéressant : la « Méthode Multigrille Complète » (*Full Multigrid Method*, « FMG » dans la littérature anglophone). Cette méthode, schématisée à la figure 5.2 dans le cas de trois niveaux de grille, combine les concepts d'enrichissement progressif et de multigrille.

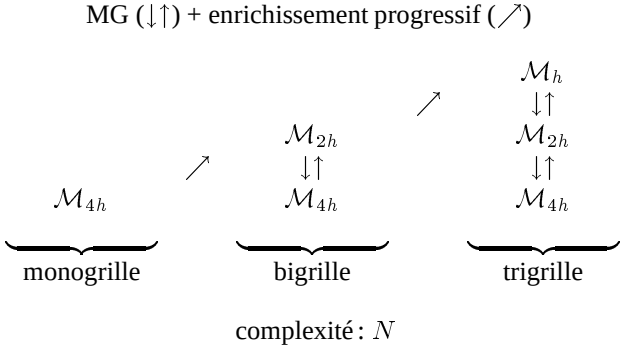


Figure 5.2. Schéma de la méthode multigrille complète dans le cas de trois niveaux

Dans le cas de la figure, on résout d'abord complètement le problème sur la grille grossière $\mathcal{M}_{4h}$ par l'itération de base monogrille. Ceci signifie qu'on interrompt cette itération dès que l'erreur itérative atteint le niveau de l'erreur d'approximation associée à la discrétisation sur $\mathcal{M}_{4h}$. On prolonge le résultat sur la grille $\mathcal{M}_{2h}$ et on résout itérativement, au moyen de l'algorithme bigrille, le problème discret correspondant, à partir d'une condition initiale déjà bonne, et seulement jusqu'à ce que l'erreur itérative (qui a changé d'espace) ait atteint le niveau de l'erreur d'approximation associée à la discrétisation sur $\mathcal{M}_{2h}$. On prolonge enfin le résultat sur la grille $\mathcal{M}_h$ et on résout itérativement, au moyen de l'algorithme trigrille, le problème discret souhaité. Dans ce processus, chaque algorithme de type « ℓ-grille » ($\ell = 1, 2, 3$ dans l'exemple) n'opère que pour réduire l'erreur itérative d'un facteur constant égal au rapport d'ordres de grandeur entre les erreurs d'approximation associées à deux niveaux de grille successifs. Le coût correspondant est donc directement proportionnel au nombre de degrés

de liberté correspondant, et le coût global du calcul est donné par :

$$\text{coût} = \underbrace{C\,N}_{\substack{\nu \text{ cycles MG avec} \\ (k+1) \text{ grilles}}} + \underbrace{\frac{C\,N}{2}}_{\substack{\nu \text{ cycles MG avec} \\ k \text{ grilles}}} + \underbrace{\frac{C\,N}{4}}_{\substack{\nu \text{ cycles MG avec} \\ (k-1) \text{ grilles}}} + \ldots \qquad (5.42)$$

ce qui fournit le résultat suivant, qui est le plus fondamental dans la théorie des multigrilles :

$$\boxed{\text{coût} \leq\ 2\,C\,N} \qquad (5.43)$$

et permet de conclure :

Théorème 5.2 (Vitesse de convergence théorique de la méthode multigrille complète)

Dans le cas du modèle théorique (elliptique et linéaire), la méthode multigrille complète est idéalement de complexité proportionnelle au nombre de degrés de liberté N, ou de manière équivalente, sa vitesse de convergence est indépendante de la densité du maillage fin (c'est-à-dire indépendante de h).

REMARQUE : l'enrichissement de maillage réduit d'un facteur $\ln\ N$ la complexité de la méthode multigrille qui est $N\ \ln\ N$, d'où le résultat.

Pour conclure cette section, on a rassemblé au tableau 5.1 les résultats concernant la complexité des méthodes étudiées.

5.3. Une bibliothèque sur le réseau Internet

Signalons l'existence de la bibliothèque 'MGGHAT' (MultiGrid Galerkin Hierarchical Adaptive Triangles) due à W.F. Mitchell issue de ses travaux de thèse [72]. Cette bibliothèque permet la résolution par une méthode multigrille des équations associées à la discrétisation Galerkin sur une triangulation quelconque fournie par l'utilisateur d'une EDP elliptique linéaire générale en dimension deux mise sous la forme

$$\boxed{-\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(q\frac{\partial u}{\partial y}\right) + ru = f \quad (\text{dans } \Omega)} \qquad (5.44)$$

soumise aux conditions aux limites

$$u = g \quad (\text{sur } \partial\Omega_1) \qquad (5.45)$$

$$p\frac{\partial u}{\partial x}\frac{\partial y}{\partial s} - q\frac{\partial u}{\partial y}\frac{\partial x}{\partial s} + cu = g \quad (\text{sur } \partial\Omega_2) \qquad (5.46)$$

Rappel des notations

Solution du problème continu :	u
Solution du problème discret :	u_h
Itéré n de la méthode considérée :	u_h^n
Pas d'espace :	$h = 1/M$ ($= \Delta x$, Δy ou Δz)
Nombre de degrés de liberté :	$N = M^d$ (d : dimension d'espace) M : nombre de modes / direction spatiale (pour simplifier : $M_x = M_y = M_z = M$)
Coût d'une itération grille fine :	$\mathcal{C} = \lambda\, N = \lambda\, M^d$
Erreur d'approximation :	$u_h - u = C_a\, h^\alpha + \ldots$ (généralement $\alpha = 2$)
Erreur itérative :	$u_h^n - u_h$ $\|u_h^n - u_h\| / \|u_h^0 - u_h\| = O\,(\rho^n)$ ρ : rayon spectral de la méthode de base
Critère d'arrêt :	$\|u_h^n - u_h\| = O\,(\|u_h - u\|)$

Méthode de base, « MB »

$$\rho = 1 - C_b\, h^\beta + \ldots \text{et :}$$
$$n \times (-\ln\, \rho) = \alpha \times (-\ln\, h) + \ldots$$
$$\Downarrow$$
$$n \sim \text{const.} \times M^\beta\, \ln\, M$$
$$\Downarrow$$
COMPLEXITÉ :
$$O\left(M^{\beta+d}\, \ln\, M\right) = O\left(N^{\beta/d+1}\, \ln\, N\right)$$

MB + enrichissement progressif

Même ρ, mais ici il suffit que :
$$\rho^n \sim \|u_h - u\| / \|u_{2h} - u\| \sim 2^{-\alpha} = \text{const.}$$
$$\Downarrow$$
$$n \sim \text{const.} \times M^\beta$$
$$\Downarrow$$
COMPLEXITÉ :
$$O\left(M^{\beta+d}\right) = O\left(N^{\beta/d+1}\right)$$

Multigrille, « MG »

$\rho = \text{const.}$
i.e. $\beta = 0$
$$\Downarrow$$
$$n \sim \text{const.} \times \ln\, M$$
$$\Downarrow$$
COMPLEXITÉ :
$$O\left(M^d\, \ln\, M\right) = O\,(N\, \ln\, N)$$

MG + enrichissement = « FMG »

$\rho = \text{const.}$ ($\beta = 0$), mais ici,
pour le dernier cycle, il suffit que :
$$\rho^n \sim \|u_h - u\| / \|u_{2h} - u\| \sim 2^{-\alpha} = \text{const.}$$
$$\Downarrow$$
$$n = \text{const.}$$
$$\Downarrow$$
COMPLEXITÉ :
$$O\left(M^d\right) = O\,(N)$$

Tableau 5.1. Complexité de diverses méthodes de résolution

où $p > 0$, $q > 0$, r, f, c et g sont des fonctions données de x et y, Ω est un domaine polygonal de $\mathbb{R}^2$, $\partial\Omega_1 \cup \partial\Omega_2 = \partial\Omega$ (le bord complet) et $\partial/\partial s$ est la dérivation partielle par rapport à une variable de paramétrisation du bord dans le sens direct.

Cette bibliothèque appartient au domaine public et on peut accéder aux sources à travers *netlib*, *mgnet* ou *ftp anonyme*. Par exemple, sur *mgnet*, MGGHAT peut être copiée par ftp anonyme à partir de casper.cs.yale.edu dans le répertoire mgnet/mgghat.

Exercice 5.1 (Complexité de la méthode multigrille complète)

Dans le but de résoudre le problème modèle discret unidimensionnel, on a conduit une première campagne de calculs par la « méthode multigrille complète », en utilisant 3 grilles emboîtées uniformes $\mathcal{M}_{h_i}$ ($i = 1, 2, 3$) définies par les valeurs suivantes du pas d'espace $h = \Delta x$:

$$\begin{aligned} h_1 &= \tfrac{1}{45} \text{ (grille fine)}, \\ h_2 &= \tfrac{1}{15} \text{ (grille moyenne)}, \\ h_3 &= \tfrac{1}{5} \text{ (grille grossière)}. \end{aligned} \tag{5.47}$$

On note C le coût global de cette campagne de calculs.

Des tests révèlent que la solution de grille fine ainsi calculée est insuffisamment précise et on décide de conduire une seconde campagne de calculs, initialisée par les résultats de la précédente, afin de prendre en compte une nouvelle grille, encore plus fine, $\mathcal{M}_{h_0}$, où :

$$h_0 = \tfrac{1}{135} \tag{5.48}$$

Quel est le coût de cette nouvelle campagne de calculs ?

Exercice 5.2 (Variante de la méthode bigrille idéale)

On se place à nouveau dans le cadre de la résolution du modèle discret unidimensionnel 1.10. On souhaite construire une variante de la méthode bigrille du cours dans laquelle la grille fine $\mathcal{M}_h$ serait 3 fois plus dense que la grille grossière notée $\mathcal{M}_{3h}$. On note M_h et M_{3h} les valeurs du nombre de degrés de liberté correspondant à ces grilles de sorte que :

$$\begin{cases} M_h = \mu\, 3^I - 1 \\ M_{3h} = \mu\, 3^{I-1} - 1 \end{cases} \tag{5.49}$$

où μ et I sont des entiers. Par exemple avec $\mu = 4$ et $I = 1$, on a $M_h = 11$ (soit 12

intervalles de discrétisation) et $M_{3h} = 3$ (4 intervalles) comme le schématise la figure 5.3 ci-dessous :

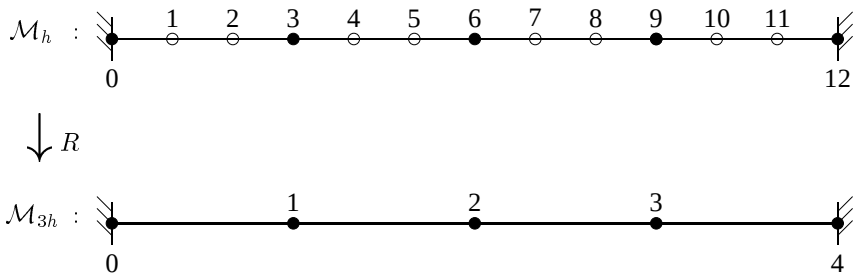


Figure 5.3. Maillages unidimensionnels $\mathcal{M}_h$ et $\mathcal{M}_{3h}$ dans le cas où $M_h = 11$.

On rappelle que les vecteurs propres de la matrice A_h peuvent être ordonnés suivant les valeurs croissantes d'un paramètre de fréquence,

$$\theta_m = \frac{m\pi}{M_h + 1} \quad (m = 1, 2, ..., M_h) \tag{5.50}$$

qui varie (à peu près) de 0 à π et que la valeur propre associée est donnée par :

$$\lambda_{h\,m} = \frac{2 - 2\,\cos\,\theta_m}{h^2} \tag{5.51}$$

Plus précisément, la j-ième composante du m-ième vecteur propre $s^{(m)}$ est donnée par :

$$s_j^{(m)} = s_{j,m} = \psi^{(m)}(x_j) = C\,\sin\,(j\theta_m) = C\,\sin\,(m\pi x_j) \tag{5.52}$$

où x_j est l'abscisse du noeud j, $\psi^{(m)}(x)$ la m-ième fonction propre (sinusoïdale) du problème continu et C une constante de normalisation.

(1) Pour quelles valeurs de m la restriction à la grille grossière (par injection directe) du vecteur propre $s^{(m)}$ est-elle un vecteur propre de la matrice A_{3h} associée à la discrétisation sur la grille grossière $\mathcal{M}_{3h}$? Conventionnellement, on désigne les modes correspondants comme ceux de « basses fréquences ».

En conséquence, quelles sont les bornes de variation du paramètre $h^2\,\lambda_{h\,m}$ correspondant aux :

(i) basses fréquences, et
(ii) hautes fréquences ?

(2) On adopte comme opérateur de lissage, la méthode de Richardson avec k pseudo-pas de temps. On souhaite que les modes de hautes fréquences soient atténués par ce cycle au moins du facteur 10^{-2}.

Choisir k pour remplir cette condition, expliciter les pseudo-pas de temps correspondants et le facteur d'atténuation des hautes fréquences effectivement réalisé.

Exercice 5.3 (Etude d'un problème aux limites anisotrope)

On souhaite résoudre le problème suivant

$$\begin{cases} -100\dfrac{\partial^2 u}{\partial x^2} - \dfrac{\partial^2 u}{\partial y^2} = 1 \quad \Big((x,y) \in \Omega =]0,1[\times]0,1[\Big) \\ u = 0 \quad \text{sur } \partial\Omega \end{cases} \tag{5.53}$$

par une « méthode bigrille idéale » en utilisant des maillages cartésiens uniformes mais *étirés*. Plus précisément, le maillage fin est 10 fois plus dense dans la direction de y, à savoir :

$$\begin{cases} h_x = \Delta x = \dfrac{1}{M+1} = \dfrac{1}{10} \\ h_y = \Delta y = \dfrac{1}{L+1} = \dfrac{1}{100} \end{cases} \tag{5.54}$$

où $M = 9$ (resp. $L = 99$) désigne le nombre de degrés de liberté suivant la direction des x (resp. des y). On discrétise par les différences finies centrées usuelles

$$-100\,\frac{u_{j-1,k} - 2u_{j,k} + u_{j+1,k}}{h_x^2} - \frac{u_{j,k-1} - 2u_{j,k} + u_{j,k+1}}{h_y^2} = 1 \tag{5.55}$$

ce qui conduit au système discret suivant à résoudre :

$$A_h\, u_h = f_h \tag{5.56}$$

où l'on a posé :

$$A_h = \text{“Penta”}\Big(\ldots, -1, \ldots, -1, 4, -1, \ldots, -1, \ldots\Big) \tag{5.57}$$

et

$$f_h = 10^{-4}\Big(1, 1, \ldots, 1\Big)^T \tag{5.58}$$

En conséquence, la valeur au nœud (j,k) du vecteur propre dont les paramètres de fréquence ont pour indices m et ℓ est donnée par (3.142), à savoir :

$$(s_{xy})_{j,k}^{(m,\ell)} = C\ \sin\, j\theta_{x\,m}\ \sin\, k\theta_{y\,\ell} \tag{5.59}$$

où C est une constante de normalisation et les paramètres de fréquence $\theta_{x\,m}$ et $\theta_{y\,\ell}$ ont les expressions suivantes :

$$\theta_{x\,m} = \frac{m\pi}{M+1} = \frac{m\pi}{10}\,, \quad m = 1, 2, ..., M = 9, \tag{5.60}$$

$$\theta_{y\,\ell} = \frac{\ell\pi}{L+1} = \frac{\ell\pi}{100}\,, \quad \ell = 1, 2, ..., L = 99, \tag{5.61}$$

et les valeurs propres de la matrice A_h sont données par :

$$\lambda^h_{m,\ell} = \Big(2 - 2\cos\,\theta_{x\,m}\Big) + \Big(2 - 2\cos\,\theta_{y\,\ell}\Big) \tag{5.62}$$

On construit une grille grossière *emboîtée* en déraffinant massivement dans la direction de y (d'un facteur 10) mais seulement dans cette direction. Autrement dit, la grille grossière contient 9×9 nœuds et les paramètres qui lui sont associés sont les suivants :

$$h'_x = h_x = \frac{1}{10} \quad (M' = M = 9) \tag{5.63}$$

$$h'_y = 10\,h_y = \frac{1}{10} \quad (L' = 9) \tag{5.64}$$

(1) Déterminer les limites du spectre complet grille fine :

$$\begin{array}{c} \lambda^h_{\text{min}} = \min\,\lambda_{m,\ell}\,, \quad \lambda^h_{\text{max}} = \max\,\lambda_{m,\ell}\,, \\ (m = 1, 2, ..., 9\,,\ \ell = 1, 2, ..., 99) \end{array} \tag{5.65}$$

On a vu qu'un vecteur propre quelconque associé à la grille fine est le discrétisé sur cette grille d'un produit de fonctions sinusoïdales ; ce discrétisé par injection sur la grille grossière subit le phénomène d'*aliasing* lorsque les paramètres de fréquence qui lui correspondent sont trop grands. En particulier, justifier que les modes de « basses fréquences » correspondent ici aux valeurs suivantes des indices :

$$\textbf{BF :}\ m = 1, 2, ..., 9\,,\ \ell = 1, 2, ..., 9 \tag{5.66}$$

A l'inverse, les modes de « hautes fréquences » sont associés aux indices :

$$\textbf{HF :}\ m = 1, 2, ..., 9\,,\ \ell = 10, 11, ..., 99 \tag{5.67}$$

Pour ces modes, déterminer :

$$\begin{array}{c} a = \min\,\lambda^h_{m,\ell}\,, \quad b = \max\,\lambda^h_{m,\ell} \\ (m, \ell)\ \in\ \{1, 2, ..., 9\} \times \{10, 11, .., 99\} \end{array} \tag{5.68}$$

(2) Sur la grille fine, on « lisse » par la méthode usuelle de Richardson en utilisant un cycle de k pseudo-pas-de-temps $\{\tau_\ell\}$ optimisés aux HF :

$$u_h^{n+l} = u_h^{n+l-1} - \tau_\ell\ \left(A_h\,u_h^{n+l-1} - f_h\right) \quad (\ell = 1, 2, ..., k) \tag{5.69}$$

– Rappeler l'expression formelle des paramètres optimaux $\{\tau_\ell^*\}$ en fonction de a, b et k.

– Comment choisir k pour assurer que les composantes fréquentielles de hautes fréquences de l'erreur soient atténuées par le cycle de Richardson au moins du facteur $\frac{1}{10}$?

Chapitre 6

Quelques applications des méthodes multigrilles en mécanique des fluides

6.1. Introduction

Nos analyses jusqu'ici ont porté sur le modèle fondamental unidimensionnel qui nous a permis de mettre en évidence les principales caractéristiques théoriques des méthodes multigrilles. Bien évidemment, la portée des méthodes multigrilles en calcul scientifique généralement, et en mécanique des fluides en particulier, dépasse très largement le contexte de ce modèle simplifié. Dans ce chapitre, sans chercher à faire une présentation d'ensemble des méthodes multigrilles en mécanique des fluides, on discute des principales difficultés rencontrées lorsqu'on cherche à incorporer une stratégie multigrille à un code de calcul d'écoulement de type « éléments finis non structurés » tels qu'ils sont couramment utilisés à l'INRIA [1] depuis une quinzaine d'années. On y présente les solutions qui ont été éprouvées au sein du Projet SINUS [2] à Sophia-Antipolis ou dans les équipes de leurs proches collaborateurs. On y aborde particulièrement les aspects suivants :

– le traitement de problèmes non linéaires ;

– l'usage de maillages non structurés ;

– l'incidence de mailles étirées, le semi-déraffinement, les multigrilles adaptatives ;

1. INRIA : Institut National de Recherche en Informatique et en Automatique
2. SINUS : Simulation Numérique dans les Sciences de l'Ingénieur
(`http://www.inria.fr/sinus/sinus.html`)

– l'impact de termes hyperboliques dominants ;

– le traitement de problèmes algébriques en l'absence d'interprétation géométrique de domaine.

Avant cela, on présente brièvement une méthode d'approximation en maillages non structurés des équations d'Euler qui régissent un écoulement de fluide parfait compressible en régime permanent ou instationnaire.

6.2. Méthode hybride par éléments finis/volumes finis pour les écoulements compressibles

Dans le cadre de ses recherches sur les méthodes numériques visant les applications à la mécanique des fluides compressibles, le Projet SINUS a développé depuis une quinzaine d'années, sous l'impulsion de A. Dervieux, et en collaboration avec ses partenaires des milieux académiques et industriels, plusieurs méthodologies numériques incorporant notamment les ingrédients suivants :

– formulation hybride par éléments finis/volumes finis applicable aux maillages non structurés ;

– approximation décentrée des termes de convection par utilisation de solveurs de Riemann approchés et extrapolation MUSCL ;

– approximation centrée de type P1-Lagrange des termes de diffusion (dans le cas des équations de Navier-Stokes en régime d'écoulement laminaire ou turbulent, éventuellement réactif) ;

– résolution des équations discrètes par un schéma pseudo-instationnaire implicite linéarisé de type « *Defect-Correction* », ou par relaxation non linéaire ;

– alternative Jacobi/multigrille pour la résolution du système linéaire ;

– adaptation aux architectures parallèles.

Ces options numériques ont été largement utilisées pour mener à bien des études scientifiques d'aide à la modélisation ou au développement en aérodynamique compressible, combustion et hypersonique notamment, mais servent également de base au code N3S-Natur [76] à vocation industrielle. Nous référons au rapport d'activité du Projet depuis le début des années 1980 pour une bibliographie exhaustive, ainsi qu'à [35]. Dans le but d'inclure dans cet ouvrage introductif aux techniques de résolution, un aperçu du « noyau » de la méthode de base utilisée dans les calculs d'aérodynamique présentés ultérieurement, on définit ici très succinctement ces principaux in-

grédients dans le cadre déjà très représentatif de la résolution des équations d'Euler stationnaires en deux dimensions d'espace ; celles-ci peuvent s'écrire sous la « forme divergence » suivante :

$$\text{div}\,\overrightarrow{H}(W) \stackrel{\text{déf}}{=} \frac{\partial F(W)}{\partial x} + \frac{\partial G(W)}{\partial y} = 0\,, \quad \overrightarrow{H}(W) = \begin{pmatrix} F \\ G \end{pmatrix} \tag{6.1}$$

où W est le vecteur des « variables conservatives », dont les champs sont les inconnues du problème :

$$W = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix} \tag{6.2}$$

où ρ est la masse volumique locale du fluide, u et v les composantes cartésiennes de la vitesse matérielle, et E l'énergie totale (interne + cinétique) spécifique. Les vecteurs $F(W)$ et $G(W)$ sont les fonctions de flux suivantes :

$$F(W) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E+p) \end{pmatrix}\,, \quad G(W) = \begin{pmatrix} \rho v \\ \rho vu \\ \rho v^2 + p \\ v(E+p) \end{pmatrix} \tag{6.3}$$

où p est la pression locale qui s'exprime en fonction des autres variables grâce à l'équation d'état qui, pour un gaz parfait diatomique, ou pour l'air, considéré comme un mélange de tels gaz, se traduit comme suit :

$$p = \rho r T = (\gamma - 1)\Big(E - \frac{u^2 + v^2}{2}\Big) \tag{6.4}$$

où $r = \mathcal{R}/M$ est la constante du gaz, M sa masse molaire, $\mathcal{R}$ la constante universelle des gaz parfaits et γ le rapport des capacités calorifiques à pression et volume constants ($\gamma = \frac{7}{5}$ pour un mélange de gaz parfaits diatomiques).

Ces équations traduisent, pour un écoulement bidimensionnel permanent (ou stationnaire) de fluide continu compressible, la conservation de la masse (continuité), de la quantité de mouvement (loi de Newton) et de l'énergie (premier principe de la thermodynamique). Bien évidemment, le système ci-dessus est soumis à un jeu de conditions aux limites qui dépendent du problème et que nous ne détaillerons pas ici.

La nature mathématique du système stationnaire précédent dépend de manière essentielle de la valeur locale du « nombre de Mach » dont la définition est la suivante :

$$M = \frac{V}{c} \tag{6.5}$$

où $V = \sqrt{u^2 + v^2}$ est le module de la vitesse et $c = \sqrt{\gamma p/\rho}$ est la célérité du son.

Lorsque l'écoulement est localement supersonique ($M > 1$) le système stationnaire est hyperbolique. Lorsque cette condition est réalisée uniformément dans le domaine, on peut calculer la solution à partir d'une ligne de données située en amont de l'écoulement par des techniques d'avancement en espace (discrétisation de type différences finies dans la direction d'avancement, ou méthode des caractéristiques, par exemple).

Par contre, le système stationnaire est elliptique dans les zones où l'écoulement est subsonique ($M < 1$). Par conséquent, si le domaine de calcul contient de telles zones, d'étendue *a priori* inconnue, on ne peut uniformément procéder par intégration en espace et les équations discrètes doivent être résolues itérativement.

Une manière courante de construire une méthode itérative pour résoudre le problème stationnaire consiste à intégrer en temps, à partir d'une condition initiale :

$$W(x, y, 0) = W_0(x, y) \tag{6.6}$$

en principe arbitraire (mais admissible), les équations d'Euler instationnaires :

$$W_t + \frac{\partial F(W)}{\partial x} + \frac{\partial G(W)}{\partial y} = 0 \tag{6.7}$$

jusqu'à convergence asymptotique ($t \to \infty$). Ces équations régissent un écoulement instationnaire de fluide parfait compressible, et sont *hyperboliques* quel que soit le régime de cet écoulement. Cette propriété provient du fait que les matrices jacobiennes (de dimension 4×4 en 2D),

$$A = A(W) = \frac{\partial F(W)}{\partial W}\,, \quad B = B(W) = \frac{\partial G(W)}{\partial W} \tag{6.8}$$

qui admettent des diagonalisations connues et sont simultanément symétrisables (sur $\mathbb{R}$), sont telles que pour tout couple de réels (k_1, k_2) la matrice $k_1\, A + k_2\, B$ est diagonalisable sur $\mathbb{R}$; ses valeurs propres réelles sont les nombres [94] :

$$\lambda_1 = \lambda_2 = k_1\, u + k_2\, v \quad \text{(double)} \tag{6.9}$$

$$\lambda_{3,4} = k_1\, u + k_2\, v + \pm c \sqrt{k_1^2 + k_2^2} \tag{6.10}$$

Cette condition équivaut à dire que la solution du système résultant d'une linéarisation autour d'un état constant résulte de la superposition d'ondes simples du type

$$\widehat{W}(x, y, t) = \widehat{W}_0\, e^{k_1\, x + k_2\, y - i\, \lambda\, t} \tag{6.11}$$

Exercice 6.1 (Nature mathématique des équations d'Euler stationnaires)

Utiliser (6.9)-(6.10) pour établir la nature mathématique du système des équations d'Euler stationnaires suivant le nombre de Mach, M.

Notons enfin que les expressions mêmes des fonctions de flux en font des fonctions homogènes de degré 1 des variables conservatives ; en vertu du théorème d'Euler, on a donc les identités suivantes :

$$F(W) = A(W)\,W\,, \quad G(W) = B(W)\,W \tag{6.12}$$

On souhaite maintenant donner quelques éléments descriptifs du schéma d'approximation spatiale du terme divergence, div $\overrightarrow{H}(W)$, et de la technique de résolution par intégration en temps de (6.7) ou par relaxation non linéaire.

6.2.1. Approximation spatiale décentrée en maillage non structuré

Lorsqu'on vise notamment les applications à l'aérodynamique industrielle, on souhaite développer une méthode d'approximation suffisamment souple pour s'appliquer à une discrétisation non structurée du domaine. En effet, dans les applications complexes, le problème de la génération d'un maillage initial est en soi un problème de calcul scientifique non trivial, coûteux en heures d'ingénieur autant qu'en temps de calcul, et l'approche « maillage non structuré » qui permet de discrétiser le domaine en « simplex » simplifie le cadre de cette construction initiale. De plus cette approche permet de rendre les nécessaires procédures d'adaptation de maillage à la solution bien plus locales, et donc plus faciles à mettre en œuvre, et plus économiques en nombre d'éléments rajoutés. Ces considérations ont conduit le Projet SINUS comme d'autres équipes à considérer en 2D la donnée d'une triangulation arbitraire (en 3D un maillage constitués de tétraèdres) et à chercher une approximation spatiale du second ordre, ce qui milite pour une représentation des fonctions de type P1-Lagrange, c'est-à-dire linéaire par élément et s'appuyant sur des degrés de liberté placés aux sommets (x_i, y_i) des triangles :

$$W_i = W(x_i, y_i) \tag{6.13}$$

L'approximation repose sur la construction préalable d'un « maillage dual » obtenu comme suit : on trace les médianes des triangles ; on identifie autour de chaque nœud i une cellule de contrôle C_i dont les arêtes sont des portions des médianes des triangles dont le nœud i est un sommet (voir figure 6.1).

Appliquant la formule de Green à l'intégrale étendue à la cellule du terme de

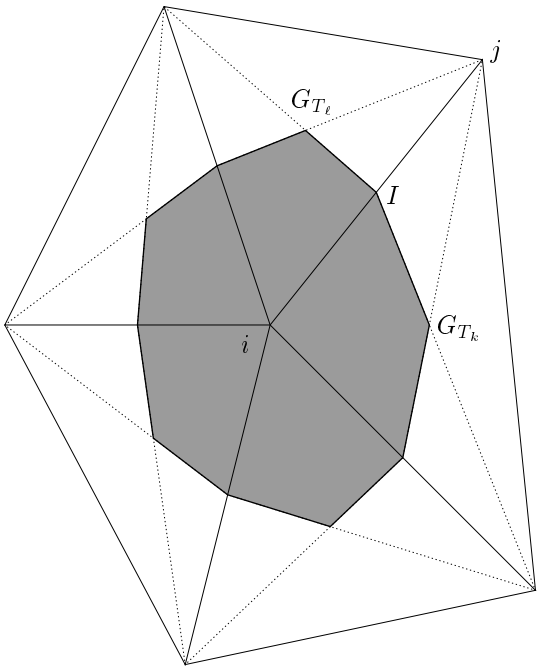


Figure 6.1. Cellule du maillage dual de volumes finis construit à partir d'une triangulation quelconque

divergence, il vient :

$$\iint_{C_i} \text{div}\, \overrightarrow{H}(W)\, dx\, dy = \iint_{C_i} \left(\frac{\partial F(W)}{\partial x} + \frac{\partial G(W)}{\partial y} \right) dx\, dy \tag{6.14}$$

$$= \int_{\partial C_i} \Big(n_x\, F(W) + n_y\, G(W) \Big)\, ds \tag{6.15}$$

$$= \sum_{j \text{ voisin de } i} \phi_{i,j} \tag{6.16}$$

où s est l'abscisse curviligne le long du bord de la cellule, le vecteur

$$\vec{n} = \begin{pmatrix} n_x \\ n_y \end{pmatrix} \tag{6.17}$$

est la normale unitaire extérieure à la cellule, et :

$$\phi_{i,j} = \int_{\partial C_i \cap \partial C_j} \Big(n_x \, F(W) + n_y \, G(W) \Big) \, ds \tag{6.18}$$

est le flux à travers l'interface $\partial C_i \cap \partial C_j$ entre le nœud i et le nœud voisin j ; cette interface est composée de 2 segments reliant le milieu I du segment ij aux centres de gravité G_{T_k} et G_{T_ℓ} des triangles dont le segment ij est le côté commun. Ce flux est ensuite approché par l'expression suivante :

$$\phi_{i,j} \approx \eta_x \, F_I + \eta_y \, G_I \tag{6.19}$$

où l'indice I portant sur les flux F et G indique une approximation (définie ultérieurement) consistante à la valeur au point I, et le vecteur

$$\vec{\eta} = \int_{\partial C_i \cap \partial C_j} \vec{n} \, ds \tag{6.20}$$

est la normale intégrée.

En raison de l'invariance des équations d'Euler lors d'un changement de base euclidienne, il existe une rotation $\mathcal{R}_\gamma$ pour laquelle l'expression précédente du flux se ramène formellement à celle d'un écoulement 1D :

$$\phi_{i,j} = \| \, \eta \, \| \, \mathcal{R}_\gamma^{-1} \, F\Big(\mathcal{R}_\gamma \, W_I \Big) \tag{6.21}$$

Exercice 6.2 (Invariance des équations d'Euler par rotation)

Démontrer la relation ci-dessus en explicitant la matrice $\mathcal{R}_\gamma$ pour laquelle elle est vraie.

En conséquence, une fois choisie une certaine décomposition de flux, par exemple celle de Van Leer ([90] ou [91]),

$$F(U) = F^+(U) + F^-(U) \tag{6.22}$$

où les flux $F^\pm$ ont la propriété spectrale suivante :

$$\forall \, U \text{ (admissible)} \ , \ \forall \, \lambda \ \in \ \sigma\Big(\frac{\partial F^\pm(U)}{\partial U} \Big) \ , \ \lambda \ \in \ \mathbb{R}^\pm \ , \tag{6.23}$$

une approximation décentrée d'ordre 1 est obtenue en complétant (6.21) de la définition :

$$F\Big(\mathcal{R}_\gamma \, W_I \Big) = \underbrace{F^+\Big(\mathcal{R}_\gamma \, W_i \Big)}_{\substack{\text{influence} \\ \text{amont}}} + \underbrace{F^-\Big(\mathcal{R}_\gamma \, W_j \Big)}_{\substack{\text{influence} \\ \text{aval}}} . \tag{6.24}$$

Enfin, le passage à l'ordre 2 se réalise par une adaptation (aux maillages non structurés) de la technique MUSCL (*Monotone Upstream-centred Schemes for Conservation Laws*) de Van Leer ([89], ou [91]), dans laquelle les vecteurs W_i et W_j sont remplacés dans (6.24) par des vecteurs $W_{i,j}$ et $W_{j,i}$ obtenus par extrapolation linéaire à partir des nœuds i et j respectivement :

$$W_{i,j} = W_i + \frac{\overrightarrow{ij}}{2}.\overrightarrow{\nabla W}_i \tag{6.25}$$

$$W_{j,i} = W_j - \frac{\overrightarrow{ij}}{2}.\overrightarrow{\nabla W}_j \tag{6.26}$$

Dans ces extrapolations, le gradient $\overrightarrow{\nabla W}_i$ (resp. $\overrightarrow{\nabla W}_j$) peut se calculer en faisant la moyenne étendue aux triangles T_k dont le nœud i (resp. j) est un sommet, du gradient constant par triangle de l'approximation P1-Lagrange de W pondérée par les aires de ces triangles, c'est-à-dire :

$$\overrightarrow{\nabla W}_i = \frac{\sum_k \text{aire}(T_k)\,\overrightarrow{\nabla W}_{T_k}}{\sum_k \text{aire}(T_k)} \tag{6.27}$$

Une alternative à cette formule centrée est d'utiliser le gradient associé au « triangle amont », $T_{i,j}$ (resp. $T_{j,i}$), c'est-à-dire le triangle contenant le point $i - \varepsilon\,\overrightarrow{ij}$ (resp. $j + \varepsilon\,\overrightarrow{ij}$) pour tout $\varepsilon > 0$ suffisamment petit (voir figure 6.2) ; on peut aboutir dans ce cas à une approximation totalement décentrée.

Figure 6.2. Construction de « triangles amont »

A chaque type d'approximation (centrée ou totalement décentrée) correspond des vecteurs

$$\delta_i^- = W_{i,j} - W_i \tag{6.28}$$

$$\delta_j^+ = W_{j,i} - W_j \tag{6.29}$$

dont les 4 composantes sont des perturbations à apporter aux composantes nodales en i ou j pour obtenir un extrapolé à l'interface. En pratique, pour chaque nœud, la perturbation effectivement apportée à une composante donnée est calculée par une formule

non linéaire de moyenne des deux types de correction citées ; ce type de moyenne est dû à Van Albada et al. [88]. Par cette construction, le schéma d'approximation, dans le contexte plus simple d'une équation hyperbolique scalaire en une dimension d'espace, aurait la propriété d'être « TVD » (à variation totale décroissante) ce qui assure certaines propriétés de préservation de la monotonie (voir [7]). Alternativement, on peut utiliser des procédures de « limitation de pente » (voir par exemple [35]).

6.2.2. Résolution par intégration pseudo-instationnaire implicite

La formulation instationnaire (6.7) a été introduite un peu artificiellement pour construire une itération dont le point fixe, atteint lorsque $t \to \infty$, est la solution du problème. Il est donc souhaitable que l'itération converge aussi vite que possible ; autrement dit que la stabilité de l'intégration en temps autorise l'utilisation de pas de temps Δt aussi grands que possible. Pour cette raison, on choisit la méthode d'Euler implicite qui appliquée à l'équation modèle de l'advection pure est inconditionnellement stable. Appliquée à (6.7) après intégration à la cellule C_i, le schéma en temps s'écrit :

$$\mathcal{A}_i \, \frac{W_i^{n+1} - W_i^n}{\Delta t} + \Big(\Phi_h(W^{n+1}) \Big)_i = 0 \quad (i \ : \ \text{nœud courant}) \tag{6.30}$$

où

$$\mathcal{A}_i = \text{aire}(C_i) \tag{6.31}$$

est l'aire de la cellule C_i, n est l'indice d'itérations (ou nombre de pas de temps effectués), le vecteur W_i est considéré comme la moyenne du champ W à la cellule :

$$W_i^n = \frac{1}{\mathcal{A}_i} \iint_{C_i} W(x, y, t^n) \, dx \, dy \tag{6.32}$$

et, d'après le paragraphe précédent, le flux Φ_h a la forme suivante :

$$\Big(\Phi_h(W^{n+1}) \Big)_i = \sum_{j \text{ voisin de } i} \phi_{i,j}^{n+1} \tag{6.33}$$

Cette méthode est d'ordre 1 en temps, mais la précision de la solution stationnaire, définie par :

$$\Phi_h(W) = 0 \tag{6.34}$$

ne dépend que du schéma d'approximation spatiale dont le paragraphe précédent fournit une option. Sans changer l'ordre de l'erreur de troncature on peut linéariser Φ_h :

$$\Phi_h(W^{n+1}) = \Phi_h(W^n) + \Phi_h'(W^n) \, (W^{n+1} - W^n) \tag{6.35}$$

où Φ'_h est le jacobien de l'approximation spatiale. On rassemble les équations sous la forme incrémentale suivante parfois appelée « forme Δ » :

$$\left(\frac{\mathcal{A}}{\Delta t} + \Phi'_h(W^n)\right)(W^{n+1} - W^n) = -\Phi_h(W^n) \tag{6.36}$$

où $\mathcal{A}$ est la matrice diagonale ayant la structure par bloc scalaire suivante :

$$\mathcal{A} = \text{BDiag}\ (\mathcal{A}_i\ I_4) \tag{6.37}$$

D'un point de vue algorithmique, la résolution se décompose en 3 phases :

1. Phase explicite – Calcul du second membre

$$\boxed{f_h = -\Phi_h(W^n)} \tag{6.38}$$

(calcul des flux eulériens à l'ordre 2) ;

2. Phase implicite – Calcul des blocs coefficients de la matrice

$$\boxed{A_h = \frac{\mathcal{A}}{\Delta t} + \Phi'_h(W^n)} \tag{6.39}$$

et résolution du système

$$\boxed{A_h\ \delta W = f_h} \tag{6.40}$$

3. Réactualisation

$$\boxed{W^{n+1} = W^n + \delta W} \tag{6.41}$$

La phase explicite contrôle la précision de la solution convergée. On la conduit à l'ordre 2 (au moins).

La phase implicite est un préconditionneur contrôlant la stabilité et la vitesse de convergence de l'itération. Il est possible de la conduire également à l'ordre 2. Cela nécessite la délicate construction du jacobien d'une approximation décentrée d'ordre 2. Lorsque ceci est fait, la convergence itérative est proche de quadratique [84], ce qui s'explique par le fait que si Δt était infini, l'algorithme s'identifierait dans ce cas à la méthode de Newton. Alternativement, le choix d'une approximation du jacobien à l'ordre 1 seulement présente plusieurs avantages :

1. simplicité algorithmique de la construction de la matrice A_h ;

2. erreur de troncature à caractère plus dissipatif favorisant la convergence vers l'état stationnaire ;

3. (et surtout) structure « à diagonale dominante » de la matrice A_h dans le cas d'un modèle hyperbolique linéaire, permettant donc la résolution de la phase implicite par *relaxation*.

Ce dernier point est extrêmement important dans le cas où l'on utilise une formulation en maillage non structuré pour laquelle la matrice A_h n'a pas en général de structure « bande » pour laquelle une résolution directe du système (6.40) serait viable.

En pratique, on résout le système linéaire partiellement seulement par quelques itérations de Jacobi, Gauss-Seidel ou par une méthode multigrille, avant de réactualiser la matrice au pas de temps suivant. Au cours de l'itération non linéaire en temps, le pas de temps est graduellement augmenté jusqu'à atteindre de manière stable des valeurs très grandes, parfois même sans limitation de stabilité suivant le problème.

Lorsque le pas de temps Δt est infini, l'itération non linéaire prend la forme :

$$\Phi_h^1(W^{n+1}) = \Phi_h^1(W^n) - \Phi_h^2(W^n) \tag{6.42}$$

où Φ_h^2 est l'approximation précise (ici à l'ordre 2 au moins) et Φ_h^1 une approximation moins précise (ici à l'ordre 1) mais plus facile à inverser (ici quasi-linéaire et à diagonale dominante pour le problème modèle). Certains auteurs de la littérature anglophone désignent ce type d'itération comme un algorithme de « *Defect-Correction* » [13]. Les propriétés itératives de cet algorithme ont été étudiées dans le cas de modèles hyperboliques et évaluées sur les équations d'Euler dans [36] et [38]. En particulier, pour le problème modèle linéaire de l'advection pure soumis à des conditions de type Dirichlet (en amont) et Neumann (bord libre aval), et certaines approximations décentrées, le rayon spectral de l'itération est égal à $\frac{1}{2}$, indépendamment de la finesse du maillage. Voir également [69] pour l'étude d'une méthode implicite de même type mais précise au second ordre en temps.

6.2.3. Résolution par relaxation non linéaire

Avec les notations du paragraphe précédent, le système discret associé au problème stationnaire s'écrit :

$$\left[\Phi_h(W_h)\right]_i = 0 \quad i = 1, 2, ..., i_{\max} \tag{6.43}$$

où i est l'indice du nœud courant, et W_h le vecteur qui rassemble les 4 inconnues du problème (en 2D), W_i, aux différents nœuds du maillage. Dans le cas d'une approximation de type « MUSCL », outre W_i, l'équation (6.43) fait explicitement intervenir seulement les inconnues W_j aux nœuds j voisins du nœud i ou voisins d'un voisin.

Nous venons de voir qu'une manière de résoudre ce système consistait à le plonger dans une formulation pseudo-instationnaire que l'on « relaxe » par intégration temporelle, par exemple par un schéma implicite de type « *Defect-Correction* ». Une alternative à cette approche consiste à « relaxer » encore, mais par une itération non linéaire, que l'on ne construit pas cette fois-ci à partir de la discrétisation consistante d'un problème d'évolution.

Par exemple, une « méthode de Gauss-Seidel non linéaire », consiste à effectuer un balayage des nœuds suivant une certaine numérotation ($i = 1, 2, ..., i_{\max}$). Au nœud i, on résout (6.43) par rapport aux inconnues W_i seulement, les autres, W_j ($j \neq i$), étant fixées aux valeurs les plus récemment réactualisées ; cette résolution locale des 4 équations peut se faire par une itération de type Newton ou quasi-Newton.

Noter que dans les applications à la mécanique des fluides, la convection joue souvent un rôle dominant, faisant apparaître dans l'écoulement des directions préférentielles de propagation de l'information. En conséquence, l'efficacité de cet algorithme peut être essentiellement conditionnée par le choix de la numérotation des nœuds.

Une itération de ce type constitue l'ingrédient de base de la méthode multigrille non linéaire « FAS » de la section suivante.

6.3. Traitement multigrille de problèmes non linéaires

Eludant pour l'instant le problème du choix d'une suite de grilles de densités différentes discrétisant le domaine, $\mathcal{M}_{h_0}$ (grille fine ; $h_0 = h$), $\mathcal{M}_{h_1}$,...,$\mathcal{M}_{h_k}$ (grille grossière), des schémas d'approximation applicables sur ces différentes grilles aux équations retenues (équations d'Euler ou de Navier-Stokes, en compressible, laminaire ou turbulent), ainsi que des opérateurs de transfert de grille-à-grille,

$$I_{h_\ell}^{h_m} \ : \ \mathcal{M}_{h_\ell} \longrightarrow \mathcal{M}_{h_m} \tag{6.44}$$

comment incorporer une stratégie multigrille dans la résolution numérique du système non linéaire des équations discrètes ?

Deux méthodes au moins se sont révélées efficaces pour réaliser cet objectif :

Méthode multigrille linéaire

Dans cette approche, on calcule un écoulement permanent par intégration pseudo-temporelle de la formulation instationnaire au moyen d'un schéma implicite linéarisé de type « *Defect-Correction* » (cf. paragraphe 6.2.2) (méthode d'Euler implicite). A chaque pas de temps, on inverse le système linéaire par une méthode multigrille. Dans la littérature anglophone, ce type de méthode porte le nom de « Correction Scheme », et son extension au cas non linéaire de « Newton-MG ».

Méthode multigrille non linéaire

Dans cette approche, on applique un algorithme dû à Brandt [20] connu sous le nom de « *Full-Approximation Scheme* » (« FAS »). Pour cela, on définit sur les différentes grilles, $\mathcal{M}_{h_\ell}$ ($\ell = 0, 1, 2, ...$), un ensemble de discrétisations compatibles non linéaires des équations stationnaires correspondant à la modélisation retenue :

$$\Phi_{h_\ell}\left(W_{h_\ell}\right) = 0 \tag{6.45}$$

L'algorithme consiste à résoudre partiellement sur un niveau de grille donné, $\mathcal{M}_{h_\ell}$, le système algébrique suivant par relaxation non linéaire (cf. paragraphe 6.2.3) :

$$\Phi_{h_\ell}\left(W_{h_\ell}\right) = \mathcal{S}_{h_\ell} \tag{6.46}$$

dans lequel le terme source est défini comme nul sur la grille fine,

$$\mathcal{S}_{h_0} = 0 \quad \text{(grille fine)} \tag{6.47}$$

et autrement ($\ell \geq 1$) comme suit :

$$\mathcal{S}_{h_\ell} \overset{\text{déf}}{=} \Phi_{h_\ell}\left(W_{h_\ell}^{(0)}\right) - \overline{I}_{h_{\ell-1}}^{h_\ell}\left(\Phi_{h_{\ell-1}}\left(W_{h_{\ell-1}}\right) - \mathcal{S}_{h_{\ell-1}}\right) \tag{6.48}$$

où :

$W_{h_{\ell-1}}$ est l'approximation des inconnues sur la grille de densité immédiatement supérieure $\mathcal{M}_{h_{\ell-1}}$, à l'issue de la relaxation non linéaire conduite précédemment sur cette grille dans la phase descendante du cycle multigrille ;

$W_{h_\ell}^{(0)} = I_{h_{\ell-1}}^{h_\ell} W_{h_{\ell-1}}$ est la restriction de l'estimation précédente des inconnues, servant de condition initiale à la nouvelle relaxation non linéaire.

Dans la phase ascendante du cycle multigrille, on transfère des corrections par interpolation. Lorsque les prolongements ne sont pas suivis de phase de lissage, on obtient un cycle « en dent de scie » (voir par exemple [73]).

Noter que les opérateurs de transfert ne sont pas les mêmes suivant qu'ils opèrent sur le vecteur des inconnues (ou des corrections), associées aux nœuds du maillage, ou sur les résidus (les flux) approchés aux interfaces, c'est-à-dire sur les segments, ce qui justifie la notation légèrement différente.

REMARQUE : l'algorithme FAS peut être considéré comme un algorithme de type « *Defect-Correction* » dans lequel l'approximation de bas niveau utilisée comme préconditionneur de l'approximation de grille fine est constituée de l'ensemble des opérations de relaxation et de transfert effectuées sur les grilles de moindre densité. Ces différents algorithmes sont donc intimement proches et peuvent se combiner de différentes manières. (Voir en particulier [73] et [47].)

6.4. Multigrilles en maillages non structurés

Dans cette section on se concentre sur les problèmes liés à la construction d'une hiérarchie de grilles de finesses différentes lorsque le « solveur de base » s'appuie sur une discrétisation non structurée quelconque du domaine.

On présente à titre d'illustration, certaines applications à la mécanique des fluides réalisées au sein du Projet SINUS. Afin de mieux comparer les algorithmes numériques entre eux, on a choisi de les évaluer dans le même cas d'écoulement transsonique eulérien stationnaire (nombre de Mach en amont, $M_\infty = 0.72$, incidence nulle) autour d'une géométrie classique de profil d'aile NACA 0012. Dans ce cas, l'écoulement admet un choc attaché au bord de fuite.

La résolution globale est réalisée par la « méthode multigrille non linéaire complète »[3] qui est constituée d'une itération monogrille non linéaire sur une grille grossière, suivie d'un prolongement de la solution sur une grille plus fine, puis d'une itération bigrille non linéaire de type FAS, suivie d'un prolongement de la solution sur une grille encore plus fine, puis d'une itération trigrille non linéaire de type FAS, etc. Théoriquement, si les cycles monogrilles, bigrilles, trigrilles, etc. étaient chacun interrompus lorsque l'erreur itérative atteignait juste le niveau de l'erreur d'approximation du niveau de grille-fine correspondant, et si les densités des grilles successives étaient dans un rapport donné, le coût global serait proportionnel au coût de l'ultime phase multigrille, celle qui implique toutes les grilles utilisées. Pour cette raison, on se contente de fournir comme paramètre indicateur du coût global le nombre observé de cycles de l'ultime phase multigrille nécessaire à la satisfaction d'un critère d'arrêt, atteint lorsque la norme du résidu a au moins été réduite d'un facteur égal à une tolérance prescrite (typiquement égale à 10^{-4}).

L'algorithme de relaxation non linéaire des systèmes algébriques (6.46) associés aux différents niveaux de grille consiste ici en l'intégration des systèmes pseudo-instationnaires suivants :

$$\left(W_{h_\ell}\right)_t + \Phi_{h_\ell}\left(W_{h_\ell}\right) - \mathcal{S}_{h_\ell} = 0 \quad (\ell = 0, 1, 2, \ldots) \tag{6.49}$$

sur un nombre donné, souvent égal à 1, de pas de temps. Cette intégration est conduite par une méthode de Runge-Kutta à 4 pas intermédiaires ajustés pour optimiser les propriétés de lissage [57] [59]. Dans sa thèse, E. Morano [73] a aussi évalué comme lisseur la méthode d'Euler implicite linéarisée en formulation « *Defect-Correction* » définie à la section précédente.

Les divers algorithmes diffèrent principalement par la construction des niveaux de grille et des opérateurs de transfert.

3. *Full Multi-Grid method* (« FMG ») en anglais.

6.4.1. Multitriangulation

La manière la plus simple de construire une hiérarchie de grilles emboîtées *de même type* consiste à d'abord choisir une grille grossière, par exemple en 2D une triangulation du domaine, puis à diviser chaque élément ; par exemple, chaque triangle en 4. Le niveau de grille suivant est obtenu par application de la même technique de division des éléments, et ainsi de suite jusqu'à l'obtention d'une grille de la finesse souhaitée. Quelle que soit sa nature, le type d'approximation par éléments finis retenue, en particulier une approximation d'ordre deux, peut dans ce cas être appliqué à l'identique sur chaque niveau.

Cette approche est illustrée aux figures 6.3 et 6.4 (a) tirées de la thèse de M.H. Lallemand [57] [58], qui représentent des zooms de triangulations emboîtées en l'occurrence assez grossières, du domaine. Ces maillages ont respectivement 121, 442 et 1 684 nœuds.

Dans une première expérience numérique, on utilise sur chaque niveau de grille, le même type d'approximation au second ordre des équations d'Euler. Pour les transferts, les valeurs nodales de l'inconnue sont restreintes par simple injection (nœuds communs aux 2 niveaux), $I^{h_\ell}_{h_{\ell-1}}$, et par moyenne pondérée par les aires pour les résidus, $\overline{I}^{h_\ell}_{h_{\ell-1}}$, ce qui revient à calculer une intégrale de surface de manière consistante. Les prolongements se font par interpolation P1-Lagrange. Une réduction du résidu de grille-fine de 10^{-4} est réalisée par 40 cycles trigrilles (de la phase ultime de la méthode complète) [57].

Les approximations associées aux grilles moyenne et grossière n'affectant pas la solution convergée associée à la grille fine, on peut simplifier l'algorithme précédent en construisant l'approximation seulement à l'ordre 1 sur les niveaux inférieurs. Il en résulte une moins bonne compatibilité entre les approximations de grille fine et ici de grille moyenne, et en conséquence, la nécessité de prolonger la phase ultime trigrille plus longuement (80 cycles) pour satisfaire le même critère d'arrêt [57].

Malheureusement, cette approche admet de sérieuses limitations. En effet, à mesure que l'on construit ainsi des grilles de plus en plus fines, celles-ci deviennent en quelque sorte de moins en moins « non structurées ». Si le nombre de niveaux est grand on aboutit à une grille fine présentant des « macro-éléments » structurés, et ceci est en contradiction profonde avec les options adoptées pour le solveur de base. En réalité, dans les grands problèmes de calcul scientifique industriels, l'option « maillage non structuré » a pour objet d'une part de réussir à mailler des geométries complexes (ce qui en soi est un problème difficile) et d'autre part, d'y parvenir « économiquement » à la fois en heures d'ingénierie, et en nombre d'éléments du maillage produit. La règle du jeu naturelle est donc de supposer qu'un *maillage de finesse maximale* est fourni au responsable de la résolution des équations et qu'il lui appartient de construire uni-

quement les grilles grossières en conséquence. Une exception à cette règle se produit lorsqu'en résolvant on constate la nécessité de raffiner le maillage pour l'adapter localement à la solution (voir plus loin les paragraphes sur les triangulations non emboîtées et les multigrilles adaptatives).

Dans l'impossibilité de construire une hiérarchie de grilles emboîtées à partir d'une grille fine arbitrairement non structurée, on peut ou bien construire des grilles grossières par agglomération de cellules de volumes finis (maillages duaux), ou bien reconstruire des maillages non emboîtés (par exemple par triangulation de Delaunay) et adapter l'algorithme multigrille en conséquence.

6.4.2. Agglomération

Pour fixer les idées, plaçons nous dans le cas d'un calcul en 2 dimensions d'espace lorsque le maillage fin donné est une triangulation arbitraire. On trace les médianes de chaque triangle, ce qui permet d'identifier autour de chaque nœud une cellule de contrôle. Ces cellules constituent le « maillage dual fin ». On a vu qu'une approximation de type « volumes finis » est construite par bilan de flux aux interfaces, c'est-à-dire sur les arêtes de la cellule, et l'ordre deux est atteint en calculant ces flux après une extrapolation de type MUSCL ([89]). En mécanique des fluides, ce type de construction a le très grand mérite dans le cas hyperbolique des équations d'Euler de permettre la satisfaction de la contrainte de « conservativité » de manière automatique. Un maillage moins fin de volumes finis peut ensuite être construit par agglomération des cellules par paquets de 4 ; différents algorithmes sont possibles (voir e.g. [57]). L'algorithme se répète pour construire les niveaux suivants de grilles grossières.

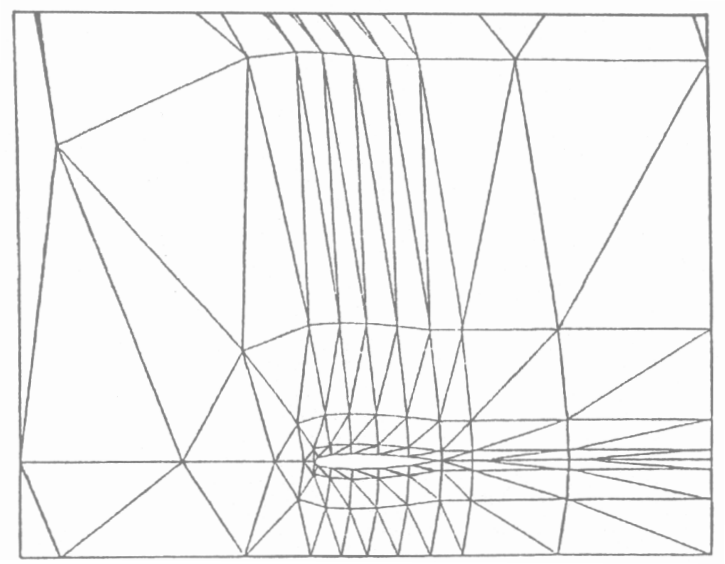
Une illustration de cette approche est fournie aux figures 6.4 (b) et 6.5 qui montrent des zooms du maillage dual de la triangulation fine précédente et des maillages « volumes finis » obtenus par deux agglomérations successives des cellules 4 à 4. Ces maillages contiennent respectivement 1 684, 411 et 111 cellules.

Les restrictions de résidu se font par sommation de flux, et les prolongements par affectation de la valeur moyenne associée à une macro-cellule aux cellules qui la constituent. Sur les maillages inférieurs, l'approximation est d'ordre 1.

En reprenant le cas-test précédent d'écoulement transsonique, on constate que le critère d'arrêt est satisfait par la méthode multigrille complète qui résulte de ces choix, avec le même nombre de cycles (80) dans la phase ultime trigrille que dans le cas de multitriangulation lorsque seule l'approximation de grille fine est du second ordre.

Une certaine faiblesse de cette approche réside dans la difficulté qui en résulte à construire des approximations d'ordre deux sur les maillages grossiers duaux. En effet, le bord d'une cellule d'un maillage grossier n'est plus l'interface entre 2 nœuds (à moins d'utiliser des techniques complexes de reconstruction) et l'extrapolation

(a)



(b)

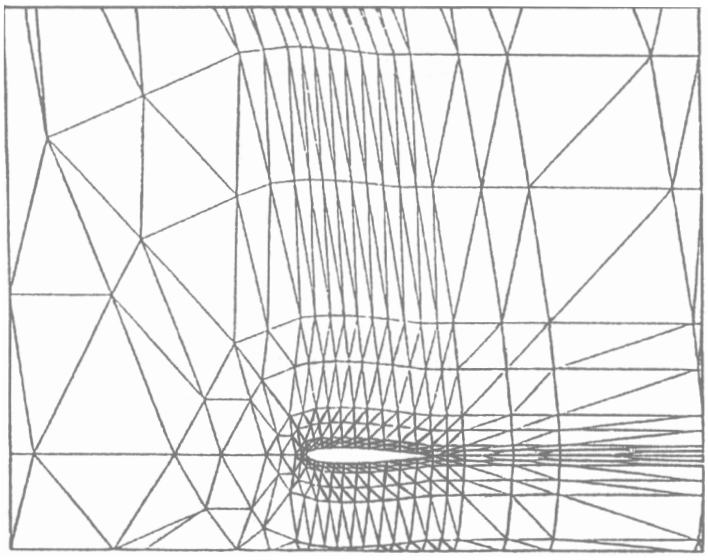
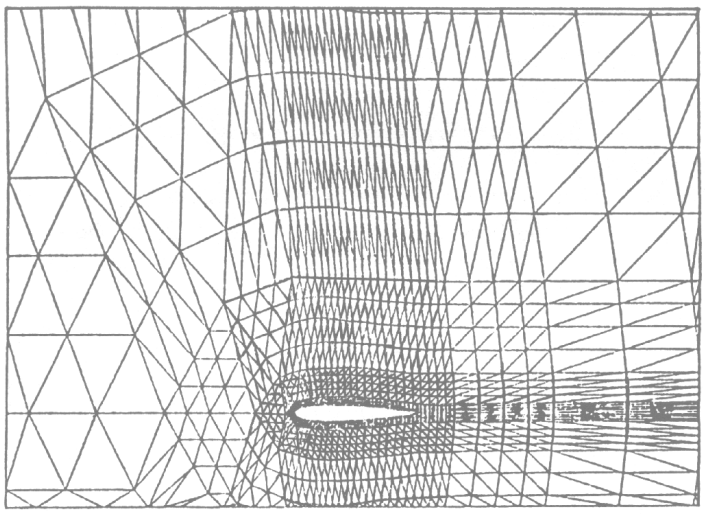


Figure 6.3. Zooms de triangulations grossière, (a) (121 nœuds), et moyenne, (b) (442 nœuds) autour d'une géométrie de profil d'aile NACA 0012

(a)



(b)

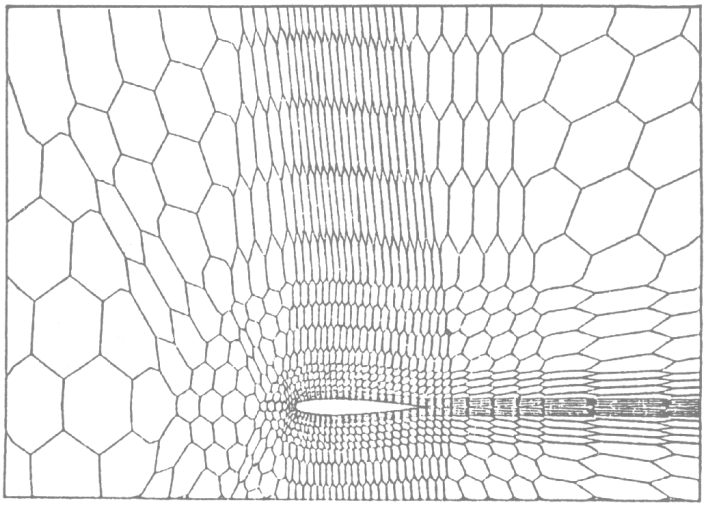


Figure 6.4. Zooms de triangulations fine, (a) (1684 nœuds), et de maillage dual, (b) (1684 cellules), autour d'une géométrie de profil d'aile NACA 0012

MUSCL n'est pas applicable directement. On vient de voir que le léger défaut de compatibilité qui en résulte entre les approximations associées à la grille fine et la plus dense des autres, entraîne une certaine perte en efficacité algorithmique de la méthode multigrille.

Le cas test précédent a été refait (cf. [57] ou [58]) en partant cette fois-ci d'une triangulation fine « vraiment non structurée » à 800 nœuds générée par une méthode dite frontale (figure 6.6 (a)). Les maillages de volumes finis successifs sont représentés aux figures 6.6 (b) et 6.7.

La figure 6.8 présente les courbes de convergence du résidu de différents algorithmes de résolution.

Lorsque l'approximation est du 1er ordre sur tous les niveaux, les phases monogrille, bigrille et trigrille ont des vitesses de convergence très proches. Par contre, dans le cas d'une approximation d'ordre 2 (sur la grille fine seulement), la phase trigrille doit être prolongée afin de satisfaire le même critère d'arrêt.

Soulignons en conclusion que cette méthode se révèle, malgré cette nuance, très efficace. Voir notamment : [57], [58], [52], [53], [73], [28], [29], [40], [41], etc.

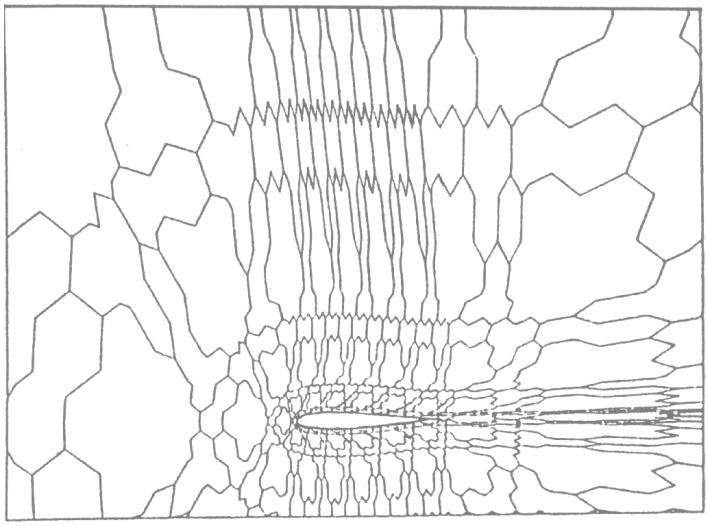
6.4.3. Triangulations non emboîtées

Une alternative à l'agglomération consiste à construire une hiérarchie de grilles dont les éléments ne sont pas emboîtés, mais qui s'appuient sur des sous-ensembles emboîtés de l'ensemble initial des nœuds du maillage fin. Par exemple dans le cas bidimensionnel, on peut utiliser des triangulations de Delaunay [43].

Ce type d'approche comporte deux difficultés principales. La première est la nécessité de définir un algorithme d'appauvrissement du nuage de points qui produise les densités souhaitables de nœuds sur les grilles successives, et, plus difficile encore, qui respecte la géométrie du domaine, ses singularités, par exemple ses éventuelles arêtes vives ; ce problème admet néanmoins quelques solutions maîtrisées (voir par exemple [15] [16]). La deuxième difficulté est de construire des opérateurs de transfert de grille à grille suffisamment précis, pour que les approximations associées aux différents niveaux soient cohérentes entre elles.

Par exemple, la figure 6.9 représentent des zooms de triangulations de Delaunay du domaine obtenues par applications successives d'un algorithme d'appauvrissement dû à H. Guillard [44]. Dans cet algorithme, on part de la liste des nœuds dans un certain ordre ; on retient le premier nœud, mais on élimine ses voisins ; puis, on retient le premier nœud suivant restant dans la liste et on élimine ses voisins (s'ils n'ont pas déjà été éliminés) ; et ainsi de suite jusqu'à épuration de la liste. L'algorithme est en fait amendé de manière à respecter au mieux la géométrie initiale du domaine. Dans cet

(a)



(b)

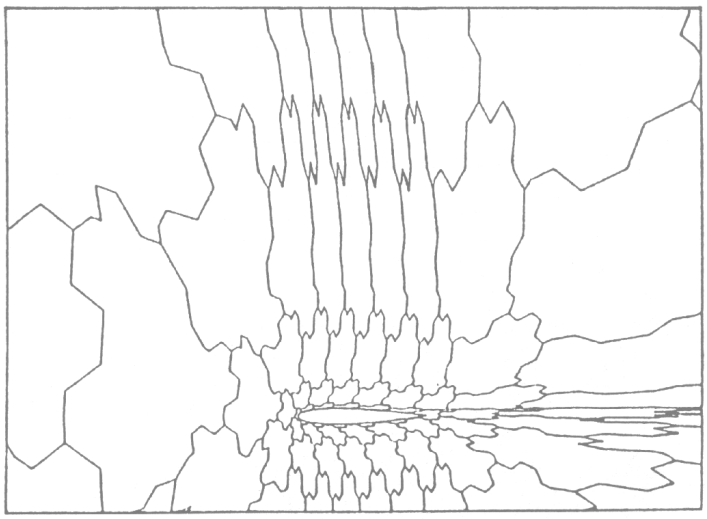
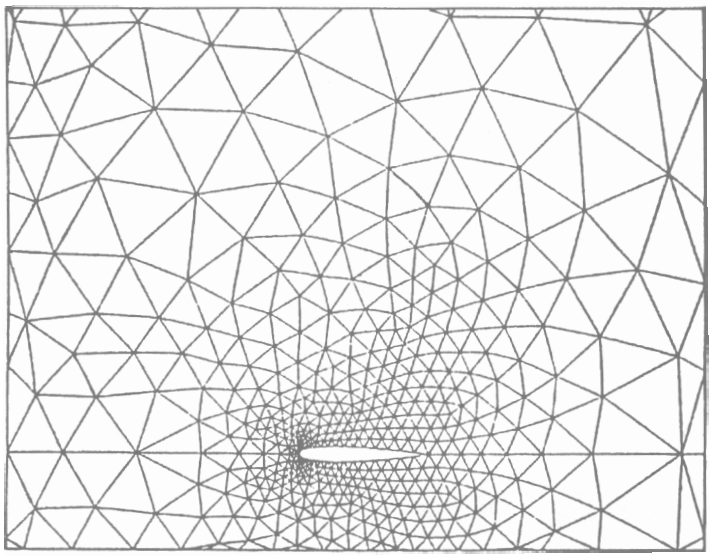


Figure 6.5. Zooms de maillages de volumes finis agglomérés moyen (411 cellules) et grossier (111 cellules) autour d'une géométrie de profil d'aile NACA 0012

(a)



(b)

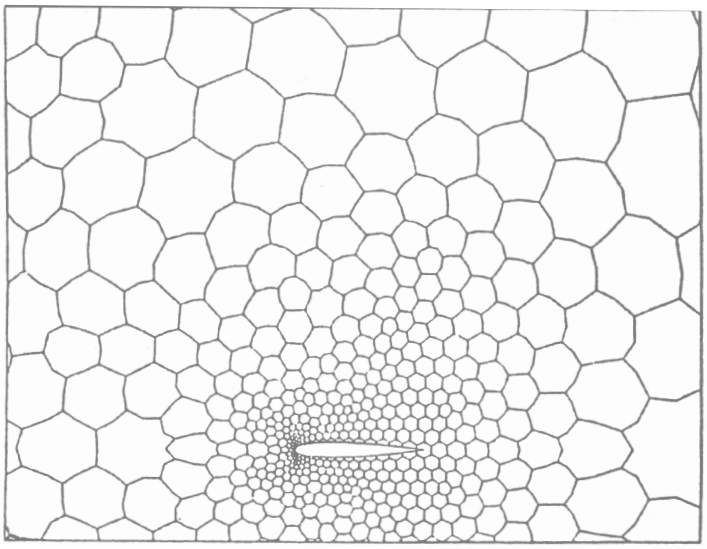
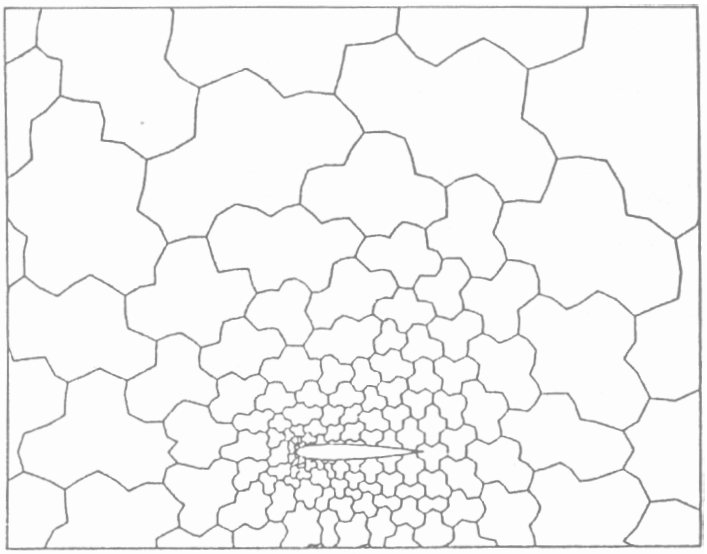


Figure 6.6. Zooms de triangulation fine non structurée autour d'une géométrie de profil d'aile NACA 0012 et du maillage dual

(a)



(b)

Figure 6.7. Zooms de maillages de volumes finis agglomérés moyen, (a), et grossier, (b), autour d'une géométrie de profil d'aile NACA 0012

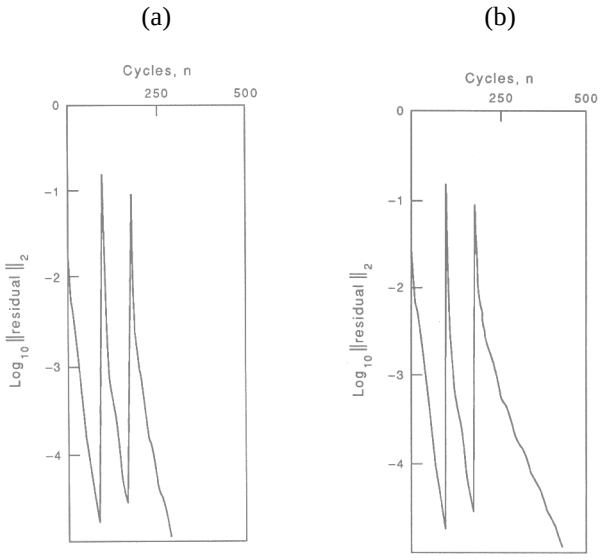


Figure 6.8. Courbes de convergence de la méthode multigrille complète non linéaire : (a) approximation du 1er ordre, (b) approximation (de grille fine) du 2e ordre

exemple, les triangulations successives ont respectivement 800, 223, 67 et 19 nœuds.

E. Morano dans sa thèse [73] a refait le calcul d'écoulement transsonique précédent par la méthode multigrille non linéaire complète en utilisant 6 niveaux de grille constitués des 4 maillages de Delaunay précédents complétés par 2 maillages de 3 114 et 12 284 nœuds respectivement ; ces maillages sont représentés à la figure 6.10, et ont été obtenus par 2 raffinements successifs par division des éléments à partir du maillage fin précédent. La figure 6.11 démontre la grande efficacité de la méthode, lorsque le nombre de niveaux de grille est grand. La convergence a quasiment la même qualité pour les cycles bigrilles, trigrilles, etc. que dans l'ultime phase 6-grilles. Selon les auteurs de [74], ce calcul a été réalisé par la méthode multigrille complète en 42 « unités de travail », c'est-à-dire 42 évaluations du flux Φ_h des équations d'Euler discrétisées. On peut estimer que la méthode de Runge-Kutta 4 appliquée seule sur la grille fine aurait atteint un niveau de convergence itérative comparable en un millier d'itérations, pour un coût environ deux ordres de grandeur supérieur. En comparaison à une méthode d'intégration pseudo-temporelle implicite de type « *Defect-Correction* », le gain en efficacité serait moindre, néanmoins de l'ordre de plusieurs dizaines. La figure 6.12 fournit les lignes de niveaux du nombre de Mach de la solution calculée mettant en évidence un choc symétrique attaché au bord de fuite.

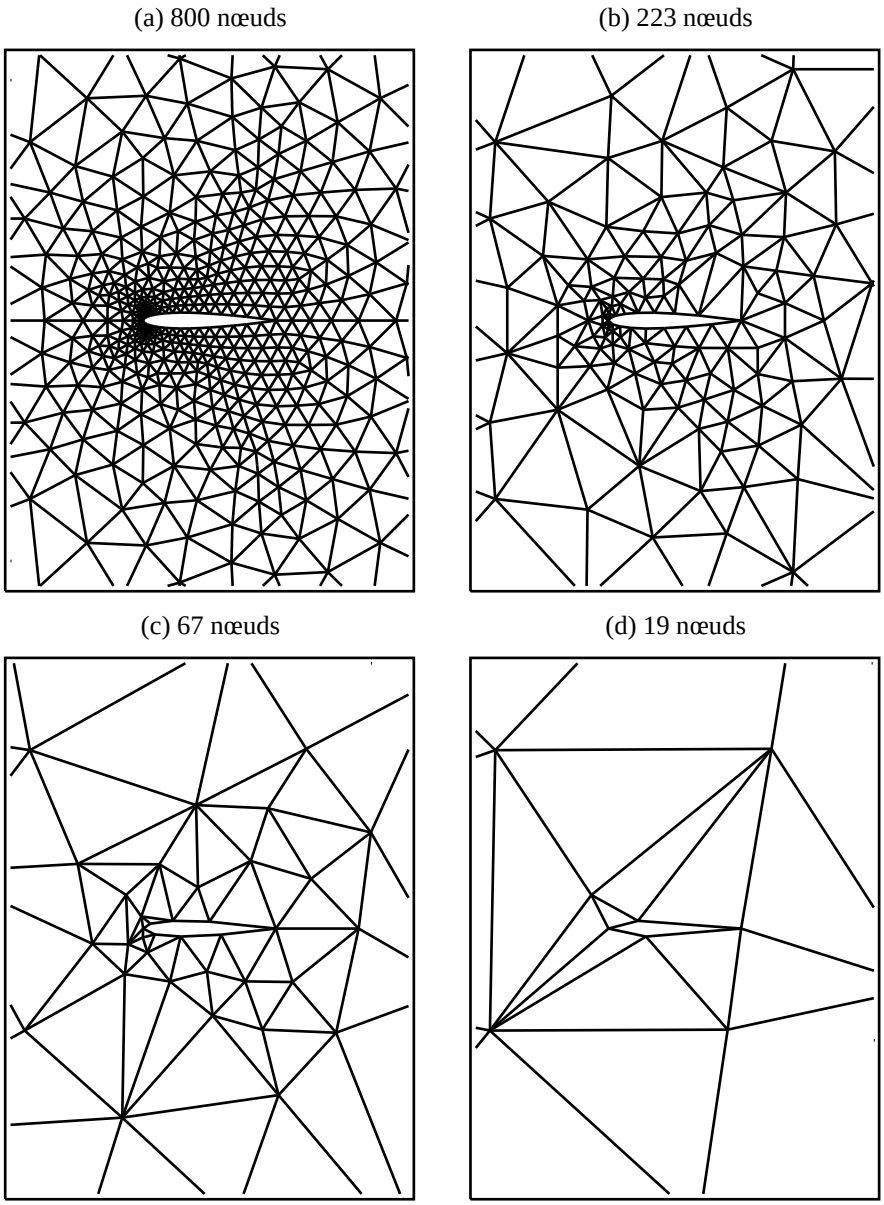


Figure 6.9. Zooms de triangulations de Delaunay associées à des sous-ensembles emboîtés de nœuds

(a) 3 114 nœuds (b) 12 284 nœuds

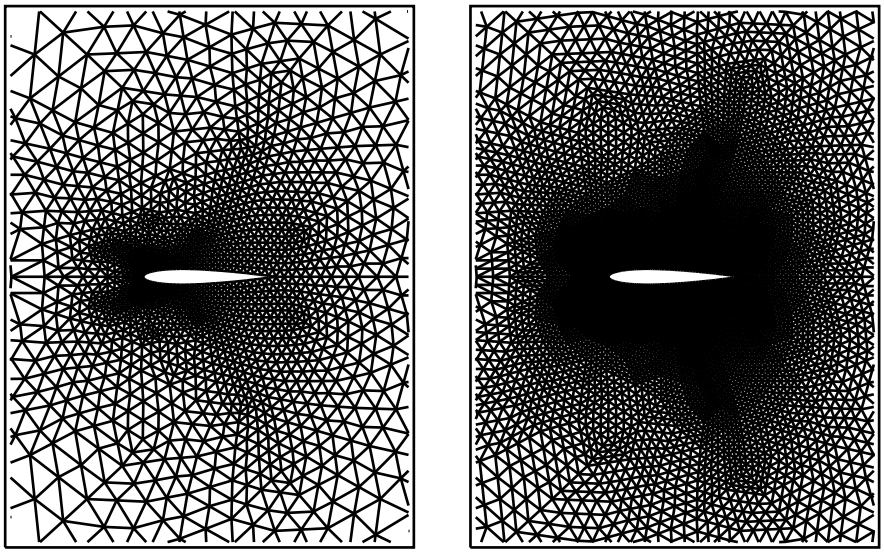


Figure 6.10. Maillages fins autour d'une géométrie de profil d'aile NACA 0012

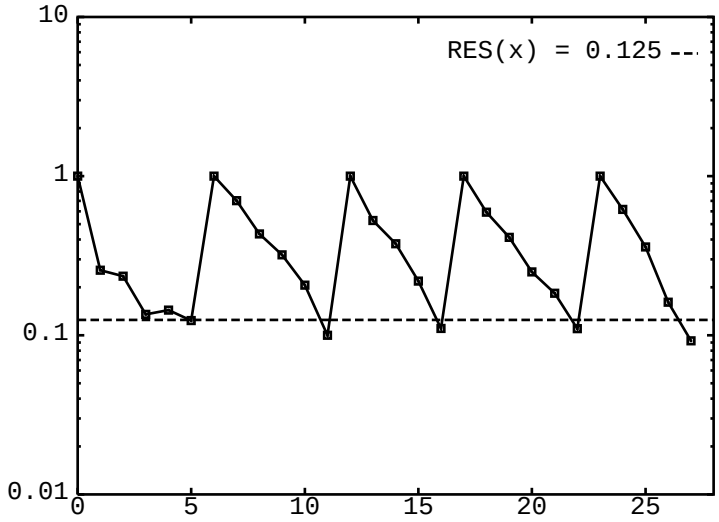
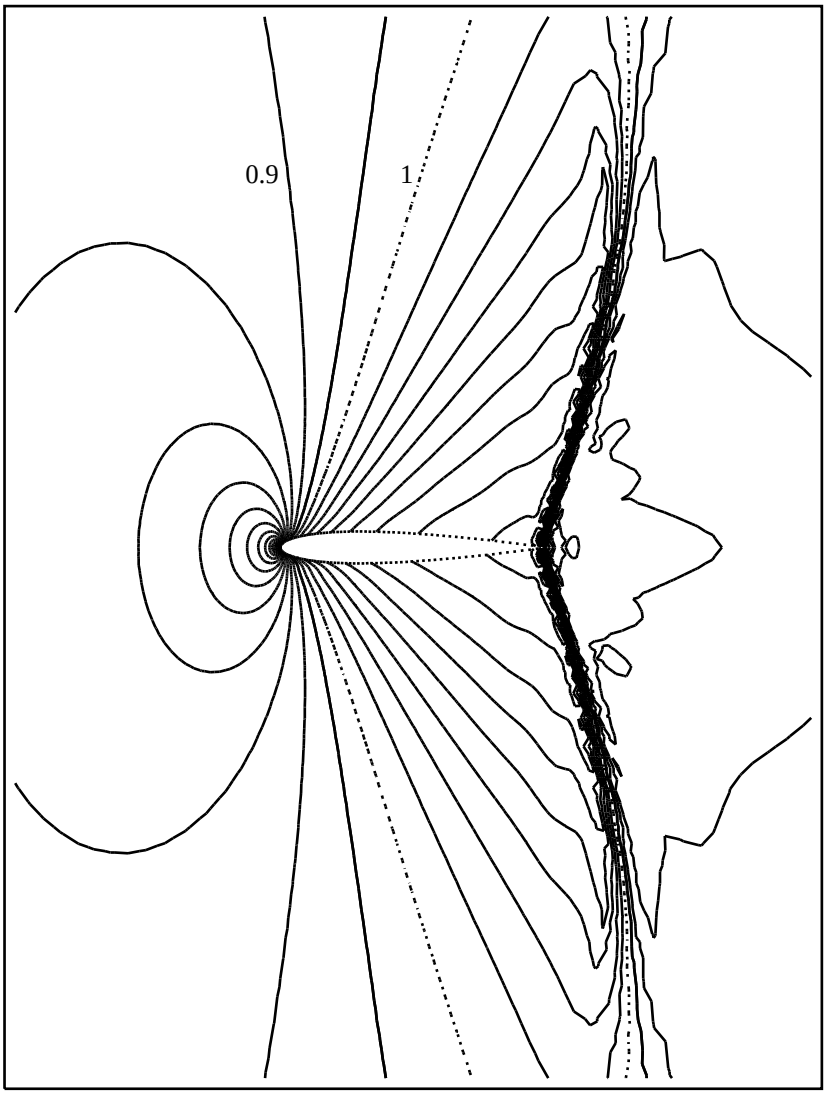


Figure 6.11. Convergence de la méthode multigrille non linéaire avec 6 niveaux de grille



LIGNES ISOMACH
Intervalle entre isovaleurs : 0.50000E-01
Min./ Max. Mach 0.45549E-01 , 1.5196

Figure 6.12. Lignes de niveaux du nombre de Mach dans l'écoulement transsonique autour d'une géométrie de profil d'aile NACA 0012 ($M_\infty = 0.72\,;\ \alpha = 0^o$).

6.4.4. Mailles étirées, semi-déraffinement, multigrilles adaptatives

Une difficulté majeure des calculs en mécanique des fluides est la présence potentielle dans la solution de structures physiques de dimension inférieure à la dimension d'espace. Il s'agit principalement des couches limites visqueuses près des parois, et des chocs. La présence de ces structures amène le praticien à utiliser des maillages localement raffinés de manière non isotrope, par adaptation statique ou dynamique. En particulier, il n'est pas rare en aérodynamique industrielle que le nombre de Reynolds Re soit de l'ordre de 10^6, et comme l'épaisseur de couche limite varie comme $Re^{-1/2}$, on peut être amené à utiliser près d'une paroi des mailles très étirées, ayant des facteurs de forme de plusieurs milliers. Si on applique sans modification particulière un algorithme de déraffinement isotrope pour construire les maillages grossiers, on risque de détruire le caractère adapté de la grille fine. Les approximations de grilles grossières ne pourront alors servir de « bons préconditionneurs » à l'approximation de grille fine, et on peut prévoir une performance médiocre de l'algorithme multigrille.

Dans le cas d'un choc, la difficulté peut être de nature différente selon la procédure utilisée pour raffiner. Notons que dans ce cas, en général, ni la localisation, ni l'intensité de la singularité n'est connue *a priori*.

Pour remédier au problème des mailles étirées, on peut construire les grilles grossières par une technique de semi-déraffinement local. Pour en comprendre le mécanisme, revenons momentanément au problème modèle d'un opérateur elliptique linéaire à coefficients constants mais très différents en grandeur, dont on a établi le spectre discret,

$$\underline{\lambda}^{h}_{m,\ell} = \underline{\lambda}^{h_x}_{m} + \underline{\lambda}^{h_y}_{\ell} = \frac{2 - 2\,\cos\theta_{x\,m}}{h_x^2} + \frac{2 - 2\,\cos\theta_{y_\ell}}{h_y^2} \tag{6.50}$$

au paragraphe 3.5.2. Dans un plan $(\underline{\lambda}^{h_x}_{m}, \underline{\lambda}^{h_y}_{\ell})$, les modes fréquentiels sont représentés dans un rectangle très allongé verticalement si le maillage est tassé en y (voir figure 6.13). Dans ce plan, les lignes de niveaux de la valeur propre générique $\underline{\lambda}^{h}_{m,\ell}$ sont les parallèles à la deuxième bissectrice des axes. Dans la méthode bigrille idéale classique dans laquelle on déraffine d'un facteur 2 en x comme en y, on identifie les basses fréquences comme celles occupant le rectangle homothétique de rapport $\frac{1}{2}$. On constate alors qu'à cause de l'étirement, les valeurs propres associées aux hautes fréquences seules occupent l'intervalle $\left[2, 4 + 4/\varepsilon^2\right]$ c'est-à-dire la presque totalité de l'intervalle de variation complet. Dans ce cas, un lisseur classique sera très inefficace, d'où la nécessité d'itérer plus longuement sur la grille fine, ou de subir une dégration de performance de l'algorithme bigrille. Par contre, si on déraffine dans la direction de y seulement, on est quasiment ramené à une situation unidimensionnelle, les valeurs propres associées aux hautes fréquences occupant à peine plus de 50 % de l'intervalle (voir figure 6.13 (b)). Bien évidemment, la performance est maintenue au prix de la résolution d'un plus grand nombre d'équations sur la grille grossière, ce qui n'altère

pas l'évaluation de la « complexité » de l'algorithme seulement si la procédure de semi-déraffinement est appliquée de manière très locale. Il s'agit là d'un sujet d'étude encore ouvert.

Par exemple dans la méthode d'agglomération précédente, J. Francescatto dans sa thèse [40] évalue un facteur de forme local pour chaque cellule. L'algorithme d'agglomération de 4 cellules est appliqué seulement si le facteur de forme est proche de 1 (disons entre $\frac{1}{2}$ et 2) ; sinon, on identifie une direction principale d'étirement en calculant pour chaque arête reliant le nœud courant i à un voisin j une quantité scalaire mesurant l'intensité de la connexion ij ; puis on agglomère seulement la cellule $\mathcal{C}_i$ à la cellule $\mathcal{C}_j$ correspondant à la direction principale d'étirement. Une illustration de cette méthode est fournie par la figure 6.14 tirée de [40]. En (a) on y visualise un maillage dual fin discrétisant un domaine carré dont les mailles dans la direction verticale, ont été tassées à proximité du bord inférieur pour prendre en compte une couche limite, et à l'inverse allongées à proximité du bord opposé. Le maillage grossier correspondant est indiqué en (b) ; il met en évidence un semi-déraffinement par agglomération des cellules 2 à 2 à proximité de ces deux bords, mais chacun dans une direction opposée de l'autre ; entre ces deux bords, les cellules sont agglomérées 4 à 4 de manière isotrope grâce à une transition automatique.

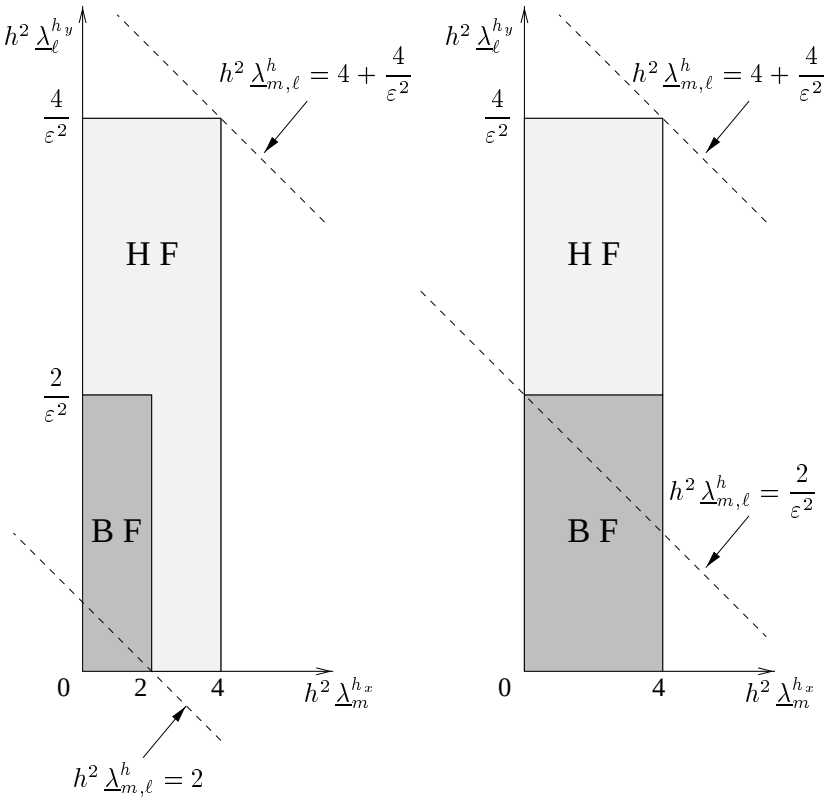
Notons qu'il existe d'autres situations où le passage d'une grille à la suivante résulte d'un déraffinement qui n'est pas local, mais zonal. On se réfère ici aux « multigrilles adaptatives » dans lesquelles un maillage fin peut différer du maillage grossier dans une zone du domaine seulement ; par exemple, dans une zone où la viscosité joue vraiment un rôle. Ce type d'approche a été notamment étudié dans le cadre d'approximations éléments finis/volumes finis du type de la section précédente par F. Angrand et P. Leyland (voir par exemple [5] [6]).

Ce type d'approche proposé par S. McCormick et J. Thomas d'abord pour des équations elliptiques sous le nom de « méthode FAC » (Fast Adaptive Composite method) ([70] [71]) constitue une passerelle entre les méthodes multigrilles proprement dites et les techniques de coordination de domaines introduites au chapitre 7 et connaît depuis quelques années d'intéressants développements notamment pour les équations paraboliques au sein de la communauté numéricienne française (cf. en particulier [17] [78] [50] [30]).

6.4.5. Impact de termes hyperboliques dominants

Dans les applications à la mécanique des fluides, qu'on résolve les équations d'Euler ou de Navier-Stokes, les termes de convection jouent un rôle prépondérant. Dans le cadre des multigrilles, il est donc assez naturel de chercher à contruire des opérateurs de transfert de grille à grille qui tiennent compte des directions préférentielles locales de l'écoulement. Cette question a été étudiée notamment par M.P. Leclercq dans sa thèse ([59] [60]).

(a) déraffinement isotrope (b) semi-déraffinement

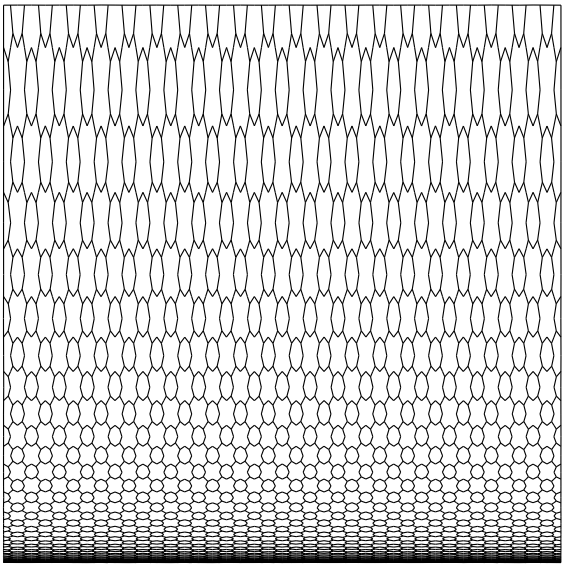


$$h_x = \Delta x = h\,, \quad h_y = \Delta y = \varepsilon\, h\,, \quad \varepsilon \ll 1$$

$$\underline{\lambda}_{m,\ell}^h = \underline{\lambda}_m^{h_x} + \underline{\lambda}_\ell^{h_y} = \frac{2 - 2\,\cos\theta_{x\,m}}{h_x^2} + \frac{2 - 2\,\cos\theta_{y_\ell}}{h_y^2}$$

Figure 6.13. Domaine des paramètres de fréquence dans un cas de forte anisotropie

(a) maillage fin de volumes finis



(b) maillage « semi-déraffiné »

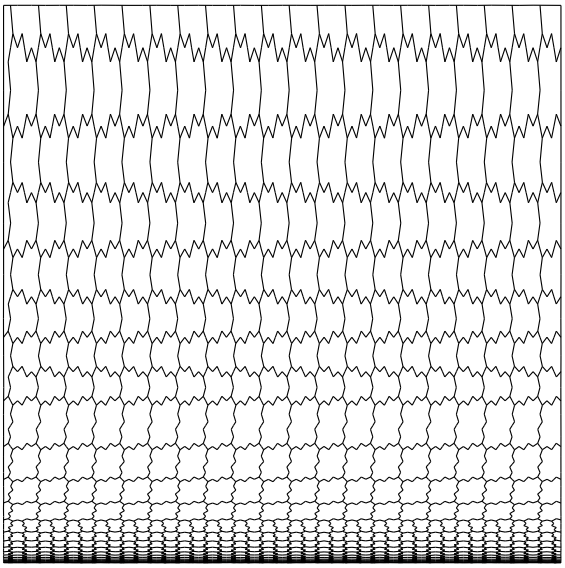


Figure 6.14. Agglomération non isotrope de cellules de volumes finis

6.4.6. Multigrilles algébriques

Dans une méthode multiniveau ou multigrille algébrique, on construit une hiérarchie d'approximations qui ne sont pas nécessairement les discrétisations d'une EDP sur différents niveaux de grille, mais simplement des systèmes algébriques de moindres degrés de liberté sur les niveaux inférieurs. Les connexions ne sont plus entre les nœuds d'un maillage mais entre des variables plus ou moins fortement liées par les coefficients d'influence $\{a_{j,k}\}$ que représentent les coefficients du système algébrique à résoudre. Se posent alors de nombreuses questions dont celles de la définition abstraite du système discret réduit et de la construction du lisseur. Il est notamment possible de définir un concept d'agglomération « algébrique » consistant à sommer les équations 2 à 2 (sommation des lignes) et en rassemblant les degrés de liberté 2 à 2 (sommation des colonnes). La justification théorique de ce type d'approche reste encore du domaine de la recherche (cf. notamment [21] [22] [81]).

6.5. Quelques remarques finales

Certaines manifestations scientifiques internationales sont entièrement consacrées au thème des multigrilles. C'est le cas en particulier de la conférence de Copper Mountain, Colorado (*Copper Mountain Conference on Multigrid Methods*) qui se tient les années impaires, en alternance avec une conférence sur un thème voisin (*Copper Mountain Conference on Iterative Methods*), ainsi que de certains ateliers (tels que les « *GAMM workshops on parallel multigrid* »). Pour de plus amples informations sur ces évènements scientifiques, on pourra consulter le serveur suivant :

`http://www.cerfacs.fr/~douglas/mgnet.html`

Par ailleurs, on recommande spécialement la lecture de l'ouvrage de H. Guillard [45], membre du Projet SINUS, pour une présentation complémentaire des méthodes multigrilles.

D'une manière générale, les principales questions que se posent le théoricien ou le praticien sont les suivantes :

– Comment construire la hiérarchie des niveaux de grille suivant le type d'approximation (éléments Finis en maillage structuré ou non, méthode spectrale, multigrille géométrique ou algébrique, etc.) ?

– Comment définir les opérateurs de grille grossière ? (quels opérateurs de transfert de grille à grille en particulier?)

– Comment traiter les problèmes non linéaires ?

– Comment construire des lisseurs adaptés aux discrétisations retenues ?

Quelques illustrations de ces problématiques en mécanique des fluides sont fournies par les actes d'un atelier ERCIM [8].

Notons que si la théorie a aujourd'hui atteint un niveau satisfaisant pour le traitement de problèmes aux limites elliptiques [46] [18] [71] [95], le champ d'investigation reste encore largement ouvert en hyperbolique. En particulier, pour une équation hyperbolique du premier ordre, certains auteurs avancent que les méthodes multigrilles souffrent de limitations intrinsèques, notamment lorsque le schéma d'approximation est d'ordre élevé (voir [87]).

Remarquons enfin que le concept de « hiérarchie d'approximations », pivot d'efficacité quasi-optimale, peut être étendu à des domaines autres que la résolution stricte des systèmes discrets, en particulier à leur optimisation (voir notamment [12] et [68]).

Chapitre 7

Méthodes multidomaines

7.1. Introduction

On se place à nouveau dans le cas où, après avoir discrétisé un problème d'EDP sur un domaine ouvert Ω de frontière $\Gamma = \partial\Omega$, on aboutit à un grand système (pas nécessairement linéaire) de M équations algébriques à M inconnues,

$$\boxed{A_h\, u_h = f_h} \tag{7.1}$$

Pour la résolution de ce système, on suppose que l'on dispose d'une « méthode de base » qui peut être *directe* (par exemple par résolution formelle, élimination de Gauss, etc.) ou *itérative* (itérations de Jacobi, Gauss-Seidel, Richardson, annihilation, etc.). On fait l'hypothèse suivante : « *Le coût de la résolution du système par la méthode de base est asymptotiquement une fonction sur-linéaire du nombre N de degrés de liberté* ». On exclut donc en principe le cas où la méthode de base serait l'« algorithme multigrille complet ».

Supposons maintenant que l'on partitionne le domaine Ω en deux sous-domaines Ω_1, Ω_2 disjoints ayant une frontière commune γ dite « interface ». Si une information suffisamment riche sur $u_{h\,/\gamma}$ (ou ses dérivées partielles) était connue *a priori*, on pourrait résoudre le problème global par deux résolutions indépendantes sur Ω_1 et Ω_2 ce qui entraînerait une réduction du coût en vertu de l'hypothèse faite.

En pratique la nature de l'information, c'est-à-dire le jeu précis de conditions à l'interface qu'il faudrait connaître pour que les sous-problèmes restreints aux sous-domaines soient bien posés dépend de l'EDP considérée, éventuellement des valeurs

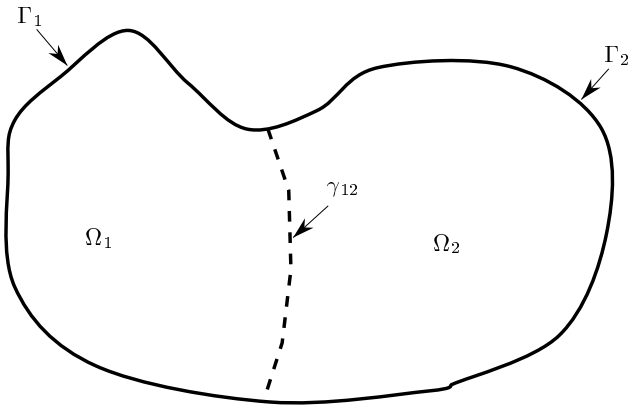


Figure 7.1. Partition de domaine

locales des inconnues dans le cas d'un problème non linéaire, et n'est généralement pas unique. Souvent, en particulier dans le cas de problème ayant une composante hyperbolique (propagation d'onde), une partie de l'information contenue dans l'inconnue se propage de Ω_1 vers Ω_2 et une partie complémentaire dans l'autre sens. Ce transfert d'information est en quelque sorte un moyen de communication entre les sous-domaines permettant aux solutions partielles $u_{h\,/\Omega_1}$ et $u_{h\,/\Omega_2}$ de se « coordonner », ou se « raccorder » pour composer une solution globale de la régularité voulue.

Une « méthode de résolution par partition de domaine » est un algorithme de résolution dans lequel après avoir partitionné le domaine de calcul en sous-domaines disjoints, on itère sur des conditions d'interface et on résout à chaque itération (« itération de coordination ») plusieurs problèmes, chacun restreint à un sous-domaine. La résolution sur les sous-domaines est effectuée par la méthode de base, et une fois raccordées les solutions partielles reconstituent la solution globale.

Une variante de cette approche consiste à décomposer le domaine Ω en sous-domaines Ω_1, Ω_2 se recouvrant partiellement. Souvent dans ce cas, on cherchera à minimiser une forme quadratique mesurant le « défaut de raccord » entre les deux solutions partielles (« Moindres Carrés »). On parle alors de « méthode par décomposition de domaine avec recouvrement ».

Dans ce cours introductif, on se propose d'étudier sommairement les principales méthodes par décomposition de domaine, appelées plus brièvement « méthodes multidomaines », dans le cas où la décomposition en partition ou recouvrement a pour but de réaliser une économie du coût global de résolution, le modèle mathématique

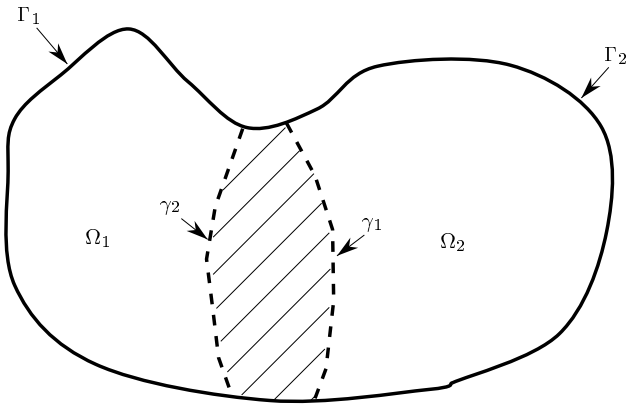


Figure 7.2. Décomposition de domaine avec recouvrement

discret étant le même partout dans le domaine global Ω.

Notons que ces méthodes sont *naturellement parallélisables* du fait que chaque sous-problème peut être résolu sur une unité de calcul propre. C'est pourquoi on observe aujourd'hui un intérêt accru pour ces techniques.

Bien que cet aspect ne soit pas traité ici, il convient de noter qu'il existe des méthodes conceptuellement voisines dans lesquelles on utilise une décomposition du domaine global dans le but d'utiliser des modèles physiques (c'est-à-dire mathématiques) différents dans chacun des sous-domaines. Dans ce cas, une économie en coût peut être réalisée par le fait que la résolution du modèle physique le plus complet (et donc le plus onéreux à résoudre) est restreinte à un sous-domaine où il est vraiment nécessaire. Par exemple, en mécanique des fluides, lorsqu'on calcule un écoulement autour d'un profil d'aile, ce qui constitue un problème « externe », c'est-à-dire un problème où le domaine de calcul est l'extérieur d'un domaine connexe, on peut restreindre la résolution des équations complètes de Navier-Stokes à une bande autour du profil, et se contenter de résoudre les équations d'Euler à l'extérieur de cette bande, sachant que les phénomènes visqueux y sont négligeables. Se pose alors le problème de l'identification de conditions d'interface pour lesquelles les sous-problèmes correspondants sont bien posés et se raccordent dans un certain sens. [1] On qualifiera ces méthodes d'« algorithmes de raccord de modèles ».

1. Les modèles mathématiques n'étant pas les mêmes dans les différents sous-domaines, les formulations variationnelles qui leur sont associées sont elles-mêmes distinctes. En particulier, les solutions partielles (au sens faible) n'appartiennent pas nécessairement à un même espace. Par conséquent, définir la régularité du raccord est une difficulté intrinsèque à ce type d'algorithme de coordination.

7.2. La méthode de Schwarz

Schwarz, Hermann Amandus (1843-1921), a utilisé cet algorithme comme outil théorique pour prouver l'existence de solutions à l'équation de Laplace dans des domaines généraux (voir plus loin).

7.2.1. Algorithme multiplicatif en 1D

Le problème test usuel (« Laplacien 1D » soumis à des conditions de Dirichlet homogènes aux deux bords),

$$\boxed{\begin{array}{c} -u_{xx} = f \quad (x \in \Omega =]0,1[) \\ u(0) = u(1) = 0 \end{array}} \tag{7.2}$$

est ici considéré en continu afin d'isoler le problème de la coordination de sous-domaines de celui de la discrétisation.

Soit δ un réel strictement positif suffisamment petit. Le domaine $\Omega =]0,1[$ est partagé en deux sous-domaines,

$$\Omega_1 =]0, \tfrac{1}{2} + \delta[\quad \Omega_2 =]\tfrac{1}{2} - \delta, 1[\tag{7.3}$$

($\Omega_1 \cup \Omega_2 = \Omega$) qui se recouvrent sur l'intervalle

$$\Omega_1 \cap \Omega_2 =]\tfrac{1}{2} - \delta, \tfrac{1}{2} + \delta[\tag{7.4}$$

On pose

$$\boxed{\phi(x) = \int_0^x f(t)\,dt\,,\ \psi(x) = \int_0^x \phi(t)\,dt} \tag{7.5}$$

de sorte que la solution du problème est donnée par :

$$u(x) = -\psi(x) + Ax + B \tag{7.6}$$

où les constantes A et B sont réglées de manière que la fonction $u(x)$ satisfasse les conditions de Dirichlet homogènes aux deux bords :

$$\left\{ \begin{array}{l} B = 0 \\ A = \psi(1) \end{array} \right. \tag{7.7}$$

ce qui donne :

$$\boxed{u(x) = -\psi(x) + x\,\psi(1)} \tag{7.8}$$

et en particulier :

$$\boxed{u(\tfrac{1}{2}+\delta) = -\psi(\tfrac{1}{2}+\delta) + \left(\tfrac{1}{2}+\delta\right)\ \psi(1)} \tag{7.9}$$

On se propose de construire une suite $\{u^n\}$ d'approximations de $u(\frac{1}{2}+\delta)$,

$$\boxed{u^n = u(\tfrac{1}{2}+\delta) + e^n} \tag{7.10}$$

où e^n est l'erreur à l'itération n. L'itération $n+1$ consiste à résoudre successivement [2] deux sous-problèmes respectivement sur Ω_1 et Ω_2.

Plus précisément, on résout d'abord sur Ω_1 le problème suivant :

$$\boxed{\begin{array}{c} -v_{xx} = f \quad \left(0 < x < \frac{1}{2}+\delta\right) \\ v(0) = 0\,,\ v(\frac{1}{2}+\delta) = u^n \end{array}} \tag{7.11}$$

dont on tire l'approximation suivante de $u(\frac{1}{2}-\delta)$:

$$\boxed{v^{n+1} = v(\tfrac{1}{2}-\delta)} \tag{7.12}$$

On résout ensuite le problème suivant sur Ω_2 :

$$\boxed{\begin{array}{c} -w_{xx} = f \quad \left(\frac{1}{2}-\delta < x < 1\right) \\ w(\frac{1}{2}-\delta) = v^{n+1}\,,\ w(1) = 0 \end{array}} \tag{7.13}$$

dont on tire la nouvelle approximation de $u(\frac{1}{2}+\delta)$ suivante :

$$\boxed{u^{n+1} = w(\tfrac{1}{2}+\delta) = u(\tfrac{1}{2}+\delta) + e^{n+1}} \tag{7.14}$$

(voir figure 7.3).

2. On examinera plus tard des variantes parallélisables de cet algorithme.

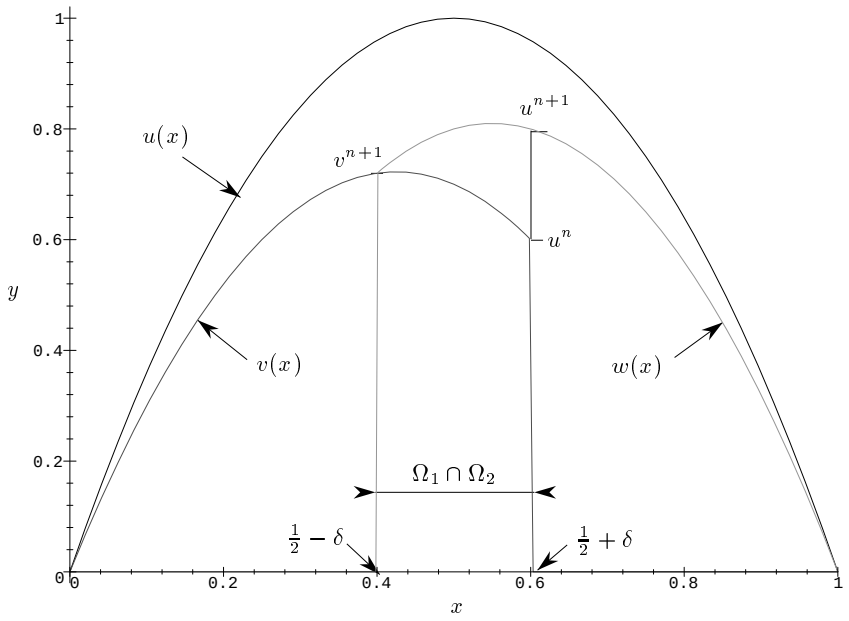


Figure 7.3. Algorithme de Schwarz multiplicatif pour le problème modèle unidimensionnel

Les calculs qui suivent ont pour objectif d'établir la relation entre les deux valeurs successives de l'erreur e^n et e^{n+1}.

On a :

$$v(x) = -\psi(x) + A_1 x + B_1 \tag{7.15}$$

où les constantes A_1 et B_1 sont telles que l'expression ci-dessus satisfait bien les conditions de Dirichlet imposées à la fonction $v(x)$:

$$\begin{cases} 0 = B_1 \\ u^n = -\psi\left(\frac{1}{2} + \delta\right) + A_1\left(\frac{1}{2} + \delta\right) + B_1 \end{cases} \tag{7.16}$$

D'où :

$$v(x) = -\psi(x) + \frac{x}{\frac{1}{2} + \delta}\left[u^n + \psi\left(\frac{1}{2} + \delta\right)\right] \tag{7.17}$$

et en particulier :

$$\boxed{v^{n+1} = -\psi\left(\tfrac{1}{2} - \delta\right) + \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left[u^n + \psi\left(\tfrac{1}{2} + \delta\right)\right]} \tag{7.18}$$

De même :

$$w(x) = -\psi(x) + A_2(1 - x) + B_2 \tag{7.19}$$

où les constantes A_2 et B_2 sont telles que l'expression ci-dessus satisfait bien les conditions de Dirichlet imposées à la fonction $w(x)$:

$$\left\{ \begin{array}{l} 0 = -\psi(1) + B_2 \\ v^{n+1} = -\psi\left(\tfrac{1}{2} - \delta\right) + A_2\left(\tfrac{1}{2} + \delta\right) + B_2 \end{array} \right. \tag{7.20}$$

D'où :

$$w(x) = -\psi(x) + \psi(1) + \frac{1 - x}{\frac{1}{2} + \delta}\left[v^{n+1} + \psi\left(\tfrac{1}{2} - \delta\right) - \psi(1)\right] \tag{7.21}$$

ce qui donne :

$$\begin{aligned} u^{n+1} &= -\psi\left(\tfrac{1}{2} + \delta\right) + \psi(1) + \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left[v^{n+1} + \psi\left(\tfrac{1}{2} - \delta\right) - \psi(1)\right] \\ &= -\psi\left(\tfrac{1}{2} + \delta\right) + \psi(1) + \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left\{\frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left[u^n + \psi\left(\tfrac{1}{2} + \delta\right)\right] - \psi(1)\right\} \end{aligned} \tag{7.22}$$

soit encore :

$$e^{n+1} + \left(\tfrac{1}{2} + \delta\right)\psi(1) = \psi(1) + \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left\{\frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left[e^n + \left(\tfrac{1}{2} + \delta\right)\psi(1)\right] - \psi(1)\right\} \tag{7.23}$$

et enfin :

$$\boxed{e^{n+1} = \left(\frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\right)^2 e^n} \tag{7.24}$$

La suite :

$$\boxed{e^n = \left(\frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\right)^{2n} e^0} \tag{7.25}$$

est donc contractante (et convergente vers 0) si et seulement si :

$$-\left(\tfrac{1}{2}+\delta\right) < \tfrac{1}{2}-\delta < \left(\tfrac{1}{2}+\delta\right) \tag{7.26}$$

ce qui équivaut à la condition naturelle :

$$\boxed{0<\delta<\tfrac{1}{2}} \tag{7.27}$$

En conclusion, on constate que la convergence (de la coordination seule) est d'autant meilleure que le recouvrement est large.

Exercice 7.1 (Application de l'algorithme de Schwarz)

On souhaite résoudre le problème modèle discret unidimensionnel (1.10) dans le cas particulier où

$$f(x) = e^{-2(x-\frac{1}{2})^2} \tag{7.28}$$

par la discrétisation centrée habituelle sur un maillage uniforme $\mathcal{M}_h$ pour lequel

$$h = \frac{1}{10} \tag{7.29}$$

(1) Estimer l'ordre de grandeur ε de l'erreur d'approximation

$$\| u - u_h \|_\infty \tag{7.30}$$

(norme infinie).

(2) On résout les équations discrètes par l'algorithme de Schwarz (multiplicatif) après avoir décomposé le domaine $\Omega = [0,1]$ en 2 sous-domaines se recouvrant partiellement :

$$\Omega_1 = [0, \tfrac{6}{10}] \quad \Omega_2 = [\tfrac{4}{10}, 1]\,. \tag{7.31}$$

Estimer le nombre de cycles complets (1 cycle = 1 résolution sur Ω_1 + 1 résolution sur Ω_2) nécessaires à ce que les solutions partielles

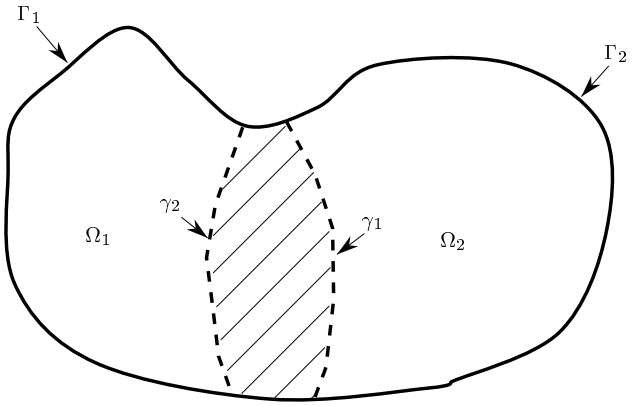
$$u_1 = u_{/\Omega_1} \quad u_1 = u_{/\Omega_2} \tag{7.32}$$

se raccordent à ε près.

7.2.2. Généralisation au cas multidimensionnel

Nous venons de constater explicitement la convergence du processus de coordination dans le cas particulier d'un problème mono-dimensionnel. On généralise ici ce résultat au cas multidimensionnel dans lequel Ω est un ouvert connexe de $\mathbb{R}^d$, Ω_1, Ω_2 deux sous-domaines ouverts et connexes se recouvrant sur une « bande » de taille finie. On introduit les notations suivantes :

$$\left\{\begin{array}{l} \Omega = \Omega_1 \cup \Omega_2 \\ \gamma_1 = \partial\Omega_1 \cap \Omega_2 \\ \Gamma_1 = \partial\Omega_1 - \gamma_1 \\ \gamma_2 = \partial\Omega_2 \cap \Omega_1 \\ \Gamma_2 = \partial\Omega_2 - \gamma_2 \end{array}\right. \tag{7.33}$$



Le problème global à résoudre est ici le suivant :

$$\boxed{\begin{array}{c} -\Delta u = f \text{ sur } \Omega \\ u = g \text{ sur } \partial\Omega \end{array}} \tag{7.34}$$

L'itération $n+1$ de l'algorithme de Schwarz consiste en la résolution des deux sous-problèmes suivants :

$$\left\{\begin{array}{l} -\Delta u_1^{n+1} = f \text{ sur } \Omega_1 \\ u_1^{n+1} = g \text{ sur } \Gamma_1 \\ u_1^{n+1} = u_2^n \text{ sur } \gamma_1 \end{array}\right. \tag{7.35}$$

et

$$\left\{\begin{array}{l} -\Delta u_2^{n+1} = f \text{ sur } \Omega_2 \\ u_2^{n+1} = g \text{ sur } \Gamma_2 \\ u_2^{n+1} = u_1^{n+1} \text{ sur } \gamma_2 \end{array}\right. \tag{7.36}$$

Au préalable, on rappelle la définition des « fonctions harmoniques » et le « principe du maximum ».

Définition 7.1 (fonction harmonique)

Etant donné un ouvert Ω de $\mathbb{R}^d$, on appelle *fonction harmonique* sur Ω toute fonction $\varphi \in C^2(\Omega)$ vérifiant

$$\Delta\varphi = 0 \tag{7.37}$$

en tout point $x \in \Omega$. (Autrement dit, une *fonction harmonique* est une *solution classique* de l'équation de Laplace avec second membre $f = 0$.)

Théorème 7.1 (Principe du maximum)

Etant donné un ouvert borné Ω de $\mathbb{R}^d$, $\partial\Omega$ sa frontière et une fonction φ harmonique sur Ω et continue sur $\overline{\Omega}$, on pose :

$$m = \min_{x \in \partial\Omega} \varphi(x) \tag{7.38}$$

$$M = \max_{x \in \partial\Omega} \varphi(x) \tag{7.39}$$

Alors :

$$\forall\, x \in \Omega\,,\ m \leq \varphi(x) \leq M \tag{7.40}$$

Si de plus, Ω est connexe et φ non constante, les inégalités strictes s'appliquent :

$$\forall\, x \in \Omega\,,\ m < \varphi(x) < M \tag{7.41}$$

En d'autres termes, avec les hypothèses faites, les extrêmas de la fonction φ sont atteints sur le bord et uniquement sur le bord si cette fonction n'est pas constante. Pour une démonstration de cet important théorème, voir par exemple [33].

On considère alors les fonctions donnant l'erreur due à la coordination :

$$e_1^n : \Omega_1 \to \mathbb{R},\ e_2^n : \Omega_2 \to \mathbb{R} \tag{7.42}$$

dont les définitions sont les suivantes :

$$e_1^n = u_1^n - u_{/\Omega_1}\,,\ e_2^n = u_2^n - u_{/\Omega_2} \tag{7.43}$$

où u désigne la solution du problème global. On observe que les fonctions e_1^{n+1} et e_2^{n+1} sont les solutions des sous-problèmes suivants :

$$\left\{ \begin{array}{l} \Delta e_1^{n+1} = 0 \text{ sur } \Omega_1 \\ e_1^{n+1} = 0 \text{ sur } \Gamma_1 \\ e_1^{n+1} = e_2^n \text{ sur } \gamma_1 \end{array} \right. \tag{7.44}$$

et

$$\left\{ \begin{array}{l} \Delta e_2^{n+1} = 0 \text{ sur } \Omega_2 \\ e_2^{n+1} = 0 \text{ sur } \Gamma_2 \\ e_2^{n+1} = e_1^{n+1} \text{ sur } \gamma_2 \end{array} \right. \tag{7.45}$$

On suppose que les ouverts Ω_1 et Ω_2 sont connexes. D'autre part on élimine par hypothèse le cas trivial où $e_{2\,/\gamma_1}^n$ serait identiquement nul. Dans ce cas, la fonction harmonique e_1^{n+1} n'est pas constante sur l'ouvert connexe Ω_1. En conséquence, en vertu du *Principe du Maximum* :

$$\forall\, x \in \Omega_1\ ,\ \min_{\partial\Omega_1}\ e_1^{n+1} < e_1^{n+1}(x) < \max_{\partial\Omega_1}\ e_1^{n+1} \tag{7.46}$$

En particulier, puisque $\gamma_1 \cap \gamma_2 = \emptyset$ (« largeur de bande non nulle »), ce résultat est applicable à γ_2 :

$$\forall\, x \in \gamma_2\ ,\ \min_{\partial\Omega_1}\ e_1^{n+1} < e_1^{n+1}(x) < \max_{\partial\Omega_1}\ e_1^{n+1} \tag{7.47}$$

En conséquence :

$$\|e_{1\,/\gamma_2}^{n+1}\|_\infty < \|e_{1\,/\partial\Omega_1}^{n+1}\|_\infty = \|e_{2\,/\gamma_1}^{n}\|_\infty \tag{7.48}$$

En appliquant le même raisonnement au domaine Ω_2, on aboutit à la conclusion analogue suivante :

$$\|e_{2\,/\gamma_1}^{n+1}\|_\infty < \|e_{1\,/\gamma_2}^{n+1}\|_\infty \tag{7.49}$$

Finalement, on a :

$$\|e_{2\,/\gamma_1}^{n+1}\|_\infty < \|e_{2\,/\gamma_1}^{n}\|_\infty \tag{7.50}$$

Considérons maintenant l'opérateur S suivant :

$$S\ :\ e_{2\,/\gamma_1}^{n} \rightarrow e_{2\,/\gamma_1}^{n+1} \tag{7.51}$$

Cet opérateur est linéaire. Pour simplifier bornons-nous au cas d'espaces fonctionnels de dimension finie. Dans ce cas, l'opérateur S est représenté par une matrice et l'inégalité précédente qui est vraie quelles que soient les conditions initiales implique que le rayon spectral de S est strictement inférieur à 1, ce qui établit la convergence de l'itération.

La preuve s'étend au cas d'espaces fonctionnels de dimension infinie [64].

7.2.3. Aperçu de la preuve de Schwarz

A l'époque de Schwarz, on savait prouver l'existence d'une solution à l'équation de Laplace dans un rectangle en la calculant formellement par la technique de séparation des variables et l'utilisation de séries de Fourier. En utilisant une version particularisée à cette solution du *principe du maximum*, Schwarz a pu démontré la convergence de l'algorithme de coordination lorsque les deux sous-domaines sont des rectangles se recouvrant sur un carré. La fonction obtenue à la limite est une solution de l'équation de Laplace sur le domaine fait de la réunion de ces sous-domaines. Ensuite par généralisations successives, il aboutit à l'existence de la solution dans un domaine connexe quelconque pourvu que la frontière soit polygonale, puis dans un domaine connexe quelconque par passage à la limite.

7.2.4. Algorithme additif

On considère ici une variante « naturellement parallélisable » de l'algorithme précédent. Dans ce nouvel algorithme, on résout simultanément les problèmes de gauche et de droite. Autrement dit, à partir d'une solution à l'itération n définie par les fonctions suivantes (définies sur des domaines qui se recouvrent partiellement) :

$$\left\{ \begin{array}{l} \mathbf{v}^n \ : \ \bar{\Omega}_1 = [0, \frac{1}{2} + \delta] \to \mathbb{R} \\ \mathbf{w}^n \ : \ \bar{\Omega}_2 = [\frac{1}{2} - \delta, 1] \to \mathbb{R} \end{array} \right. \tag{7.52}$$

qui fournissent en particulier les valeurs :

$$\left\{ \begin{array}{l} v^n = \mathbf{v}^n \left(\frac{1}{2} - \delta\right) \\ w^n = \mathbf{w}^n \left(\frac{1}{2} + \delta\right) \end{array} \right. \tag{7.53}$$

on calcule une nouvelle approximation $\left(\mathbf{v}^{n+1} = v,\ \mathbf{w}^{n+1} = w\right)$ en résolvant les deux systèmes suivants :

$$\boxed{\begin{array}{c} -v_{xx} = f \quad \left(0 < x < \frac{1}{2} + \delta\right) \\ v(0) = 0\,,\ v(\frac{1}{2} + \delta) = w^n \end{array}} \tag{7.54}$$

et

$$\boxed{\begin{array}{c} -w_{xx} = f \quad \left(\frac{1}{2} - \delta < x < 1\right) \\ w(\frac{1}{2} - \delta) = v^n\,,\ w(1) = 0 \end{array}} \tag{7.55}$$

En changeant certains symboles dans les résultats précédents, on obtient facilement :

$$v(x) = -\psi(x) + \frac{x}{\frac{1}{2} + \delta} \left[w^n + \psi\left(\tfrac{1}{2} + \delta\right)\right] \tag{7.56}$$

et en particulier

$$\boxed{v^{n+1} = -\psi\left(\tfrac{1}{2} - \delta\right) + \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left[w^n + \psi\left(\tfrac{1}{2} + \delta\right)\right]} \tag{7.57}$$

et symétriquement

$$w(x) = -\psi(x) + \psi(1) + \frac{1-x}{\frac{1}{2} + \delta}\left[v^n + \psi\left(\tfrac{1}{2} - \delta\right) - \psi(1)\right] \tag{7.58}$$

ce qui donne

$$\boxed{w^{n+1} = -\psi\left(\tfrac{1}{2} + \delta\right) + \psi(1) + \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta}\left[v^n + \psi\left(\tfrac{1}{2} - \delta\right) - \psi(1)\right]} \tag{7.59}$$

On introduit ensuite les définitions suivantes :

$$\left\{ \begin{array}{l} v^n = u\left(\frac{1}{2} - \delta\right) + \varepsilon_1^n = -\psi\left(\frac{1}{2} - \delta\right) + \left(\frac{1}{2} - \delta\right)\psi(1) + \varepsilon_1^n \\ w^n = u\left(\frac{1}{2} + \delta\right) + \varepsilon_2^n = -\psi\left(\frac{1}{2} + \delta\right) + \left(\frac{1}{2} + \delta\right)\psi(1) + \varepsilon_2^n \end{array} \right. \tag{7.60}$$

que l'on substitue dans les deux équations encadrées précédentes afin d'aboutir aux relations suivantes :

$$\left\{ \begin{array}{l} \varepsilon_1^{n+1} = \rho'\,\varepsilon_2^n \\ \varepsilon_2^{n+1} = \rho'\,\varepsilon_1^n \end{array} \right. \tag{7.61}$$

où l'on a posé

$$\rho' = \frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta} = \sqrt{\rho} \tag{7.62}$$

Ce résultat peut s'écrire sous forme vectorielle de la manière suivante :

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}^{n+1} = \rho' \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}^n \tag{7.63}$$

ce qui fait apparaître les valeurs propres $\pm\rho'$ de la matrice d'amplification.

En conclusion, cet algorithme de coordination dont le rayon spectral est égal à ρ' est deux fois plus lent que le précédent mais permet la résolution simultanée des deux sous-systèmes sur une architecture parallèle.

La relation simple que nous venons d'obtenir entre les rayons spectraux des algorithmes de Schwarz additif (identifié par $\nu = 0$ dans ce qui suit) et multiplicatif ($\nu = 1$) est générale dans le cas de 2 sous-domaines. En effet, la seule hypothèse de linéarité de la discrétisation intérieure et des informations transférées d'un sous-domaine à l'autre permet de donner à ces algorithmes la forme générale suivante :

$$A_{11}\, u_I^{n+1} = f_I - A_{12}\, u_{II}^{n} \tag{7.64}$$

$$A_{22}\, u_{II}^{n+1} = f_{II} - A_{12}\, u_{II}^{n+\nu} \tag{7.65}$$

pour certaines définitions appropriées des blocs A_{ij} $(i, j = 1, 2)$ et des seconds membres f_I et f_{II}. On voit donc que ces algorithmes équivalent respectivement aux itérations de Jacobi et Gauss-Seidel par blocs (étudiées au chapitre 3) appliquées au système linéaire suivant :

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_I \\ u_{II} \end{pmatrix} = \begin{pmatrix} f_I \\ f_{II} \end{pmatrix} \tag{7.66}$$

pour lequel nous avions déjà démontré que $\rho_1 = \rho_0^2$. Bien évidemment, dans le cas d'un partionnement des inconnues correspondant à plus de 2 sous-domaines, la relation entre les vitesses de convergence des algorithmes additif et multiplicatif n'est pas aussi simple.

7.2.5. Gain en efficacité

Dans le cas du laplacien, on constate en pratique que le gain en efficacité est maximal lorsque le recouvrement (c'est-à-dire, en 1D, le paramètre δ), bien qu'essentiel à la convergence de l'algorithme de coordination, est aussi petit que possible. De plus, lorsqu'on résout par relaxation les sous-systèmes discrets associés aux différents sous-domaines (itération de Jacobi), on constate que le temps de calcul de la résolution globale est moindre lorsqu'on effectue une seule itération de Jacobi par itération de coordination. Dans ce cas l'algorithme de Schwarz équivaut à la méthode de relaxation initiale *répartie*. Autrement dit, une programmation parallèle de l'algorithme additif permet alors de réaliser un gain en efficacité à peu près égal au nombre de sous-domaines, c'est-à-dire d'unités de calcul. Ce résultat ne s'étend pas nécessairement aux cas d'autres équations ou d'autres choix de la méthode de résolution.

7.3. Méthodes pour l'advection-diffusion

De très nombreux phénomènes physiques sont modélisés par des EDP combinant des termes de convection, dits d'advection en linéaire, aux termes de diffusion. Il est par conséquent intéressant d'examiner de plus près le cas de l'équation parabolique d'advection-diffusion. Par exemple, en 2D, on considère l'équation suivante :

$$u_t + \vec{V}(x, y)\,.\,\nabla u = \varepsilon\, \Delta u \quad (\text{sur } \Omega) \tag{7.67}$$

dans laquelle le champ de vitesses $\vec{V}(x,y)$ est donné. Cette équation est soumise à un jeu de conditions aux limites.

Exercice 7.2 (Advection pure et conditions aux limites)

On considère le problème d'advection pure ($\varepsilon = 0$) dans le cas où l'espace est de dimension 2 ($\Omega \subseteq \mathbb{R}^2$), et le vecteur $\vec{V}(x,y)$ est constant :

$$\vec{V}(x,y) = \begin{pmatrix} a \\ b \end{pmatrix} \tag{7.68}$$

(1) Etablir (ou rappeler) que la solution du problème associé au domaine non borné ($\Omega = \mathbb{R}^2$) correspondant aux valeurs initiales

$$u(x,y,0) = u_0(x,y) \tag{7.69}$$

est la suivante :

$$\begin{aligned} u(x,y,t) &= u(M,t) \\ &= u_0(M - \vec{V}t) \\ &= u_0(x - at, y - bt) \end{aligned} \tag{7.70}$$

(2) Pour le problème associé à un domaine borné convexe Ω, on partitionne la frontière $\Gamma = \partial\Omega$ en deux arcs :

$$\Gamma^- = \left\{ (x,y) \in \Gamma \text{ tq } \vec{V} \,.\, \vec{n} \leq 0 \right\} \tag{7.71}$$

$$\Gamma^+ = \left\{ (x,y) \in \Gamma \text{ tq } \vec{V} \,.\, \vec{n} > 0 \right\} \tag{7.72}$$

Montrer que le problème suivant :

$$u_t + \vec{V} \,.\, \nabla u = 0 \quad (\text{sur } \Omega) \tag{7.73}$$

$$u(x,y,t) = g(x,y) \quad (\forall\, (x,y) \in \Gamma^*,\ \forall\, t > 0) \tag{7.74}$$

$$u(x,y,0) = u_0(x,y) \quad (\forall\, (x,y) \in \Omega) \tag{7.75}$$

est bien ou mal posé suivant que $\Gamma^* = \Gamma^-$ ou Γ^+.

Revenant au cas général de l'advection-diffusion ($\varepsilon > 0$), on souhaite fixer des conditions aux limites pour lesquelles le problème reste bien posé lorsque la diffusion est évanescente ($\varepsilon \ll 1$). Pour cela, on s'inspire du problème hyperbolique de l'advection pure que l'on obtient pour $\varepsilon = 0$ (cf. exercice 7.2). On voit qu'il est naturel de fixer u (condition de Dirichlet) sur la partie Γ^- du bord $\Gamma = \partial\Omega$ où l'advection est « entrante », par exemple :

$$u = g \quad (\text{sur } \Gamma^-) \tag{7.76}$$

où :

$$\Gamma^- = \left\{ (x,y) \in \Gamma \text{ tq } \vec{V}(x,y) \,.\, \vec{n} \leq 0 \right\} \tag{7.77}$$

A l'inverse sur la partie Γ^+ du bord où l'advection est « sortante », on impose une condition de Neumann :

$$\frac{\partial u}{\partial n} = \vec{n} \,.\, \nabla u = \phi \quad (\text{sur } \Gamma^+) \tag{7.78}$$

où

$$\Gamma^+ = \left\{ (x,y) \in \Gamma \text{ tq } \vec{V}(x,y) \,.\, \vec{n} > 0 \right\} \tag{7.79}$$

On considère ici un problème d'évolution où apparaît la dérivée partielle par rapport au temps u_t car il est souvent commode, même dans les situations où l'on ne s'intéresse qu'à la solution stationnaire, de construire une méthode de résolution itérative où les itérations sont des intégrations en temps. Par conséquent, on sous-entend qu'on a également spécifié une condition initiale.

Supposons pour fixer les idées que l'intégration en temps est effectuée par la méthode d'Euler implicite. A chaque pas de temps, on doit alors résoudre le système suivant :

$$\boxed{\begin{array}{c} \alpha\, u + \vec{V}(x,y) \,.\, \nabla u - \varepsilon\, \Delta u = f \quad (\text{sur } \Omega) \\ u = g \quad (\text{sur } \Gamma^-) \\ \dfrac{\partial u}{\partial n} = \phi \quad (\text{sur } \Gamma^+) \end{array}} \tag{7.80}$$

où l'on a posé :

$$\alpha \overset{\text{déf}}{=} \frac{1}{\Delta t}\,, \; f \overset{\text{déf}}{=} \frac{u^n(x,y)}{\Delta t} \tag{7.81}$$

et

$$u^{n+1}(x,y) = u \tag{7.82}$$

Lorsque les termes de diffusion dominent (ε grand), l'algorithme de Schwarz converge pour ce problème, par continuité avec celui de l'équation de Laplace.

Revenons maintenant au cas inverse représentatif de nombreux problèmes d'écoulement où la convection, ici l'advection, est dominante (ε petit). Imaginons qu'on applique alors l'algorithme de Schwarz en utilisant un étroit recouvrement, de sorte

que les interfaces γ_1 et γ_2 sont proches (voir figure 7.4). Dans ce cas, on voit que par continuité il n'est pas possible que les conditions

$$\vec{V}(x,y)\,.\,\vec{n}_1 < 0 \quad (\text{sur } \gamma_1)\,,\ \vec{V}(x,y)\,.\,\vec{n}_2 < 0 \quad (\text{sur } \gamma_2) \tag{7.83}$$

dans lesquelles $\vec{n}_1$ et $\vec{n}_2$ sont les normales extérieures aux sous-domaines Ω_1 et Ω_2 respectivement prises le long des interfaces γ_1 et γ_2 ($\vec{n}_1 + \vec{n}_2 \approx 0$), soient satisfaites simultanément. Par conséquent au moins l'un des sous-problèmes de Dirichlet, et peut-être les deux, est mal posé. On est donc amené à modifier la méthode de coordination pour en faire un « algorithme de Dirichlet-Neumann » défini ci-après.

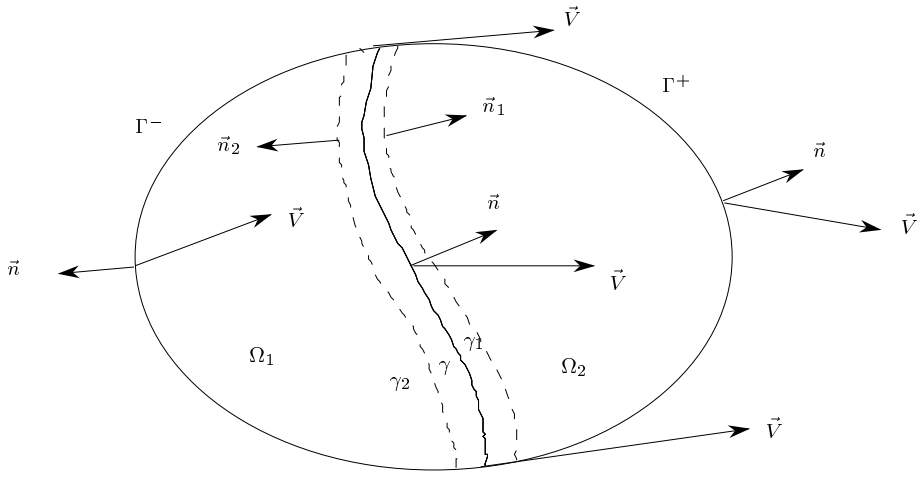


Figure 7.4. Décomposition de domaine pour un problème d'advection

On suppose désormais une décomposition du domaine Ω en partition

$$\Omega = \Omega_1 \cup \Omega_2\,,\ \Omega_1 \cap \Omega_2 = \emptyset \tag{7.84}$$

on note $\gamma = \gamma_{12}$ l'interface et $\vec{n} = \vec{n}_{12}$ la normale extérieure dans le sens de Ω_1 vers Ω_2. Considérons d'abord le cas où :

$$\forall\,(x,y)\ \in\ \gamma\,,\ \vec{V}(x,y)\,.\,\vec{n} > 0 \tag{7.85}$$

Dans ce cas, à l'interface γ, il est naturel d'imposer une condition de Neumann à l'inconnue u_1 du sous-domaine Ω_1, et à l'inverse, une condition de Dirichlet à l'inconnue u_2 du sous-domaine Ω_2. Par exemple, dans l'algorithme additif, les sous-systèmes sont indépendants et les conditions à l'interface de l'un sont calculées à partir de l'itéré

précédent de la solution de l'autre. On résout donc sur Ω_1 :

$$\alpha\, u_1 + \vec{V}(x,y)\,.\,\nabla u_1 - \varepsilon\,\Delta u_1 = f \quad (\text{sur } \Omega_1) \tag{7.86}$$

$$u_1 = g \quad (\text{sur } \Gamma^- \cap \partial\Omega_1) \tag{7.87}$$

$$\frac{\partial u_1}{\partial n} = \phi \quad (\text{sur } \Gamma^+ \cap \partial\Omega_1) \tag{7.88}$$

$$\frac{\partial u_1}{\partial n} = \frac{\partial u_2^n}{\partial n} \quad (\text{sur } \gamma) \tag{7.89}$$

et sur Ω_2 :

$$\alpha\, u_2 + \vec{V}(x,y)\,.\,\nabla u_2 - \varepsilon\,\Delta u_2 = f \quad (\text{sur } \Omega_2) \tag{7.90}$$

$$u_2 = g \quad (\text{sur } \Gamma^- \cap \partial\Omega_2) \tag{7.91}$$

$$\frac{\partial u_2}{\partial n} = \phi \quad (\text{sur } \Gamma^+ \cap \partial\Omega_2) \tag{7.92}$$

$$u_2 = u_1^n \quad (\text{sur } \gamma) \tag{7.93}$$

On peut également définir une variante multiplicative, en utilisant sur Ω_2 les valeurs fraîchement calculées sur Ω_1 :

$$u_2 = u_1^{n+1} \quad (\text{sur } \gamma) \tag{7.94}$$

Lorsque la condition jusqu'ici supposée vraie à l'interface, (7.85), n'est pas satisfaite uniformément sur γ, on peut transformer légérement la méthode ci-dessus en un « algorithme de Dirichlet-Neumann adaptatif ». Pour cela, dans la résolution sur le sous-domaine Ω_1, on applique la condition de Neumann seulement sur la partie γ^+ de l'interface γ où la condition (7.85) est satisfaite et celle de Dirichlet sur la partie γ^- complémentaire. On procède à l'inverse dans la résolution sur le sous-domaine Ω_2. Dans l'algorithme additif, ceci donne pour u_1 :

$$\frac{\partial u_1}{\partial n} = \frac{\partial u_2^n}{\partial n} \quad (\text{sur } \gamma^+) \tag{7.95}$$

$$u_1 = u_2^n \quad (\text{sur } \gamma^-) \tag{7.96}$$

et pour u_2 :

$$u_2 = u_1^n \quad (\text{sur } \gamma^+) \tag{7.97}$$

$$\frac{\partial u_2}{\partial n} = \frac{\partial u_1^n}{\partial n} \quad (\text{sur } \gamma^-) \tag{7.98}$$

On réfère à [79] pour une discussion approfondie de ces diverses variantes, que nous allons maintenant étudier dans le cas d'une seule dimension d'espace par le biais de l'exercice suivant. (Voir aussi [42].)

Exercice 7.3 (Algorithme de Dirichlet-Neumann)

On souhaite étudier la méthode de coordination dite de « Dirichlet-Neumann » appliquée au problème d'advection-diffusion suivant :

$$\left\{ \begin{array}{l} c\,U_X = \varepsilon\,U_{XX} \quad (0 < X < a) \quad (c > 0\,,\ \varepsilon > 0) \\ U(0) = U_0\,,\ U(a) = 0 \end{array} \right. \tag{7.99}$$

(1) On pose :

$$x = \frac{X}{a}\,,\ u(x) = \frac{U(X)}{U_0} \tag{7.100}$$

et on introduit le nombre de Reynolds (sans dimension)

$$r = \frac{c\,a}{\varepsilon} \tag{7.101}$$

que l'on suppose grand :

$$r \gg 1 \tag{7.102}$$

Montrer que le problème initial équivaut au suivant

$$\left\{ \begin{array}{l} u_x = \frac{1}{r}\,u_{xx} \quad (0 < x < 1) \\ u(0) = 1\,,\ u(1) = 0 \end{array} \right. \tag{7.103}$$

et établir que la « solution générale » est de la forme :

$$u(x) = A\,e^{rx} + B \tag{7.104}$$

Montrer que pour r grand,

$$u'(1) \approx -r \tag{7.105}$$

Tracer schématiquement $u(x)$. En pratique, si on résout par un schéma discret sur un maillage non uniforme, où convient-il d'utiliser un maillage fin ?

(2) Algorithme additif :

On partitionne l'intervalle

$$\Omega =]0, 1[\tag{7.106}$$

en deux sous-intervalles disjoints

$$\Omega_1 =]0, \delta[\,,\ \Omega_2 =]\delta, 1[\quad (0 < \delta < 1) \tag{7.107}$$

On construit une suite de solutions approchées au problème global, $\{u^n(x)\}$, dont le $(n+1)$-ème itéré $u(x) = u^{n+1}(x)$ se calcule à partir du précédent en résolvant (de manière exacte) les sous-problèmes suivants :

– sur $\Omega_1 =]0, \delta[$, le problème de Neumann suivant :

$$\left\{ \begin{array}{l} u_x = \frac{1}{r}\, u_{xx} \quad (0 < x < \delta) \\ u(0) = 1\ ,\ u_x(\delta) = p \end{array} \right. \tag{7.108}$$

dans lequel la valeur de p provient de la restriction à Ω_2 de l'itéré précédent $\{u^n(x)\}$:

$$p = u_x^n(\delta^+) = \frac{d}{dx} u^n(\delta^+) \tag{7.109}$$

(l'indice supérieur $^+$ sur δ correspond à une évaluation du côté droit)

– sur $\Omega_2 =]\delta, 1[$, le problème de Dirichlet suivant :

$$\left\{ \begin{array}{l} u_x = \frac{1}{r}\, u_{xx} \quad (\delta < x < 1) \\ u(\delta) = v\ ,\ u(1) = 0 \end{array} \right. \tag{7.110}$$

dans lequel la valeur de v provient de la restriction à Ω_1 de l'itéré précédent $\{u^n(x)\}$:

$$v = u^n(\delta^-) \tag{7.111}$$

(l'indice supérieur $^-$ sur δ correspond à une évaluation du côté gauche).

Utiliser la forme de la solution générale, (7.104), pour résoudre explicitement ces deux sous-problèmes en fonction de (v, p). En déduire les quantités :

$$v' = u^{n+1}(\delta^-)\ ,\ p' = u_x^{n+1}(\delta^+) \tag{7.112}$$

et montrer que :

$$\left(\begin{array}{c} v' \\ p' \end{array} \right) = G \left(\begin{array}{c} v \\ p \end{array} \right) + \vec{b} \tag{7.113}$$

où G est la matrice 2×2 suivante :

$$G = \left(\begin{array}{cc} 0 & g_{12} \\ g_{21} & 0 \end{array} \right) \tag{7.114}$$

où

$$g_{12} = \frac{1 - e^{-r\delta}}{r}\ ,\ g_{21} = \frac{-re^{r\delta}}{e^r - e^{r\delta}} \tag{7.115}$$

et $\vec{b}$ un vecteur constant. Calculer les valeurs propres de G et montrer que pour r grand son rayon spectral satisfait la condition :

$$\rho(G) < 1 \tag{7.116}$$

Qu'en concluez-vous ?

Expliquer comment cet algorithme de coordination permet d'utiliser un maillage adapté présentant une forte variation de densité de points.

(3) Algorithme multiplicatif :

On modifie l'algorithme précédent en résolvant séquentiellement sur Ω_1 puis sur Ω_2 et en utilisant lors de la résolution sur Ω_2 la valeur actualisée

$$v = u^{n+1}(\delta^-) \tag{7.117}$$

que l'on vient juste de calculer sur Ω_1.

Recalculer le rayon spectral. Qu'a-t-on gagné, qu'a-t-on perdu ?

(4) On reprend l'algorithme de la question (2) que l'on modifie en inversant le type de condition à l'interface. Plus précisément, dans le calcul de $u(x) = u^{n+1}(x)$, le sous-problème associé à Ω_1 est désormais de type "Dirichlet", la condition à l'interface étant la suivante :

$$u(\delta) = v \tag{7.118}$$

où la valeur v provient de la restriction au sous-domaine Ω_2 de la solution précédente :

$$v = u^n(\delta^+) \tag{7.119}$$

A l'inverse, le sous-problème associé à Ω_2 est désormais de type "Neumann", la condition à l'interface étant la suivante :

$$u_x(\delta) = p \tag{7.120}$$

où la valeur p provient de la restriction au sous-domaine Ω_1 de la solution précédente :

$$p = u_x^n(\delta^-) \tag{7.121}$$

Calculer le rayon spectral de cet algorithme et commenter le résultat.

7.4. Techniques issues du contrôle – Equation adjointe

Une fois n'étant pas coutume, on considère ici une équation modèle de convection-diffusion *non linéaire* :

$$u_t + auu_x + buu_y - \varepsilon(u_{xx} + u_{yy}) = 0\,,\ (x,y) \in]-1,1[^2 \tag{7.122}$$

L'intégration en temps s'effectue simplement par la méthode d'Euler. Par exemple, en implicite, le $n+1$-ème itéré $u = u^{n+1}$ est solution du système suivant :

$$\begin{array}{l} \alpha u + auu_x + buu_y - \varepsilon(u_{xx} + u_{yy}) = f\,,\ (x,y) \in]-1,1[^2 \\ \\ \qquad u(-1,y) = g(y)\,,\ \forall y \in]-1,1[\\ \qquad u_y(x,-1) = u_y(x,1) = 0\,,\ \forall x \in]-1,1[\\ \qquad u_x(1,y) = 0\,,\ \forall y \in]-1,1[\end{array} \tag{7.123}$$

où l'on a posé

$$\alpha = 1/\Delta t\,,\ f = u^n(x,y)/\Delta t\,. \tag{7.124}$$

On se place dans le cas où ce système est bien posé (e.g. $a > 0 : g > 0 : \varepsilon$ suffisamment grand).

Une « méthode de gradient » en partition pour le problème de base, (7.123), consiste à adopter les options suivantes :

- Partition du domaine

$$\Omega =]-1,1[\times]-1,1[\tag{7.125}$$

  en deux sous-domaines, par exemple

$$\Omega_1 =]-1,0[\times]-1,1[\,,\ \Omega_2 =]0,1[\times]-1,1[\tag{7.126}$$

  et introduction des valeurs de u à l'interface γ (=]-1,1[de l'axe des y) comme fonction de contrôle $v(y)$

- Définition d'une fonctionnelle de coût

$$J(v) = \frac{1}{2}\int_{-1}^{1} \left(u_x^{(1)} - u_x^{(2)}\right)^2 (0,y)\,\omega(y)\,dy \tag{7.127}$$

  mesurant la violation de régularité à l'interface (voir figure 7.5) et identification de son gradient [3]

3. L'utilisation d'une fonction de pondération $\omega(y) > 0$ n'est pas essentielle.

- Construction d'une méthode d'itération sur $v(y)$ ayant pour critère d'arrêt

$$J(v) < \epsilon \tag{7.128}$$

où ϵ mesure la tolérance en précision.

$$v \longrightarrow \begin{cases} u^{(1)} \text{ sur } \Omega_1 \\ u^{(2)} \text{ sur } \Omega_2 \end{cases} \longrightarrow J(v)$$

fonctionnelle J

Figure 7.5. Schéma fonctionnel du critère J

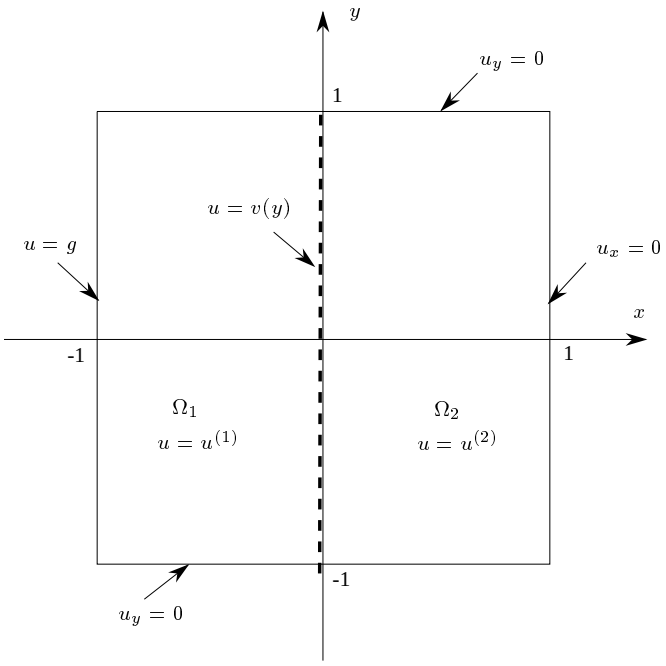


Figure 7.6. Partition de domaine et fonction de contrôle à l'interface

Par conséquent, les sous-problèmes se formulent respectivement comme suit,

sur Ω_1 :

$$\begin{array}{l} \alpha u^{(1)} + a u^{(1)} u_x^{(1)} + b u^{(1)} u_y^{(1)} - \varepsilon (u_{xx}^{(1)} + u_{yy}^{(1)}) = f \,, \\ \qquad (x,y) \in]-1,0[\times]-1,1[\\ \\ \quad u^{(1)}(-1,y) = g(y) \,,\ \forall y \in]-1,1[\\ \quad u_y^{(1)}(x,-1) = u_y^{(1)}(x,1) = 0 \,,\ \forall x \in]-1,0[\\ \quad u^{(1)}(0,y) = v(y) \,,\ \forall y \in]-1,1[\end{array} \tag{7.129}$$

et sur Ω_2 :

$$\begin{array}{l} \alpha u^{(2)} + a u^{(2)} u_x^{(2)} + b u^{(2)} u_y^{(2)} - \varepsilon (u_{xx}^{(2)} + u_{yy}^{(2)}) = f \,, \\ \qquad (x,y) \in]0,1[\times]-1,1[\\ \\ \quad u^{(2)}(0,y) = v(y) \,,\ \forall y \in]-1,1[\\ \quad u_y^{(2)}(x,-1) = u_y^{(2)}(x,1) = 0 \,,\ \forall x \in]0,1[\\ \quad u_x^{(2)}(1,y) = 0 \,,\ \forall y \in]-1,1[\end{array} \tag{7.130}$$

Dans cet exemple, le sous-problème sur Ω_1 est de type Dirichlet-Dirichlet en x et de type $Neumann - Neumann$ en y, alors que le sous-problème sur Ω_2 est de type Dirichlet-Neumann en x.

A l'issue de la résolution (séparée) des sous-problèmes posés sur Ω_1 et Ω_2, on évalue la fonctionnelle de coût $J(v)$ définie en (7.127). On souhaite maintenant identifier le gradient de cette fonctionnelle par rapport à $v(y)$ de manière à pouvoir modifier $v(y)$ dans un sens qui réduise cette fonctionnelle. Pour cette identification, on note $\delta u^{(1)}$ et $\delta u^{(2)}$ les variations que subissent les solutions partielles $u^{(1)}$ et $u^{(2)}$ lorsqu'on perturbe le contrôle $v(y)$ de $\delta v(y)$ (voir figure 7.7).

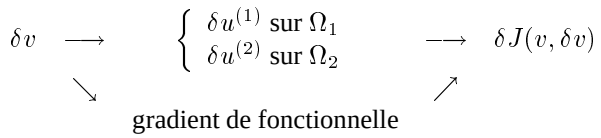


Figure 7.7. Schéma fonctionnel du gradient

Les variations $\delta u^{(1)}$ et $\delta u^{(2)}$ sont solutions des sous-problèmes linéarisés, qui sont les suivants :

sur Ω_1 :

$$\left(\alpha + a u_x^{(1)} + b u_y^{(1)}\right) \delta u^{(1)} + a u^{(1)} \delta u_x^{(1)} + b u^{(1)} \delta u_y^{(1)} = \varepsilon(\delta u_{xx}^{(1)} + \delta u_{yy}^{(1)}) \,, \quad (x,y) \in]-1,0[\times]-1,1[$$

$$\begin{aligned} &\delta u^{(1)}(-1,y) = 0 \,, \ \forall y \in]-1,1[\\ &\delta u_y^{(1)}(x,-1) = \delta u_y^{(1)}(x,1) = 0 \,, \ \forall x \in]-1,0[\\ &\delta u^{(1)}(0,y) = \delta v(y) \,, \ \forall y \in]-1,1[\end{aligned}$$

(7.131)

et sur Ω_2 :

$$\left(\alpha + a u_x^{(2)} + b u_y^{(2)}\right) \delta u^{(2)} + a u^{(2)} \delta u_x^{(2)} + b u^{(2)} \delta u_y^{(2)} = \varepsilon(\delta u_{xx}^{(2)} + \delta u_{yy}^{(2)}) \,, \quad (x,y) \in]0,1[\times]-1,1[$$

$$\begin{aligned} &\delta u^{(2)}(0,y) = \delta v(y) \,, \ \forall y \in]-1,1[\\ &\delta u_y^{(2)}(x,-1) = \delta u_y^{(2)}(x,1) = 0 \,, \ \forall x \in]0,1[\\ &\delta u_x^{(2)}(1,y) = 0 \,, \ \forall y \in]-1,1[\end{aligned}$$

(7.132)

Les deux systèmes précédents constituent une représentation implicite des fonctions $\delta u^{(1)}(x,y)$ et $\delta u^{(2)}(x,y)$ en tant que fonctionnelles de $\delta v(y)$. On en déduit l'expression implicite suivante de la variation de fonctionnelle coût :

$$\delta J = \int_{-1}^{1} \left(u_x^{(1)} - u_x^{(2)}\right) \left(\delta u_x^{(1)} - \delta u_x^{(2)}\right)(0,y)\, \omega(y)\, dy \tag{7.133}$$

Il s'agit maintenant de transformer cette relation en une expression explicite de $\delta v(y)$. Pour cela, on utilise la technique classique de l'« équation adjointe » constituée des étapes suivantes :

1. Définir une variable adjointe $\lambda(x,y)$ dont les restrictions à Ω_1 et Ω_2 sont respectivement notées $\lambda^{(1)}$ et $\lambda^{(2)}$.

2. Faire le produit scalaire de l'EDP satisfaite par $\delta u^{(1)}$ (resp. $\delta u^{(2)}$) par $\lambda^{(1)}$ (resp. $\lambda^{(2)}$), c'est-à-dire multiplier et intégrer sur le sous-domaine Ω_1 (resp. Ω_2).

3. Transformer l'identité obtenue par des intégrations par parties et des simplifications liées aux conditions aux limites homogènes satisfaites par $\delta u^{(1)}$ (resp. $\delta u^{(2)}$).

4. Poser que $\lambda^{(1)}$ (resp. $\lambda^{(2)}$) est solution d'un certain système adjoint de telle sorte que l'identité se simplifie encore et devienne informative d'une contribution jusqu'alors implicite dans l'expression de δJ.

Nous allons d'abord réaliser les étapes 2, 3 et 4 ci-dessus pour le sous-domaine Ω_1 en omettant momentanément l'indice supérieur $^{(1)}$ pour alléger l'écriture et en adoptant la notation

$$\iint_{\Omega_1} (.) \stackrel{\text{déf}}{=} \int_{-1}^{0}\int_{-1}^{1} (.)\, dx\, dy \tag{7.134}$$

Sur Ω_1 on part donc de l'équation suivante :

$$\boxed{\iint_{\Omega_1} \lambda(x,y)\left[(\alpha + au_x + bu_y)\,\delta u + au\delta u_x + bu\delta u_y - \varepsilon\,(\delta u_{xx} + \delta u_{yy})\right] dx\, dy = 0} \tag{7.135}$$

qui est une identité car elle est satisfaite quel que soit le choix que l'on fera ultérieurement (étape 4) de la fonction $\lambda(x,y)$. Avant cela, simplifions par des intégrations par parties pour éliminer les dérivées partielles par rapport à x ou y portant sur la variable d'état u. On a :

$$\begin{aligned}
\iint_{\Omega_1} \lambda a u \delta u_x &= \int_{-1}^{1}\left(\int_{-1}^{0} a u \lambda \delta u_x\, dx\right) dy \\
&= \int_{-1}^{1} \left[a u \lambda \delta u\right]_{x=-1}^{x=0} dy - \iint_{\Omega_1} (a u \lambda)_x\, \delta u \\
&= \int_{-1}^{1} a u \lambda(0,y) \delta v(y)\, dy - \iint_{\Omega_1} (a u \lambda)_x\, \delta u
\end{aligned} \tag{7.136}$$

où l'on a utilisé le fait que δu est nul sur le bord où u est imposé ($x = -1$) et égal à δv sur celui où u est « contrôlé » ($x = 0$). Symétriquement,

$$\begin{aligned}
\iint_{\Omega_1} \lambda b u \delta u_y &= \int_{-1}^{0}\left(\int_{-1}^{1} b u \lambda \delta u_y\, dy\right) dx \\
&= \int_{-1}^{0} \left[b u \lambda \delta u\right]_{y=-1}^{y=1} dx - \iint_{\Omega_1} (b u \lambda)_y\, \delta u \\
&= \int_{-1}^{0} \left[b u \lambda \delta u(x,1) - b u \lambda \delta u(x,-1)\right] dx - \iint_{\Omega_1} (b u \lambda)_y\, \delta u
\end{aligned} \tag{7.137}$$

Pour les termes de diffusion, il faut faire deux intégrations par parties. En particulier :

$$\begin{aligned}
\iint_{\Omega_1} \lambda\delta u_{xx} &= \int_{-1}^{1} \left(\int_{-1}^{0} \lambda\delta u_{xx}\, dx \right) dy \\
&= \int_{-1}^{1} [\lambda\delta u_x]_{x=-1}^{x=0}\, dy - \iint_{\Omega_1} \lambda_x \delta u_x \\
&= \int_{-1}^{1} [\lambda\delta u_x(0,y) - \lambda\delta u_x(-1,y)]\, dy - \int_{-1}^{1} \left(\int_{-1}^{0} \lambda_x \delta u_x\, dx \right) dy \\
&= \int_{-1}^{1} [\lambda\delta u_x(0,y) - \lambda\delta u_x(-1,y)]\, dy - \int_{-1}^{1} [\lambda_x \delta u]_{x=-1}^{x=0}\, dy \\
&+ \iint_{\Omega_1} \lambda_{xx}\delta u \\
&= \int_{-1}^{1} [\lambda\delta u_x(0,y) - \lambda\delta u_x(-1,y) - \lambda_x(0,y)\delta v(y)]\, dy \\
&+ \iint_{\Omega_1} \lambda_{xx}\delta u
\end{aligned} \tag{7.138}$$

Symétriquement,

$$\begin{aligned}
\iint_{\Omega_1} \lambda\delta u_{yy} &= \int_{-1}^{0} \left(\int_{-1}^{1} \lambda\delta u_{yy}\, dy \right) dx \\
&= \int_{-1}^{0} [\lambda\delta u_y]_{y=-1}^{y=1}\, dx - \iint_{\Omega_1} \lambda_y \delta u_y \\
&= -\int_{-1}^{0} \left(\int_{-1}^{1} \lambda_y \delta u_y\, dy \right) dx \\
&= -\int_{-1}^{0} [\lambda_y \delta u]_{y=-1}^{y=1}\, dx + \iint_{\Omega_1} \lambda_{yy}\delta u \\
&= \int_{-1}^{0} [-\lambda_y \delta u(x,1) + \lambda_y \delta u(x,-1)]\, dx + \iint_{\Omega_1} \lambda_{yy}\delta u
\end{aligned} \tag{7.139}$$

Les équations (7.136)-(7.137)-(7.138)-(7.139) permettent de transformer (7.135) en

l'expression suivante :

$$\begin{aligned}\iint_{\Omega_1} \left[(\alpha + au_x + bu_y)\,\lambda - (au\lambda)_x - (bu\lambda)_y - \varepsilon\,(\lambda_{xx} + \lambda_{yy})\right]\,\delta u \\ + \int_{-1}^{1} (au\lambda + \varepsilon\lambda_x)\,(0,y)\,\delta v(y)\,dy \\ + \int_{-1}^{0} \left[(bu\lambda + \varepsilon\lambda_y)\,\delta u(x,1) - (bu\lambda + \varepsilon\lambda_y)\,\delta u(x,-1)\right]\,dx \\ + \int_{-1}^{1} \left[-\varepsilon\lambda\delta u_x(0,y) + \varepsilon\lambda\delta u_x(-1,y)\right]\,dy \\ = 0\,,\ \forall\,\lambda\end{aligned} \tag{7.140}$$

Cette équation fait intervenir trois types de termes : une intégrale (la première) étendue au domaine dont l'intégrande fait apparaître le facteur δu et aucune de ses dérivées partielles, une intégrale de bord (la seconde) faisant intervenir explicitement la perturbation de contrôle, et deux intégrales de bord faisant intervenir des perturbations *a priori* inconnues de l'état. Puisqu'il s'agit d'une identité, on choisit $\lambda = \lambda^{(1)}$ (étape 4) de manière à éliminer la première et la troisième intégrales, et à donner à la quatrième la forme de la partie de δJ qui est proportionnelle à $\delta u_x(0,y) = \delta u_x^{(1)}(0,y)$. Plus précisément, on pose l'« équation adjointe » suivante :

$$\begin{aligned}&(\alpha + au_x + bu_y)\,\lambda^{(1)} - \left(au\lambda^{(1)}\right)_x - \left(bu\lambda^{(1)}\right)_y = \varepsilon(\lambda_{xx}^{(1)} + \lambda_{yy}^{(1)})\,, \\ &\qquad (x,y) \in]-1,0[\times]-1,1[\\ &\lambda^{(1)}(-1,y) = 0\,,\ \forall y \in]-1,1[\\ &\left[bu^{(1)}\lambda^{(1)} + \varepsilon\lambda_y^{(1)}\right](x,-1) = 0\,,\ \forall x \in]-1,0[\\ &\left[bu^{(1)}\lambda^{(1)} + \varepsilon\lambda_y^{(1)}\right](x,1) = 0\,,\ \forall x \in]-1,0[\\ &\lambda^{(1)}(0,y) = \left(u_x^{(1)} - u_x^{(2)}\right)(0,y)\,\omega(y)\,,\ \forall y \in]-1,1[\end{aligned} \tag{7.141}$$

de sorte que (7.140) fournit le résultat suivant :

$$\begin{aligned}&\varepsilon \int_{-1}^{1} \left(u_x^{(1)} - u_x^{(2)}\right)\,\delta u_x^{(1)}(0,y)\,\omega(y)\,dy \\ &= \int_{-1}^{1} \left(au\lambda^{(1)} + \varepsilon\lambda_x^{(1)}\right)(0,y)\,\delta v(y)\,dy\end{aligned} \tag{7.142}$$

On procède ensuite de manière analogue pour le sous-domaine Ω_2 pour lequel les conditions aux limites ne sont pas exactement les mêmes. Après quelques calculs de

même nature, on obtient les résultats suivants (en omettant, pour alléger l'écriture, l'indice supérieur $^{(2)}$ portant sur λ dans les 5 équations suivantes) :

$$\iint_{\Omega_2} \lambda a u \delta u_x = \int_{-1}^{1} \left[(a u \lambda \delta u(1,y)) - a u \lambda(0,y)\, \delta v(y) \right] dy - \iint_{\Omega_2} (a u \lambda)_x \, \delta u \tag{7.143}$$

Symétriquement,

$$\iint_{\Omega_2} \lambda b u \delta u_y = \int_{0}^{1} \left[b u \lambda \delta u(x,1) - b u \lambda \delta u(x,-1) \right] dx - \iint_{\Omega_2} (b u \lambda)_y \, \delta u \tag{7.144}$$

En outre,

$$\iint_{\Omega_2} \lambda \delta u_{xx} = \int_{-1}^{1} \left[-\lambda \delta u_x(0,y) - \lambda_x \delta u(1,y) + \lambda_x(0,y) \delta v(y) \right] dy + \iint_{\Omega_2} \lambda_{xx} \delta u \tag{7.145}$$

Symétriquement,

$$\iint_{\Omega_2} \lambda \delta u_{yy} = \int_{0}^{1} \left[-\lambda_y \delta u(x,1) + \lambda_y \delta u(x,-1) \right] dx + \iint_{\Omega_1} \lambda_{yy} \delta u \tag{7.146}$$

L'identité satisfaite par $\lambda = \lambda^{(2)}$ est donc la suivante :

$$\begin{aligned} \iint_{\Omega_2} \left[(\alpha + a u_x + b u_y)\, \lambda - (a u \lambda)_x - (b u \lambda)_y - \varepsilon \left(\lambda_{xx} + \lambda_{yy} \right) \right] \delta u \\ - \int_{-1}^{1} \left(a u \lambda + \varepsilon \lambda_x \right)(0,y)\, \delta v(y)\, dy \\ + \int_{0}^{1} \left[\left(b u \lambda + \varepsilon \lambda_y \right) \delta u(x,1) - \left(b u \lambda + \varepsilon \lambda_y \right) \delta u(x,-1) \right] dx \\ + \int_{-1}^{1} \left[\left(a u \lambda + \varepsilon \lambda_x \right) \delta u(1,y) + \varepsilon \lambda \delta u_x(0,y) \right] dy \\ = 0\,, \ \forall\, \lambda \end{aligned} \tag{7.147}$$

ce qui conduit à poser l'« équation adjointe » suivante pour $\lambda = \lambda^{(2)}$:

$$\boxed{\begin{array}{c}
\left(\alpha + a u_x + b u_y\right) \lambda^{(2)} - \left(a u \lambda^{(2)}\right)_x - \left(b u \lambda^{(2)}\right)_y = \varepsilon (\lambda^{(2)}_{xx} + \lambda^{(2)}_{yy}) \,, \\
(x,y) \in]0,1[\times]-1,1[\\
\\
\left[a u^{(2)} \lambda^{(2)} + \varepsilon \lambda^{(2)}_x\right](1,y) = 0 \,, \ \forall y \in]-1,1[\\
\left[b u^{(2)} \lambda^{(2)} + \varepsilon \lambda^{(2)}_y\right](x,-1) = 0 \,, \ \forall x \in]0,1[\\
\left[b u^{(2)} \lambda^{(2)} + \varepsilon \lambda^{(2)}_y\right](x,1) = 0 \,, \ \forall x \in]0,1[\\
\lambda^{(2)}(0,y) = \left(u^{(1)}_x - u^{(2)}_x\right)(0,y)\, \omega(y) \,, \ \forall y \in]-1,1[
\end{array}} \tag{7.148}$$

de sorte que (7.147) fournit le résultat suivant :

$$\boxed{\begin{array}{l}
\varepsilon \displaystyle\int_{-1}^{1} \left(u^{(1)}_x - u^{(2)}_x\right)(0,y)\, \delta u^{(2)}_x(0,y)\, \omega(y)\, dy \\
\qquad = \displaystyle\int_{-1}^{1} \left(a u \lambda^{(2)} + \varepsilon \lambda^{(2)}_x\right)(0,y)\, \delta v(y)\, dy
\end{array}} \tag{7.149}$$

Finalement, en combinant (7.142) et (7.149), on aboutit au gradient recherché :

$$\boxed{\delta J = \int_{-1}^{1} K(y)\, \delta v(y)\, dy} \tag{7.150}$$

où l'on a posé :

$$\boxed{K(y) = \left(\lambda^{(1)}_x - \lambda^{(2)}_x\right)(0,y)} \tag{7.151}$$

En résumé, les étapes de la méthode sont les suivantes :

1. *Choix initial de la fonction d'interface* $v(y)$
2. *Résolution de l'équation d'état sur les différents sous-domaines, (7.129)-(7.130), et calcul de la fonctionnelle de coût* $J(v)$ *: arrêt si* $J(v) < \epsilon$*, sinon :*
3. *Résolution de l'équation adjointe sur les différents sous-domaines, (7.141)-(7.148), et calcul du gradient de la fonctionnelle de coût* $K(y)$

4. Modification de la fonction d'interface

$$v(y) \longrightarrow v(y) + \delta v(y) \tag{7.152}$$

et retour à l'étape 2.

En ce qui concerne le choix de la perturbation $\delta v(y)$, plusieurs options sont possibles :

4.1 Un pas dans la direction opposée au gradient :

$$\delta v(y) = -\rho \frac{K(y)}{\int_{-1}^{1} K^2(y)\, dy} J \quad (0 < \rho < 1) \tag{7.153}$$

de sorte que si ρ est suffisamment petit : $\delta J \approx -\rho J$.

4.2 Une optimisation unidimensionnelle :
Une fois l'état u, le coût J_0 et le gradient $K(y)$ calculés, on perturbe le contrôle dans la direction opposée au gradient, comme ci-dessus, avec deux valeurs différentes ρ_1 et ρ_2 du pas, ce qui fournit deux valeurs supplémentaires du coût J_1 et J_2 (après deux nouvelles résolutions de l'équation d'état). On choisit ρ comme l'abscisse du minimum de la parabole qui passe par les points $(0, J_0)$, (ρ_1, J_1) et (ρ_2, J_2).

4.3 Gradients conjugués :
Dans la méthode précédente on remplace le gradient par une direction de gradient conjugué, ou toute autre direction de descente produite par un algorithme de minimisation plus savant (e.g. Méthode de Davidon, GMRES [82]).

REMARQUES FINALES :

(1) On a choisi dans l'exemple, de régulariser la dérivée normale à l'interface par le choix de l'état u comme fonction de contrôle. En conséquence, on a obtenu un gradient égal au saut de cette dérivée à travers l'interface. On peut à l'inverse, régulariser la fonction d'état (continuité) par le choix de la dérivée normale comme fonction de contrôle (pour autant que les problèmes correspondants aux sous-domaines restent bien posés, voir exercice 7.4). Dans ce cas, l'expression du gradient contient le saut de u à travers l'interface comme facteur.

Exercice 7.4 (Contrôle de la continuité par la dérivée normale)

On considère le problème suivant qui diffère du précédent par le remplacement de la

condition de Neumann sur le bord $y = -1$ par une condition de type Dirichlet :

$$\begin{array}{l} \alpha u + auu_x + buu_y - \varepsilon(u_{xx} + u_{yy}) = f\,,\ (x,y) \in]-1,1[^2 \\ \\ \quad u(-1,y) = g(y)\,,\ \forall y \in]-1,1[\\ \quad u(x,-1) = h(x)\,,\ \forall x \in]-1,1[\\ \quad u_y(x,1) = 0\,,\ \forall x \in]-1,1[\\ \quad u_x(1,y) = 0\,,\ \forall y \in]-1,1[\end{array} \tag{7.154}$$

On souhaite résoudre ce problème par partition du domaine conformément à la figure 7.8 et contrôle par la dérivée normale à l'interface ($x = 0$) :

$$v(y) = u_x(0,y) \quad (-1 \leq y \leq 1) \tag{7.155}$$

Reformuler les sous-problèmes non linéaires, la fonctionnelle de coût, sa première variation, les sous-problèmes linéarisés et les équations adjointes. En déduire l'expression du gradient.

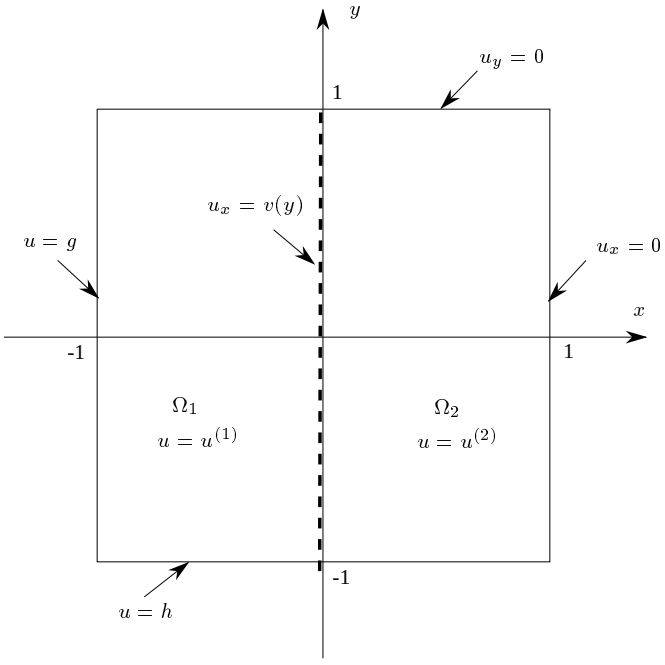


Figure 7.8. Partition et contrôle par la dérivée normale (exercice 7.4)

On peut imaginer des « méthodes adaptatives » dans lesquelles on choisirait localement le long de l'interface la « bonne variable » (ou son gradient) à contrôler.

(2) **Moindres carrés**

Dans le cas d'une décomposition en sous-domaines ayant un recouvrement dont on note γ_1 et γ_2 les bords, une approche alternative consiste à régulariser à convergence en contrôlant u (ou toute autre variable plus pertinente) sur γ_1 et γ_2 et en minimisant la fonctionnelle de coût suivante :

$$J(v_1, v_2) = \frac{1}{2} \iint_{\Omega_1 \cap \Omega_2} \left(u^{(1)} - u^{(2)} \right)^2 \omega(x, y)\, dx\, dy \quad (\omega(x, y) \geq 0) \qquad (7.156)$$

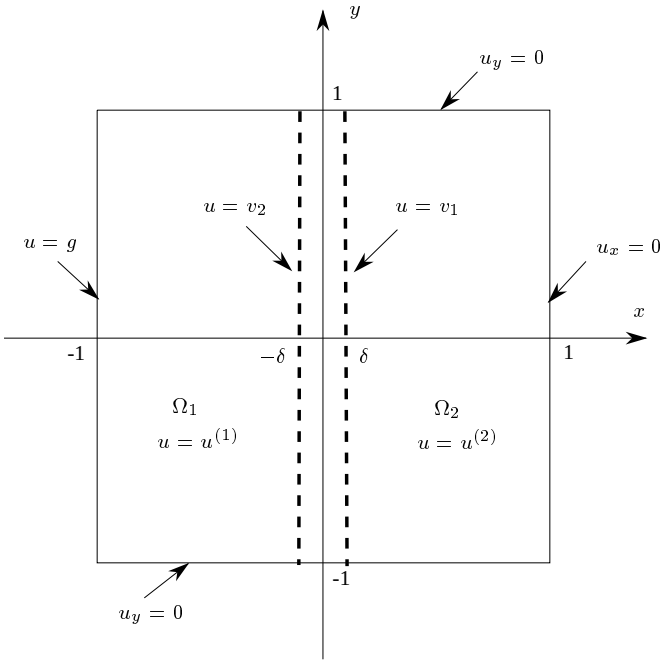


Figure 7.9. Décomposition de domaine avec recouvrement et fonction de contrôle aux interfaces dans la méthode des moindres carrés

Exercice 7.5 (Moindres carrés)

Identifier le gradient de la fonctionnelle de coût de la méthode des moindres carrés.

(3) Le mérite principal de ces méthodes de contrôle semble résider davantage dans leur approche rationnelle pour identifier des conditions d'interface viables et dans le fait qu'elles ont permis de résoudre effectivement certains problèmes de grande taille inaccessibles globalement, que dans leur potentialité à réduire le coût global de la résolution par le biais notamment du parallélisme.

Annexe A

Normes et équivalences

Dans cette annexe, on rappelle les principales propriétés des normes usuelles définies sur l'espace vectoriel de dimension finie K^M où K est le corps des nombres réels ($K = \mathbb{R}$) ou complexes ($K = \mathbb{C}$), ainsi que sur l'espace noté $\mathcal{M}_{M\times M}$ des matrices de dimension $M \times M$ dont les éléments sont dans K. On précise en particulier les constantes d'équivalence entre les « normes-p », p étant un réel supérieur ou égal à 1, souvent mais pas nécessairement entier, ou le symbole ∞.

A.1. Normes vectorielles

A.1.1. Définitions générales

Définition A.1

« Norme-p » ($p \geq 1$) sur K^M ($K = \mathbb{R}$ ou $\mathbb{C}$) :

$$\boxed{\|u\|_p \stackrel{\text{déf}}{=} \left(\sum_{j=1}^{M} |u_j|^p \right)^{\frac{1}{p}} \qquad (u \in K^M)} \tag{A.1}$$

En particulier, on a :

$$\begin{array}{c} \|u\|_1 = \sum_j |u_j| \\ \|u\|_2 = \sqrt{\sum_j |u_j|^2} = \sqrt{u^* u} \\ \|u\|_\infty = \lim_{p\to\infty} \|u\|_p = \max_j |u_j| \end{array} \tag{A.2}$$

où l'on a simplifié l'écriture des $\sum$ et du $\max$ sachant que tous les indices prennent les valeurs 1, 2, ..., M. De plus l'indice supérieur * correspondant à la transposition-conjugaison :

$$u^* = \bar{u}^T \tag{A.3}$$

La norme-2 est également appelée « norme euclidienne ».

A.1.2. Relations d'équivalence avec la norme infinie

D'après les définitions, il est évident que :

$$\forall\, u \in K^M\,,\ \|u\|_p \leq \left(M \max_j |u_j|^p\right)^{\frac{1}{p}} = M^{\frac{1}{p}} \|u\|_\infty \tag{A.4}$$

et ce résultat donne la majoration la plus stricte possible qui soit générale.

Exercice A.1 (Majoration de la norme-p par la norme infinie)

Montrer qu'il existe des vecteurs u pour lesquels il y a égalité.

Inversement,

$$\|u\|_p \geq \left(\max_j |u_j|^p\right)^{\frac{1}{p}} = \|u\|_\infty\,. \tag{A.5}$$

Exercice A.2 (Majoration de la norme infinie par la norme-p)

Montrer qu'il existe des vecteurs u pour lesquels il y a égalité.

Finalement,

$$\boxed{\forall\, u \in K^M\,,\ \frac{1}{M^{\frac{1}{p}}} \|u\|_p \leq \|u\|_\infty \leq \|u\|_p \leq M^{\frac{1}{p}} \|u\|_\infty\,.} \tag{A.6}$$

A.1.3. Relations d'équivalence entre la norme-p et la norme-q

Lemme A.1

Soit $u \in K^M$ fixé. L'application $p \in [1, +\infty] \to \|u\|_p$ est monotone-décroissante.

DÉMONSTRATION : soient deux réels positifs p et q tels que

$$p \geq q \geq 1\,. \tag{A.7}$$

Il s'agit de prouver que l'on a :

$$\|u\|_p \leq \|u\|_q \,. \tag{A.8}$$

Si $u = 0$, c'est évident. Sinon ($u \neq 0$), posons :

$$\xi = \frac{u}{\|u\|_p} \tag{A.9}$$

de sorte que

$$\|\xi\|_p = \left(\sum_j |\xi_j|^p \right)^{\frac{1}{p}} = 1 \,. \tag{A.10}$$

En conséquence,

$$\forall\, j \,,\ |\xi_j| \leq 1 \text{ et } |\xi_j|^q \geq |\xi_j|^p \ , \tag{A.11}$$

de sorte que

$$\|\xi\|_q = \left(\sum_j |\xi_j|^q \right)^{\frac{1}{q}} \geq \left(\sum_j |\xi_j|^p \right)^{\frac{1}{q}} = 1 \,, \tag{A.12}$$

ce qui établit le résultat cherché. □

Théorème A.1

Soient p et q deux réels positifs tels que $p \geq q \geq 1$. On a :

$$\boxed{\forall\, u \in K^M \,,\ \frac{1}{M^{\frac{1}{q}-\frac{1}{p}}} \|u\|_q \leq \|u\|_p \leq \|u\|_q \leq M^{\frac{1}{q}-\frac{1}{p}} \|u\|_p \,.} \tag{A.13}$$

DÉMONSTRATION : introduisons d'abord les définitions suivantes :

Définition A.2

« Sphère de K^M de rayon unité au sens de la norme-p » :

$$\mathcal{S}_p^M \overset{\text{déf}}{=} \left\{ \xi \in K^M \text{ tq } \|\xi\|_p = 1 \right\} \tag{A.14}$$

Définition A.3

« Boule fermée de K^M de rayon unité au sens de la norme-p » :

$$\mathcal{B}_p^M \stackrel{\text{déf}}{=} \left\{ \eta \in K^M \text{ tq } \|\eta\|_p \leq 1 \right\} \tag{A.15}$$

Exercice A.3 (Identification de sphères et de boules)

Identifier ces ensembles dans les cas correspondants à $M = 2$ et $p = 1$, 2 et ∞.

Posons

$$\mu = \mu(M) \stackrel{\text{déf}}{=} \max_{u \in K^M,\ u \neq 0} \frac{\|u\|_q}{\|u\|_p} . \tag{A.16}$$

En vertu de l'homogénéité de la norme, la constante μ est également donnée par :

$$\mu = \max_{\xi \in \mathcal{S}_p^M} \|\xi\|_q \tag{A.17}$$

La détermination de la constante μ est un problème d'optimisation sous contrainte qui se résout en formant un hamiltonien par adjonction de la contrainte à la fonction à maximiser :

$$\mathcal{H} = \mathcal{H}(\xi, \lambda) = \|\xi\|_q + \lambda \left(\|\xi\|_p - 1 \right) \tag{A.18}$$

On peut sans perte de généralité se restreindre au cas où $\xi_j \geq 0\,,\ \forall\, j$, ce qui permet de supprimer certaines valeurs absolues :

$$\mathcal{H} = \left(\sum_j \xi_j^q \right)^{\frac{1}{q}} + \lambda \left[\left(\sum_j \xi_j^p \right)^{\frac{1}{p}} - 1 \right] \tag{A.19}$$

La localisation du maximum s'obtient alors en écrivant que $\mathcal{H}$ est stationnaire par rapport à toutes ses variables $\{\xi_j\}$ ($j = 1,\ 2, ...,\ M$) et λ :

$$\forall\, j\ :\ \frac{\partial \mathcal{H}}{\partial \xi_j} = \frac{\partial \mathcal{H}}{\partial \lambda} = 0\,. \tag{A.20}$$

Or l'expression de $\mathcal{H}$ est une fonction symétrique des composantes du vecteur ξ. Par conséquent les M premières des équations ci-dessus permettent de donner à ξ_j la même expression en fonction de λ ; en reportant dans la dernière on pourrait en tirer la valeur de λ. Il est inutile de faire ce calcul : on peut dès à présent conclure que le maximum est atteint lorsque toutes les composantes de ξ sont égales, ce qui donne :

$$\mu = M^{\frac{1}{q} - \frac{1}{p}}\,. \tag{A.21}$$

Ce résultat établit la première et la troisième inégalités du théorème. Quant à la seconde, c'est l'expression directe du Lemme A.1. □

REMARQUE : par passage à la limite, le théorème redonne le résultat de la sous-section précédente.

A.1.4. Inégalité de Hölder

Théorème A.2

« *Inégalité de Hölder* » :
Soient deux réels positifs p *et* q *tels que*

$$\frac{1}{p} + \frac{1}{q} = 1 \,; \tag{A.22}$$

on a :

$$\boxed{\forall\, u \in K^M \,,\ (\|u\|_2)^2 \leq \|u\|_p \, \|u\|_q \,.} \tag{A.23}$$

DÉMONSTRATION : si $p \geq q$, on a $p \geq 2 \geq q \geq 1$ et sinon on inverse les notations de p et q. Traçons alors dans un plan (x, y) la courbe $\mathcal{C}$ d'équation :

$$y = x^{p-1} \tag{A.24}$$

ce qui équivaut à :

$$x = y^{q-1} \,. \tag{A.25}$$

Soient a et b deux réels strictement positifs, A_1 l'aire entre la courbe $\mathcal{C}$ et l'axe des x de l'abscisse 0 à a, et A_2 l'aire entre la courbe $\mathcal{C}$ et l'axe des y de l'ordonnée 0 à b. On a :

$$A_1 = \int_0^a x^{p-1}\, dx = \frac{a^p}{p} \,, \quad A_2 = \int_0^b y^{q-1}\, dy = \frac{b^q}{q} \tag{A.26}$$

Il est évident d'après la figure que la somme de ces aires dépasse celle d'un rectangle d'arêtes a et b. Par conséquent :

$$a\, b \leq \frac{a^p}{p} + \frac{b^q}{q} \,. \tag{A.27}$$

On fait ensuite les substitutions suivantes :

$$a = \frac{|u_j|}{\|u\|_p} \,, \quad b = \frac{|u_j|}{\|u\|_q} \tag{A.28}$$

ce qui donne :

$$\frac{|u_j|^2}{\|u\|_p \, \|u\|_q} \leq \frac{|u_j|^p}{p\,(\|u\|_p)^p} + \frac{|u_j|^q}{q\,(\|u\|_q)^q}\,. \tag{A.29}$$

Enfin, on somme sur j :

$$\frac{(\|u\|_2)^2}{\|u\|_p \, \|u\|_q} \leq \frac{1}{p} + \frac{1}{q} = 1\,, \tag{A.30}$$

ce qui fournit le résultat. □

A.2. Normes matricielles

A.2.1. Définitions et propriétés générales

On note $\mathcal{M}_{M\times M}$ l'espace vectoriel des matrices de dimension $M \times M$ dont les éléments sont dans K ($K = \mathbb{R}$ ou $\mathbb{C}$). On rappelle que cet espace est isomorphe à celui des endomorphismes de K^M.

Les définitions faites précédemment des normes-p sur l'espace K^M *induisent* naturellement la définition suivante de la « norme-p induite » d'une matrice carrée A :

Définition A.4

« Norme-p induite » ($p \geq 1$) d'une matrice $A \in \mathcal{M}_{M\times M}$:

$$\boxed{\|A\|_p \stackrel{\text{déf}}{=} \max_{u\,\in\,K^M,\ u\neq 0} \frac{\|A\,u\|_p}{\|u\|_p} \quad (A \in \mathcal{M}_{M\times M})} \tag{A.31}$$

En vertu de l'homogénéité de la norme, lorsque $u \neq 0$ on a

$$\frac{\|A\,u\|_p}{\|u\|_p} = \left\| A\,\frac{u}{\|u\|_p} \right\|_p \tag{A.32}$$

de sorte que le résultat suivant pourrait servir de définition alternative à la norme

induite :

Théorème A.3

$$\forall\, A \in \mathcal{M}_{M\times M}\ ,\ \|A\|_p \overset{\text{déf}}{=} \max_{\xi\,\in\,\mathcal{S}_p^M} \|A\,\xi\|_p = \max_{\eta\,\in\,\mathcal{B}_p^M} \|A\,\eta\|_p \tag{A.33}$$

où $\mathcal{S}_p^M$ et $\mathcal{B}_p^M$ sont à nouveau la sphère et la boule fermée de K^M de rayon unité au sens de la norme-p.

On peut alternativement définir la norme d'une matrice de dimension $M \times M$ comme celle d'un vecteur de K^{M^2}, sans que cette définition soit *induite* par la définition faite de la norme vectorielle. En particulier, on considérera la « norme infinie standard » ou « norme du max » et la « norme euclidienne » dont les définitions sont les suivantes :

Définition A.5

« Norme du max d'une matrice $A = \{a_{j,k}\}$ de $\mathcal{M}_{M\times M}$ »

$$\|A\|_{\max} \overset{\text{déf}}{=} \max_{j,k} |a_{j,k}| \quad (A \in \mathcal{M}_{M\times M}) \tag{A.34}$$

Définition A.6

« Norme euclidienne d'une matrice $A = \{a_{j,k}\}$ de $\mathcal{M}_{M\times M}$ »

$$\|A\|_E \overset{\text{déf}}{=} \sqrt{\sum_j \sum_k |a_{j,k}|^2} \quad (A \in \mathcal{M}_{M\times M}) \tag{A.35}$$

Exercice A.4 (Norme euclidienne d'une matrice)

Montrer que :

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \|A\|_E = \sqrt{\mathrm{Trace}\ (\mathrm{A^*\,A})}} \tag{A.36}$$

Les relations d'équivalence entre ces deux normes sont évidentes :

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \frac{1}{M}\,\|A\|_E \leq \|A\|_{\max} \leq \|A\|_E \leq M\,\|A\|_{\max}\,.} \tag{A.37}$$

L'avantage des normes induites est qu'elles permettent d'écrire des majorations du type

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \forall\, u \in K^M\,,\ \|A\,u\| \leq \|A\|\,\|u\|} \tag{A.38}$$

alors que dans le cas inverse un coefficient multiplicatif apparaîtrait à droite ; par exemple :

$$\|A\,u\|_\infty \leq M\,\|A\|_{\max}\,\|u\|_\infty \tag{A.39}$$

En outre, quelle que soit la norme induite $\|.\|$:

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \forall\, B \in \mathcal{M}_{M\times M}\,,\ \|A\,B\| \leq \|A\|\,\|B\|} \tag{A.40}$$

Exercice A.5 (Norme induite d'un produit de matrices)

Justifier cette inégalité.

D'autre part, on introduit la définition suivante :

Définition A.7 (Rayon Spectral d'une matrice)

Le rayon spectral $\rho(A)$ d'une matrice carrée $A \in \mathcal{M}_{M\times M}$ dont les valeurs propres sont notées $\{\lambda_m\}$ $(m = 1, 2, ..., M)$, est défini comme suit :

$$\boxed{\rho(A) \stackrel{\text{déf}}{=} \max_{m=1,2,...,M} |\lambda_m| \quad (A \in \mathcal{M}_{M\times M})} \tag{A.41}$$

Théorème A.4

Toute norme induite de toute matrice carrée $A \in \mathcal{M}_{M\times M}$ est au moins égale à son rayon spectral

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \|A\| \geq \rho\,(A)} \tag{A.42}$$

Exercice A.6 (Norme induite et rayon spectral d'une matrice)

Justifier cette inégalité et identifier des cas où il y a égalité.

La dimension M étant fixée, les inclusions suivantes résultent du Lemme A.1 :

$$\mathcal{B}_1^M \subseteq \mathcal{B}_2^M \subseteq ... \subseteq \mathcal{B}_\infty^M \tag{A.43}$$

Par conséquent, quelle que soit la matrice carrée A, on a en vertu de (A.33) :

$$\|A\|_p = \max_{\eta\,\in\,\mathcal{B}_p^M} \|A\,\eta\|_p \leq \max_{\eta\,\in\,\mathcal{B}_\infty^M} \|A\,\eta\|_p \tag{A.44}$$

car $\mathcal{B}_\infty^M$ contient $\mathcal{B}_p^M$. De plus, en vertu de (A.6) on a aussi :

$$\|A\,\eta\|_p \leq M^{\frac{1}{p}}\,\|A\,\eta\|_\infty \tag{A.45}$$

Il en résulte la conclusion suivante :

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \|A\|_p \leq M^{\frac{1}{p}}\,\|A\|_\infty} \tag{A.46}$$

Il est plus difficile d'être général pour établir les relations en sens inverse. A cette fin, on établit d'abord comment calculer les principales normes lorsqu'on connaît les éléments de la matrice.

A.2.2. Nouvelle définition de la norme-infinie induite

On pose à nouveau $A = \{a_{j,k}\}$ $(j, k = 1, 2, ..., M)$. On a :

$$\|A\|_\infty = \max_{\xi} \max_{j} \left| \sum_k a_{j,k}\,\xi_k \right| \tag{A.47}$$

où le vecteur ξ satisfait la contrainte $\max_k\,|\xi_k| = 1$. On voit que le maximum par rapport à ξ est atteint lorsque ce vecteur est donné par :

$$\xi_k = \begin{cases} \bar{a}_{j,k}/|a_{j,k}| & \text{si } a_{j,k} \neq 0 \\ 1 & \text{sinon} \end{cases} \tag{A.48}$$

et ce, pour la valeur de l'indice j pour laquelle résultat est maximum. D'où finalement :

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\ ,\ \|A\|_\infty = \max_j \sum_k |a_{j,k}|} \tag{A.49}$$

Autrement dit, pour chaque ligne de la matrice on fait la somme des modules des éléments et on prend le plus grand des nombres ainsi obtenus. Pour cette raison, cette norme est souvent appelée dans la littérature anglophone « Maximum-Row-Sum ».

A.2.3. Nouvelle définition de la norme-1 induite

Il résulte de la définition que :

$$\begin{aligned}
\|A\|_1 &= \max_{\xi\,\in\,K^M\ /\ \sum_k\ |\xi_k|=1} \sum_j \Big|\big(A\,\xi\big)_j\Big| \\
&= \max_{\xi\,\in\,K^M\ /\ \sum_k\ |\xi_k|=1} \sum_j \left|\sum_k a_{j,k}\,\xi_k\right| \\
&\le \max_{\xi\,\in\,K^M\ /\ \sum_k\ |\xi_k|=1} \sum_j \sum_k |a_{j,k}|\ |\xi_k| \\
&\le \max_{\xi\,\in\,K^M\ /\ \sum_k\ |\xi_k|=1} \sum_k \underbrace{\left(\sum_j |a_{j,k}|\right)}_{\le\ \|A^*\|_\infty} |\xi_k| \\
&\le \|A^*\|_\infty \max_{\xi\,\in\,K^M\ /\ \sum_k\ |\xi_k|=1} \sum_k |\xi_k| = \|A^*\|_\infty\,.
\end{aligned} \tag{A.50}$$

Inversement, soit k l'indice pour lequel

$$\sum_j |a_{j,k}| = \|A^*\|_\infty \tag{A.51}$$

et ξ^k le vecteur dont les composantes sont données par

$$\xi^k_j = \delta^k_j \tag{A.52}$$

(symbole de Kronecker), de sorte que

$$\|\xi^k\|_1 = \|\xi^k\|_\infty = 1 \tag{A.53}$$

et

$$\begin{aligned} \left\| A\,\xi^k \right\|_1 &= \sum_j \left| \left(A\,\xi^k \right)_j \right| \\ &= \sum_j \left| \sum_m a_{j,m}\,\xi_m^k \right| \\ &= \sum_j |a_{j,k}| \\ &= \|A^*\|_\infty \end{aligned} \tag{A.54}$$

Par conséquent,

$$\|A\|_1 \geq \left\| A\,\xi^k \right\|_1 = \|A^*\|_\infty \tag{A.55}$$

En combinant les deux résultats principaux précédents et celui de la sous-section précédente, on aboutit à la conclusion suivante :

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \|A\|_1 = \|A^*\|_\infty = \max_k \sum_j |a_{j,k}|} \tag{A.56}$$

Autrement dit, pour chaque colonne de la matrice on fait la somme des modules des éléments et on prend le plus grand des nombres ainsi obtenus. Pour cette raison, cette norme est souvent appelée dans la littérature anglophone « Maximum-Column-Sum ».

A.2.4. Relations d'équivalence entre les normes 1 et infinie induites

En vertu de (A.46), on sait que :

$$\|A\|_1 \leq M\,\|A\|_\infty\,. \tag{A.57}$$

Les nouvelles définitions données aux normes 1 et infinie de la matrice A montrent que :

$$\|A\|_1 \geq \|A\|_{\max}\,,\ \|A\|_\infty \leq M\,\|A\|_{\max} \tag{A.58}$$

de sorte que

$$\|A\|_1 \geq \frac{1}{M}\,\|A\|_\infty\,. \tag{A.59}$$

En regroupant ces résultats, on tire les conclusions suivantes :

$$\boxed{\begin{aligned} &\forall\, A \in \mathcal{M}_{M\times M}\,,\ \tfrac{1}{M}\,\|A\|_\infty \leq \|A\|_1 \leq M\,\|A\|_\infty\,, \\ &\qquad\qquad \tfrac{1}{M}\,\|A\|_1 \leq \|A\|_\infty \leq M\,\|A\|_1\,. \end{aligned}} \tag{A.60}$$

A.2.5. Nouvelle définition de la norme-2 induite

On a :

$$\begin{aligned}
(\|A\|_2)^2 &= \left(\max_{u \in K^M,\ u \neq 0} \frac{\|A\,u\|_2}{\|u\|_2} \right)^2 \\
&= \max_{u \in K^M,\ u \neq 0} \left(\frac{\|A\,u\|_2}{\|u\|_2} \right)^2 \\
&= \max_{u \in K^M,\ u \neq 0} \frac{(A\,u)^*\ A\,u}{u^*\,u} \\
&= \max_{u \in K^M,\ u \neq 0} \frac{u^*\ (A^*\,A)\ u}{u^*\,u}
\end{aligned} \tag{A.61}$$

où $A^* = \bar{A}^T$ désigne la matrice adjointe. La matrice $A^*\,A$ est hermitienne semi-définie positive, et se diagonalise par une transformation unitaire U :

$$\begin{gathered}
A^*\,A = U^*\,\Lambda\,U \\
U^*\,U = I_M \quad \text{(matrice identité)} \\
\Lambda\ :\ \text{matrice diagonale semi-définie positive.}
\end{gathered} \tag{A.62}$$

En posant

$$v = U\,u \tag{A.63}$$

on note que :

$$\|v\|_2^2 = v^*\,v = u^*\,u = \|u\|_2^2 \tag{A.64}$$

(puisque, précisément, U qui est unitaire, conserve la norme-2), et :

$$u^*\,A^*\,A\,u = v^*\,\Lambda\,v \tag{A.65}$$

de sorte que :

$$\begin{aligned}
(\|A\|_2)^2 &= \max_{v \in K^M,\ v \neq 0} \frac{v^*\,\Lambda\,v}{v^*\,v} \\
&= \max_{v \in K^M,\ v \neq 0} \left(\lambda_1 \frac{|v_1|^2}{\sum_k |v_k|^2} + \lambda_2 \frac{|v_2|^2}{\sum_k |v_k|^2} + \ldots + \lambda_M \frac{|v_M|^2}{\sum_k |v_k|^2} \right) \\
&= \max_k \lambda_k \\
&= \rho\,(A^*\,A)
\end{aligned} \tag{A.66}$$

où ρ désigne le rayon spectral commun aux matrices semblables Λ et $A^*\,A$. En conclusion :

$$\boxed{\forall\, A \in \mathcal{M}_{M \times M}\ ,\ \|A\|_2 = \sqrt{\rho\,(A^*\,A)}} \tag{A.67}$$

A.2.6. Relations d'équivalence entre les normes 2-induite et euclidienne

On a :

$$\frac{1}{M}\,\text{Trace}\,(A^*\,A) \leq \rho\,(A^*\,A) \leq \text{Trace}\,(A^*\,A) \tag{A.68}$$

Exercice A.7 (Norme 2 induite et norme euclidienne d'une matrice)

Justifier cette assertion.

En utilisant (A.36) on en déduit :

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\,,\ \frac{1}{\sqrt{M}}\,\|A\|_E \leq \|A\|_2 \leq \|A\|_E \leq \sqrt{M}\,\|A\|_2\,.} \tag{A.69}$$

A.2.7. Relations d'équivalence entre les normes 2 et infinie induites

En vertu de (A.46), on sait que :

$$\|A\|_2 \leq M^{\frac{1}{2}}\,\|A\|_\infty\,. \tag{A.70}$$

On a également déjà vu que :

$$\|A\|_2 = \rho\,(A^*\,A) \geq \frac{\|A\|_E}{\sqrt{M}} \tag{A.71}$$

Or de plus,

$$\|A\|_E \geq \|A\|_{\max} \geq \frac{\|A\|_\infty}{M} \tag{A.72}$$

de sorte que :

$$\boxed{\begin{array}{l}\forall\, A \in \mathcal{M}_{M\times M}\,,\ \frac{1}{M^{\frac{3}{2}}}\,\|A\|_\infty \leq \|A\|_2 \leq M^{\frac{1}{2}}\,\|A\|_\infty\,,\\ \qquad\qquad \frac{1}{M^{\frac{1}{2}}}\,\|A\|_2 \leq \|A\|_\infty \leq M^{\frac{3}{2}}\,\|A\|_2\,.\end{array}} \tag{A.73}$$

Exercice A.8 (Propriétés de l'application $p \to \|A\|_p$)

Soient deux vecteurs non nuls u et v de $\mathbb{R}^M$ et la matrice $A = u\,v^T$.

(1) Démontrer que

$$\|A\|_1 = \|u\|_1 \, \|v\|_\infty \tag{A.74}$$

$$\|A\|_2 = \|u\|_2 \, \|v\|_2 \tag{A.75}$$

$$\|A\|_\infty = \|u\|_\infty \, \|v\|_1 \tag{A.76}$$

(2) L'application $p \in [1, +\infty] \to \|A\|_p$ est-elle monotone ?

(3) Donner en fonction de M les majorations les plus serrées de $\|A\|_p / \|A\|_q$ correspondant à $p,\ q = 1,\ 2$ ou ∞ de toutes les manières possibles.

A.2.8. Extension de l'inégalité de Hölder aux matrices

Soit $A \in \mathcal{M}_{M\times M}$ quelconque ; on a :

$$\begin{aligned} (\|A\|_2)^2 &= \rho\,(A^*\,A) \\ &\leq \|A^*\,A\| \\ &\leq \|A^*\| \, \|A\| \end{aligned} \tag{A.77}$$

quelle que soit la norme induite $\|\ \|$ choisie. En particulier, en prenant la norme-infinie induite on aboutit au

Théorème A.5

« *Inégalité de Hölder* »

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\ ,\ (\|A\|_2)^2 \leq \|A\|_1 \, \|A\|_\infty\,.} \tag{A.78}$$

On a également le

Théorème A.6

Soient deux réels positifs p et q tels que

$$\frac{1}{p} + \frac{1}{q} = 1\ ;\ p \geq 2 \geq q \geq 1 \tag{A.79}$$

on a :

$$\boxed{\forall\, A \in \mathcal{M}_{M\times M}\ ,\ (\|A\|_2)^2 \leq M^{\frac{1}{2}-\frac{1}{p}} \, \|A\|_p \, \|A\|_q = M^{\frac{1}{q}-\frac{1}{2}} \, \|A\|_p \, \|A\|_q\,.} \tag{A.80}$$

DÉMONSTRATION : pour un certain vecteur ξ tel que

$$\|\xi\|_2 = 1 \tag{A.81}$$

on a :

$$\begin{aligned}
(\|A\|_2)^2 &= (\|A\,\xi\|_2)^2 \\
&\leq \|A\,\xi\|_p \, \|A\,\xi\|_q \text{ (en vertu de (A.23))} \\
&\leq \frac{\|A\,\xi\|_p}{\|\xi\|_p} \, \frac{\|A\,\xi\|_q}{\|\xi\|_q} \, \|\xi\|_p \, \|\xi\|_q \\
&\leq \|A\|_p \, \|A\|_q \, \|\xi\|_p \, \|\xi\|_q
\end{aligned} \tag{A.82}$$

Or, en vertu de (A.13), on a, du fait que $p \geq 2 \geq q \geq 1$:

$$\|\xi\|_p \leq \|\xi\|_2 = 1\,, \; \|\xi\|_q \leq M^{\frac{1}{q}-\frac{1}{2}} \, \|\xi\|_2 = M^{\frac{1}{2}-\frac{1}{p}}\,, \tag{A.83}$$

d'où le résultat. □

Annexe B

Traitement algébrique de problèmes multidimensionnels

Lorsqu'un opérateur continu en plusieurs dimensions d'espace est la somme d'opérateurs unidimensionnels (conditions aux limites comprises), que le domaine est un produit cartésien, que le maillage est lui-même le produit tensoriel de discrétisations d'intervalles suivant les axes respectifs de coordonnées, la matrice d'approximation par différences finies de cet opérateur est la somme directe (ou tensorielle) des matrices d'approximation associées à ces opérateurs unidimensionnels. La structure des vecteurs propres et du spectre associé de valeurs propres peut alors se déduire immédiatement de la connaissance des diagonalisations de problèmes 1D. Cette annexe a pour but de préciser ce résultat, et pour cela, on introduit d'abord l'algèbre matricielle de Kronecker ; on étudie ensuite le cas représentatif d'un opérateur elliptique satisfaisant les hypothèses citées.

B.1. Algèbre (matricielle) de Kronecker

Définition B.1

« Produit de Kronecker » (ou « produit tensoriel ») de matrices

Etant donné deux matrices rectangulaires

$$A = \{a_{p,q}\} \ (p = 1, 2, ..., M \ ; \ q = 1, 2, ..., M') \tag{B.1}$$

et

$$B = \{b_{r,s}\} \ (r = 1, 2, ..., L \ ; \ s = 1, 2, ..., L') \tag{B.2}$$

on appelle produit tensoriel de ces matrices, la matrice de dimension $ML \times M'L'$ dont la structure est la suivante :

$$A \otimes B \stackrel{\text{déf}}{=} \begin{pmatrix} a_{1,1}B & a_{1,2}B & \dots & a_{1,M'}B \\ a_{2,1}B & a_{2,2}B & \dots & a_{2,M'}B \\ \vdots & \vdots & \dots & \vdots \\ a_{M,1}B & a_{M,2}B & \dots & a_{M,M'}B \end{pmatrix}. \tag{B.3}$$

On admet (ou on se convainc rapidement) que les relations suivantes sont vraies identiquement pourvu que les dimensions des matrices soient compatibles :

$$\boxed{\begin{array}{c}(A \otimes B)\,(C \otimes D) = (AC) \otimes (BD) \\ (A \otimes B)^T = A^T \otimes B^T\end{array}} \tag{B.4}$$

(sans inversion de l'ordre des facteurs).

Cas des matrices carrées : $M = M'$ et $L = L'$.

Si les matrices A et B sont inversibles on a :

$$\boxed{(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}} \tag{B.5}$$

(sans inversion de l'ordre des facteurs).

Si A et B sont diagonalisables,

$$\boxed{\begin{array}{c}A = X\,\mathcal{A}\,X^{-1} \\ B = Y\,\mathcal{B}\,Y^{-1}\end{array}} \tag{B.6}$$

où la matrice $X = \{X_{p,q}\}$ des vecteurs propres de A et $\mathcal{A}$ sont de dimension $M \times M$, X est inversible et $\mathcal{A}$ diagonale,

$$\mathcal{A} = \begin{pmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_M \end{pmatrix} \tag{B.7}$$

la matrice $Y = \{Y_{r,s}\}$ des vecteurs propres de B et $\mathcal{B}$ sont de dimension $L \times L$, Y est inversible et $\mathcal{B}$ diagonale,

$$\mathcal{B} = \begin{pmatrix} \beta_1 & & & \\ & \beta_2 & & \\ & & \ddots & \\ & & & \beta_L \end{pmatrix} \tag{B.8}$$

alors on a :

$$\boxed{A \otimes B = (X \otimes Y)\ (\mathcal{A} \otimes \mathcal{B})\ (X \otimes Y)^{-1}} \tag{B.9}$$

ce qui signifie qu'on obtient les vecteurs propres de $A \otimes B$ en formant le produit tensoriel du vecteur propre

$$X^{(m)} = \{X_{j,m}\}\ (j = 1, 2, ..., M)\ (m = 1, 2, ..., M) \tag{B.10}$$

de la matrice A avec le vecteur propre

$$Y^{(\ell)} = \{Y_{k,\ell}\}\ (k = 1, 2, ..., L)\ (\ell = 1, 2, ..., L) \tag{B.11}$$

de la matrice B de toutes les $M\,L$ manières possibles ; la valeur propre associée est le nombre :

$$\boxed{\lambda_{m,\ell} = \alpha_m\, \beta_\ell} \tag{B.12}$$

Définition B.2 (Somme directe de deux matrices carrées)

On appelle *somme directe* des deux matrices carrées $A \in \mathcal{M}_{M \times M}$ et $B \in \mathcal{M}_{L \times L}$ la matrice de dimension $M\,L \times M\,L$ suivante :

$$A \oplus B \stackrel{\text{déf}}{=} A \otimes I_L + I_M \otimes B \tag{B.13}$$

Enfin, on généralise cette définition comme suit :

Définition B.3

Etant donné un polynôme des indéterminées ξ et η :

$$P(\xi, \eta) = \sum_{\nu,\mu} c_{\nu,\mu}\, \xi^\nu\, \eta^\mu \tag{B.14}$$

où les $c_{\nu,\mu}$ sont des coefficients, on pose

$$P\,(A,B) = \sum_{\nu,\mu} c_{\nu,\mu}\, A^{\nu} \otimes B^{\mu} \tag{B.15}$$

(où par convention $A^0 = I_M$ et $B^0 = I_L$).

En particulier, la somme directe correspond au polynôme

$$P(\xi,\eta) = \xi + \eta \tag{B.16}$$

Théorème B.1 (Diagonalisation d'un polynôme de matrices)

Si les matrices A et B sont diagonalisables, la matrice $P(A,B)$ l'est aussi, et avec les notations de (B.6)-(B.7)-(B.8),

$$\boxed{P(A,B) = (X \otimes Y)\ P(\mathcal{A},\mathcal{B})\ (X \otimes Y)^{-1}} \tag{B.17}$$

de sorte que les valeurs propres de la matrice $P(A,B)$ sont les nombres

$$\boxed{\lambda_{m,\ell} = P\,(\alpha_m, \beta_\ell)} \tag{B.18}$$

$(m = 1,2,...,M,\ \ell = 1,2,...,L)$, et les vecteurs propres associés s'obtiennent en composant les produits tensoriels suivants :

$$\boxed{Z^{(m,\ell)} = X^{(m)} \otimes Y^{(\ell)}} \tag{B.19}$$

où le vecteur colonne $X^{(m)}$ est le vecteur propre de la matrice A associé à la valeur propre α_m et le vecteur colonne $Y^{(\ell)}$ est le vecteur propre de la matrice B associé à la valeur propre β_ℓ.

DÉMONSTRATION : on a :

$$A^{\nu} = X\,\mathcal{A}^{\nu}\,X^{-1}\,,\ B^{\mu} = Y\,\mathcal{B}^{\mu}\,Y^{-1} \tag{B.20}$$

de sorte qu'en utilisant plusieurs fois (B.4) puis (B.5), on obtient :

$$\begin{aligned} A^\nu \otimes B^\mu &= (\underbrace{X}_{A'}\ \underbrace{\mathcal{A}^\nu\, X^{-1}}_{C'}) \otimes (\underbrace{Y}_{B'}\ \underbrace{\mathcal{B}^\mu\, Y^{-1}}_{D'}) \\ &= (A' \otimes B')\,(C' \otimes D') \\ &= (X \otimes Y)[(\underbrace{\mathcal{A}^\nu}_{A''}\ \underbrace{X^{-1}}_{C''}) \otimes (\underbrace{\mathcal{B}^\mu}_{B''}\ \underbrace{Y^{-1}}_{D''})] \\ &= (X \otimes Y)\,(A'' \otimes B'')\,(C'' \otimes D'') \\ &= (X \otimes Y)\ (\mathcal{A}^\nu \otimes \mathcal{B}^\mu)\ \left(X^{-1} \otimes Y^{-1}\right) \\ &= (X \otimes Y)\ (\mathcal{A}^\nu \otimes \mathcal{B}^\mu)\ (X \otimes Y)^{-1} \end{aligned} \tag{B.21}$$

et par linéarité,

$$\begin{aligned} P(A,B) &= (X \otimes Y)\ \left(\sum c_{\nu,\mu}\, \mathcal{A}^\nu \otimes \mathcal{B}^\mu\right)\ (X \otimes Y)^{-1} \\ &= (X \otimes Y)\ P\,(\mathcal{A},\mathcal{B})\ (X \otimes Y)^{-1} \end{aligned} \tag{B.22}$$

Cette équation dans laquelle la matrice $P\,(\mathcal{A},\mathcal{B})$ est diagonale exprime la diagonalisation de la matrice $P(A,B)$, d'où le résultat. □

Corollaire B.1

(Mêmes hypothèses et notations.)
Les valeurs propres de la somme directe $A \oplus B$ sont les nombres :

$$\boxed{\lambda_{m,\ell} = \alpha_m + \beta_\ell} \tag{B.23}$$

$(m = 1,2,...,M,\ \ell = 1,2,...,L)$ *et les vecteurs propres associés sont donnés par (B.19).*

B.2. Etude d'un opérateur elliptique

On considère l'opérateur elliptique suivant :

$$\boxed{\mathcal{L} = -\sigma_x \frac{\partial^2}{\partial x^2} - \sigma_y \frac{\partial^2}{\partial y^2}} \tag{B.24}$$

dans lequel σ_x et σ_y sont deux réels strictement positifs et

$$\begin{cases} 0 \le x \le L_x \\ 0 \le y \le L_y \end{cases} \tag{B.25}$$

On suppose que l'opérateur $\mathcal{L}$ s'applique à des fonctions $u(x,y)$ satisfaisant des conditions de Dirichlet homogènes :

$$\left\{ \begin{array}{l} \forall\, y \in [0, L_y]\,,\ u(0,y) = u(L_x,y) = 0 \\ \forall\, x \in [0, L_x]\,,\ u(x,0) = u(x,L_y) = 0 \end{array} \right. \tag{B.26}$$

On considère une discrétisation uniforme du domaine en un maillage qui est un produit cartésien. En d'autres termes, le nœud courant intérieur au domaine a pour coordonnées :

$$\left\{ \begin{array}{l} x_{j,k} = j\, h_x\,,\ (j = 1, 2, ..., M) \\ y_{j,k} = k\, h_y\,,\ (k = 1, 2, ..., L) \end{array} \right. \tag{B.27}$$

où $h_x = \Delta x = L_x/(M+1)$ et $h_y = \Delta y = L_y/(L+1)$.

On note $u_{j,k}$ l'approximation de $u(x,y)$ au nœud (j,k).

Exercice B.1 (Approximation centrée d'un opérateur elliptique en 2D)

Expliciter l'approximation classique par différences finies centrées du résidu discret

$$\boxed{r_{j,k} \overset{\text{déf}}{=} \mathcal{L} u_{j,k}} \tag{B.28}$$

en fonction des valeurs nodales de u aux nœuds voisins de (j,k).

Les degrés de liberté du problème sont les valeurs aux nœuds intérieurs :

$$\{u_{j,k}\}\,,\ (j = 1, 2, ..., M)\,,\ (k = 1, 2, ..., L) \tag{B.29}$$

étant entendu que les valeurs aux bords

$$\begin{array}{l} \{u_{0,k}\}\,,\ \{u_{M+1,k}\}\,,\ (k = 0, 1, 2, ..., L+1)\,, \\ \{u_{j,0}\}\,,\ \{u_{j,L+1}\}\,,\ (j = 0, 1, 2, ..., M+1)\,, \end{array} \tag{B.30}$$

sont nulles en vertu des conditions aux limites de Dirichlet. On choisit de stocker ces

degrés de liberté « par colonnes », c'est-à-dire dans le vecteur :

$$u_h = \begin{pmatrix} u_{1,1} \\ u_{1,2} \\ \vdots \\ u_{1,L} \\ -- \\ u_{2,1} \\ u_{2,2} \\ \vdots \\ u_{2,L} \\ -- \\ \vdots \\ -- \\ u_{M,1} \\ u_{M,2} \\ \vdots \\ u_{M,L} \end{pmatrix} \tag{B.31}$$

On ordonne de manière analogue les ML valeurs nodales $\{r_{j,k}\}$ $(j = 1, 2, ..., M)$ $(k = 1, 2, ..., L)$ dans un vecteur noté r_h. On peut donc construire une matrice A_h de dimension $ML \times ML$ telle que :

$$r_h = A_h \, u_h \tag{B.32}$$

La matrice A_h est la représentation discrète de l'opérateur $\mathcal{L}$.

Exercice B.2 (Somme directe d'opérateurs, spectre)

(1) Montrer que la matrice A_h est une somme directe :

$$\boxed{A_h = A_{h\,x} \oplus A_{h\,y}} \tag{B.33}$$

dans laquelle les matrices $A_{h\,x}$ et $A_{h\,y}$ sont de dimension $M \times M$ et $L \times L$ respectivement et de structure bien connue.

(2) On note $\{\alpha_m = \underline{\lambda}_m^{h_x}\}$ $(m = 1, 2, ..., M)$ et $\{\beta_\ell = \underline{\lambda}_\ell^{h_y}\}$ $(\ell = 1, 2, ..., L)$ les valeurs propres des matrices $A_{h\,x}$ et $A_{h\,y}$ respectivement. Rappeler l'expression de

α_m et β_ℓ, et identifier le spectre de la matrice A_h :

$$\boxed{\underline{\lambda}^h_{m,\ell} = \underline{\lambda}^{h_x}_m + \underline{\lambda}^{h_y}_\ell \quad (m = 1, 2, ..., M\ ;\ \ell = 1, 2, ..., L)} \tag{B.34}$$

Annexe C

Polynômes de Tchebychev

Dans cette annexe, on passe en revue les principales propriétés des polynômes de Tchebychev qui jouent un rôle essentiel dans l'optimisation de la méthode de Richardson constituée d'un cycle d'itérations de Jacobi (cf. chapitre 3). On notera que ce problème d'optimisation est très proche du problème classique d'identification de la « meilleure approximation » d'une fonction régulière par un polynôme d'interpolation de Lagrange dans lequel on optimise les points d'interpolation (cf. tout ouvrage introductif d'analyse numérique sur l'interpolation polynômiale des fonctions tel que [31]) ; dans ce cas là, les polynômes de Tchebychev interviennent pratiquement de la même manière qu'au chapitre 3, seule change une normalisation.

On part de la « Formule de Moivre » :

$$\cos\ k\theta + i\ \sin\ k\theta = e^{i(k\theta)} = e^{k(i\theta)} = (\cos\ \theta + i\ \sin\ \theta)^k \tag{C.1}$$

de laquelle on tire :

$$\begin{aligned}
\cos\ k\theta &= \Re\left((\cos\ \theta + i\ \sin\ \theta)^k\right) \\
&= \Re\left(\sum_{n=0}^{k} C_k^n\ \cos^{k-n}\ \theta\,(i\ \sin\ \theta)^n\right) \\
&= \sum_{m=0}^{E(k/2)} C_k^{2m}\ \cos^{k-2m}\ \theta\ \times(-1)^m(1-\cos^2\ \theta)^m \\
&\stackrel{\text{déf}}{=} T_k(\cos\ \theta)
\end{aligned} \tag{C.2}$$

où $E(.)$ désigne la partie entière et $T_k(x)$ est le polynôme suivant :

$$\boxed{T_k(x) \stackrel{\text{déf}}{=} \sum_{m=0}^{E(k/2)} C_k^{2m}\, x^{k-2m}\, (x^2-1)^m} \qquad \text{(C.3)}$$

Exercice C.1 (Propriétés des polynômes de Tchebychev)

(1) Montrer que le polynôme $T_k(x)$ est de degré k exactement. (Identifier précisément le coefficient α_k de x^k.) Identifier sa parité.

(2) Montrer que :

$$\boxed{T_k(x) = \begin{cases} \cos\,(k\,\text{Arccos}\,x) \;\text{ si } x \in [-1,1]\,, \\ \text{ch}\,(k\,\text{Argch}\;\;\text{x}) \;\text{ si } x > 1\,. \end{cases}} \qquad \text{(C.4)}$$

Que peut-on dire dans le cas où $x < -1$? Calculer les valeurs de $T_k(1)$ et $T_k(-1)$?

(3) Montrer que :

$$\boxed{\forall\, x \in \mathbb{R},\; T_{k+1}(x) = 2x\, T_k(x) - T_{k-1}(x)\,.} \qquad \text{(C.5)}$$

(4) Générer les premiers éléments, identifier zéros et extrêmas. Courbes représentatives. (Vérifier que vos résultats sont conformes à ceux du tableau F.6 et de la figure F.10 qui rassemblent les résultats d'un programme en langage MAPLE qu'on pourra reconstruire.)

(5) Montrer qu'il s'agit d'une famille de polynômes orthogonaux vis-à-vis du produit scalaire suivant défini sur les fonctions continues sur [-1,1] :

$$(u,v) = \int_{-1}^{1} u(x)\, v(x)\, \frac{dx}{\sqrt{1-x^2}}\,. \qquad \text{(C.6)}$$

Annexe D

Mise en évidence du paramètre β

On revient ici sur la signification du paramètre β qui intervient dans le développement du rayon spectral en fonction de la finesse h de la maille.

On suppose que l'opérateur continu A est auto-adjoint défini positif, et que l'opérateur discret A_h lui est consistant, par exemple :

$$A = -\frac{\partial^2}{\partial x^2} \tag{D.1}$$

opérant sur un sous-ensemble de $L^2\left([a,b]\right)$ de fonctions $u(x)$ satisfaisant certaines conditions aux limites (e.g. Dirichlet homogènes, $u(a) = u(b) = 0$), et

$$A_h = h^{-2}\,\text{Trid}\,\left(-1, 2, -1\right) \tag{D.2}$$

opérant sur les discrétisés de telles fonctions. Plus généralement :

$$A_h = h^{-\beta}\,\Delta \tag{D.3}$$

où β est l'ordre de l'opérateur continu et Δ est ici un opérateur de différence (non divisée) quelconque et non le laplacien.

On sait que A admet une base dénombrable de fonctions propres que l'on peut ordonner suivant les valeurs croissantes des valeurs propres associées. Par exemple, pour l'opérateur de dérivée seconde et des conditions aux limites de Dirichlet homogènes, les fonctions propres sont sinusoïdales,

$$u_m(x) = \sin\,\frac{m\,\pi\,x}{b-a} \tag{D.4}$$

et les valeurs propres sont réelles-positives :

$$\lambda_m(A) = \left(\frac{m\,\pi}{b-a}\right)^2 \tag{D.5}$$

La matrice A_h, de dimension $M \times M$, admet M vecteurs propres et M valeurs propres associées qui, dans le cas général, sont, dans un certain sens, des approximations de leurs homologues continus. En fait, on s'attend à ce que l'approximation soit bonne pour les premiers modes seulement, c'est-à-dire les « modes de basses fréquences ». Dans le cas particulier de l'opérateur de différence-divisée seconde, les vecteurs propres sont (tous) précisément les discrétisés des M premières fonctions propres :

$$u_{j,m} = u_m(x_j) \tag{D.6}$$

et les valeurs propres associées sont données par :

$$\lambda_m(A_h) = \frac{2 - 2\cos\left(\frac{m\,\pi}{M+1}\right)}{h^2} = \frac{4\sin^2\left(\frac{m\,\pi}{2\,(M+1)}\right)}{h^2} \tag{D.7}$$

où $(M+1)\,h = b-a$. On constate que pour m petit on a bien :

$$\lambda_m(A_h) \approx \lambda_m(A) \tag{D.8}$$

Pour cette raison, on a en particulier pour $m = 1$:

$$\lambda_{\min}(A_h) \approx \lambda_{\min}(A) = O(1) \tag{D.9}$$

Par contre, les valeurs propres de la matrice A_h associées aux modes de hautes fréquences n'approchent pas celles du spectre de l'opérateur continu, qui d'ailleurs sont non bornées. On écrira plutôt :

$$\lambda_{\max}(A_h) \approx \frac{\|\Delta\|_\infty}{h^\beta} = \left(\frac{b-a}{h}\right)^\beta \frac{\|\Delta\|_\infty}{(b-a)^\beta} = O\left(\frac{b-a}{h}\right)^\beta \tag{D.10}$$

Par conséquent si κ est le nombre de conditionnement du système discret, on a :

$$\frac{1}{\kappa} = \frac{\lambda_{\min}(A_h)}{\lambda_{\max}(A_h)} = O\left(\frac{h}{b-a}\right)^\beta \tag{D.11}$$

Dans cette expression la quantité $(b-a)/h$ est un infiniment grand asymptotique au nombre de degrés de liberté.

Si maintenant on résout le système discret par la méthode de Jacobi avec un seul pseudo-pas de temps optimisé, le rayon spectral de la méthode itérative s'exprime comme suit :

$$\rho_\delta = \frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa} + \cdots = 1 - 2/\left(\frac{b-a}{h}\right)^\beta + \ldots \stackrel{\text{déf}}{=} 1 - C_b\,h^\beta + \ldots \tag{D.12}$$

Si pour résoudre on applique plutôt la méthode de Richardson avec k pseudo-pas de temps optimisés, tout se passe comme si le nombre de conditionnement était réduit d'un facteur k, et de même pour la constante C_b.

Enfin, si on résout par la méthode multigrille complète, la formule donnant ρ_δ est encore valable avec $\beta = 0$. Dans ce cas, aucune économie en coût n'est réalisée par l'algorithme multiplicatif de Schwarz.

Annexe E

Rayon spectral d'un cycle bigrille idéal

Il s'agit de vérifier que le rayon spectral du V-Cycle Bigrille Idéal admet une borne supérieure indépendante du nombre de points de discrétisation.

On se place dans le cadre du problème test usuel (« Laplacien 1D ») :

$$\boxed{\begin{array}{c} -u_{xx} = f \quad (0 \leq x \leq 1) \\ u(0) = u(1) = 0 \end{array}} \tag{E.1}$$

que l'on discrétise sur un maillage uniforme « fin » pour lequel

$$h = \frac{1}{M_h + 1} \tag{E.2}$$

où M_h est le nombre de degrés de liberté ($M_h = 2^I \mu - 1$, où μ est une constante entière) afin d'obtenir le système discret suivant :

$$\boxed{A_h \, u_h = f_h} \tag{E.3}$$

où :

$$A_h = h^{-2} \, \mathrm{Trid}_{DD} \, (-1, 2, -1) \tag{E.4}$$

L'indice $_{DD}$ est indiqué pour rappeler que les conditions aux limites sont de type Dirichlet aux deux bords. On connaît en particulier la diagonalisation de la matrice

A_h :

$$A_h = h^{-2}\, S_h\, \Lambda_h\, S_h^{-1} \tag{E.5}$$

où la matrice S_h qui représente discrètement la transformation en somme de fonctions sinus, et dont le coefficient général admet l'expression suivante :

$$(S_h)_{j,k} = \sqrt{\frac{2}{M_h+1}}\, \sin(j\,\theta_k) = \sqrt{2h}\, \sin(jk\pi h) \tag{E.6}$$

est symétrique ; d'autre part, cette matrice est orthogonale de sorte que finalement :

$$S_h^{-1} = S_h^T = S_h \tag{E.7}$$

De plus les valeurs propres sont bien connues :

$$\lambda_k = (\Lambda_h)_{k,k} = 2 - 2\cos\,\theta_k = 2 - 2\cos(k\pi h) \tag{E.8}$$

On considère d'autre part une grille « grossière », deux fois moins fine :

$$M_{2h} = 2^{I-1}\,\mu - 1 \tag{E.9}$$

On rappelle que la matrice d'amplification du V-cycle bigrille idéal admet l'expression suivante :

$$\boxed{G = G_h\, \left\{ I - I_{2h}^h \left(I_h^{2h}\, A_h\, I_{2h}^h \right)^{-1}\, I_h^{2h}\, A_h \right\}\, G_h} \tag{E.10}$$

dans laquelle :

– G_h est le facteur d'amplification de la phase de lissage ; on suppose que cette phase correspond à la méthode de Richardson avec 3 pas de temps $\{\tau_\ell\}$ choisis pour atténuer optimalement les hautes fréquences, de sorte que :

$$G_h = \left(I - \tau_3\, h^2\, A_h\right)\, \left(I - \tau_2\, h^2\, A_h\right)\, \left(I - \tau_1\, h^2\, A_h\right) \tag{E.11}$$

où :

$$\tau_\ell^{-1} = \frac{4+2}{2} + \frac{4-2}{2}\,\xi_\ell \quad (\ell = 1,\ 2,\ 3) \tag{E.12}$$

et $\xi_1 = -\sqrt{3}/2,\ \xi_2 = 0,\ \xi_3 = \sqrt{3}/2$ sont les zéros du polynôme de Tchebychev de degré trois, $T_3(\xi) = 4\xi^3 - 3\xi$.

– $I_{2h}^{h} = P$ est l'opérateur de prolongement ; on suppose qu'on utilise un point sur deux l'injection directe ou l'interpolation linéaire ; par exemple dans le cas où $M_{2h} = 3$, $M_h = 7$, cet opérateur est représenté par la matrice suivante :

$$P = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \tag{E.13}$$

– $I_{h}^{2h} = R$ est l'opérateur de restriction ; on prend

$$R = \frac{1}{2} P^T \tag{E.14}$$

On vérifie alors aisément que l'on a :

$$\boxed{R\, A_h\, P = A_{2h} = \frac{1}{4h^2}\, S_{2h}\, \Lambda_{2h}\, S_{2h}} \tag{E.15}$$

On rappelle que si A et B sont deux matrices carrées de même dimension, les matrices AB et BA sont semblables (mêmes valeurs propres). Il résulte alors des définitions que la matrice G est semblable aux matrices suivantes :

$$\begin{aligned} G &\sim \left\{ I - 2\, R^T \left(4h^2\, S_{2h}\, \Lambda_{2h}^{-1}\, S_{2h} \right)\, R\, A_h \right\} \left[\prod_{\ell=1}^{3} \left(I - \tau_\ell\, h^2\, A_h \right) \right]^2 \\ &\sim \left\{ I - 8 \underbrace{S_h\, R^T\, S_{2h}}_{\sigma^T}\, \Lambda_{2h}^{-1}\, \underbrace{S_{2h}\, R\, S_h}_{\sigma}\, \Lambda_h \right\} \left[\prod_{\ell=1}^{3} \left(I - \tau_\ell\, \Lambda_h \right) \right]^2 = \Sigma\, D^2 \\ &\sim D\, \Sigma\, D \end{aligned} \tag{E.16}$$

où l'on a posé :

$$\boxed{\begin{aligned} &\sigma = S_{2h}\, R\, S_h \\ &\Sigma = \Lambda_h^{-1} - 8\, \sigma^T\, \Lambda_{2h}^{-1}\, \sigma \quad \text{(symétrique)} \\ &D = \sqrt{\Lambda_h} \prod_{\ell=1}^{3} \left(I - \tau_\ell\, \Lambda_h \right) \quad \text{(diagonale)} \end{aligned}} \tag{E.17}$$

On est donc ramené à examiner les valeurs propres de la matrice symétrique $D\,\Sigma\,D$ pour en déduire le rayon spectral de l'opérateur G :

$$\boxed{\rho = \max_m |\,\lambda_m\,(D\,\Sigma\,D)|} \tag{E.18}$$

afin d'en observer le comportement lorsque M_h croît.

Nouvelle expression des matrices Σ et D

Pour $j = 1, 2, ..., M_{2h}$ et $k = 1, 2, ..., M_h$:

$$(R\,S_h)_{j,k} = \sum_{i=1}^{M_h} R_{j,i}\;(S_h)_{i,k} \tag{E.19}$$

et pour $m = 1, 2, ..., M_{2h}$:

$$\begin{aligned}\sigma_{m,k} &= \sum_{j=1}^{M_{2h}} (S_{2h})_{m,j}\;(R\,S_h)_{j,k} \\ &= \sum_{i=1}^{M_h}\sum_{j=1}^{M_{2h}} (S_{2h})_{m,j}\;R_{j,i}\;(S_h)_{i,k}\end{aligned} \tag{E.20}$$

Or pour j fixé, $R_{j,i} = 0$ sauf si $i = 2j-1$, $2j$ ou $2j+1$ ce qui donne :

$$\begin{aligned}\sigma_{m,k} &= \frac{1}{4}\sum_{j=1}^{M_{2h}} (S_{2h})_{m,j}\;\left[(S_h)_{2j-1,k} + 2\,(S_h)_{2j,k} + (S_h)_{2j+1,k}\right] \\ &= \frac{h}{\sqrt{2}}\,\tilde{\sigma}_{m,k}\end{aligned} \tag{E.21}$$

où l'on a posé :

$$\boxed{\tilde{\sigma}_{m,k} = \sum_{j=1}^{M_{2h}} (s_h)_{2m,j}\;\left[(s_h)_{2j-1,k} + 2\;(s_h)_{2j,k} + (s_h)_{2j+1,k}\right]} \tag{E.22}$$

et :

$$\boxed{(s_h)_{j,k} = \frac{1}{\sqrt{2h}}\;(S_h)_{j,k} = \sin(jk\pi h)} \tag{E.23}$$

Il en résulte que :

$$\Sigma = \Lambda_h^{-1} - 4h^2\,\tilde{\sigma}^T\,\Lambda_{2h}^{-1}\,\tilde{\sigma} \tag{E.24}$$

et puisque Λ_{2h} est une matrice diagonale de dimension $M_{2h} \times M_{2h}$, on a :

$$\boxed{\Sigma_{j,k} = \frac{\delta_{j,k}}{(\lambda_h)_j} - 4h^2 \sum_{m=1}^{M_{2h}} \frac{\tilde{\sigma}_{m,j}\,\tilde{\sigma}_{m,k}}{(\lambda_h)_{2m}}} \tag{E.25}$$

Enfin D est la matrice diagonale dont les éléments diagonaux sont donnés par :

$$\boxed{D_{k,k} = d_k = d\big((\lambda_h)_k\big)} \tag{E.26}$$

où par suite du choix fait des paramètres $\{\tau_\ell\}$, $d(\lambda)$ est la fonction :

$$\boxed{d(\lambda) = \sqrt{\lambda}\,\prod_{\ell=1}^{3}(1-\tau_\ell\,\lambda) = \sqrt{\lambda}\,\frac{T_3(3-\lambda)}{T_3(3)}} \tag{E.27}$$

Calcul de la matrice $\tilde{\sigma}$

On remarque que :

$$\begin{aligned}(s_h)_{2j-1,k} + 2\,(s_h)_{2j,k} + (s_h)_{2j+1,k} &\\ &= \sin[(2j-1)k\pi h] + 2\sin[2jk\pi h] + \sin[(2j+1)k\pi h]\\ &= 2\sin[2jk\pi h]\,[1+\cos(k\pi h)]\\ &= 4\cos^2\frac{k\pi h}{2}\,\sin(2jk\pi h)\end{aligned} \tag{E.28}$$

de sorte que :

$$\tilde{\sigma}_{m,k} = 4\cos^2\frac{k\pi h}{2}\sum_{j=1}^{M_{2h}}\sin(2mj\pi h)\,\sin(2kj\pi h) \tag{E.29}$$

soit encore

$$\boxed{\tilde{\sigma}_{m,k} = 2\cos^2\frac{k\pi h}{2}\,(C_{m-k} - C_{m+k})} \tag{E.30}$$

où l'on a posé :

$$\boxed{C_\nu = \sum_{j=1}^{M_{2h}} \cos\left(2\nu j\pi h\right)} \tag{E.31}$$

Calculons les coefficients C_ν. Si $\nu = 0$ (modulo $M_h + 1$), le nombre νh est un entier relatif et :

$$C_\nu = M_{2h} \tag{E.32}$$

Dans le cas inverse, $\nu \neq 0$ (modulo $M_h + 1$), on a :

$$C_\nu = \Re\left(E_\nu\right) \tag{E.33}$$

en posant :

$$E_\nu = \sum_{j=1}^{M_{2h}} \exp\left(2\nu ij\pi h\right) \tag{E.34}$$

Or :

$$\begin{aligned} E_\nu &= \exp\left(2\nu i\pi h\right) \sum_{j=1}^{M_{2h}} \exp\left[2\nu i(j-1)\pi h\right] \\ &= \exp\left(2\nu i\pi h\right) \frac{1 - \exp\left(2\nu i M_{2h}\pi h\right)}{1 - \exp\left(2\nu i\pi h\right)} \\ &= \frac{\exp\left(\nu i\pi h\right)}{-2i\sin\left(\nu\pi h\right)} \left[1 - \exp\left(2\nu i M_{2h}\pi h\right)\right] \\ &= \frac{i}{2\sin\left(\nu\pi h\right)} \left\{\exp\left(\nu i\pi h\right) - \exp\left[\left(2M_{2h}+1\right)\nu i\pi h\right]\right\} \end{aligned} \tag{E.35}$$

Or $2M_{2h} + 1 = M_h = (M_h + 1) - 1$ de sorte que :

$$\exp\left[\left(2M_{2h}+1\right)\nu i\pi h\right] = \exp\left(\nu i\pi - \nu i\pi h\right) = (-1)^\nu \exp\left(-\nu i\pi h\right) \tag{E.36}$$

ce qui donne

$$E_\nu = \frac{i}{2\sin\left(\nu\pi h\right)} \left\{\exp\left(\nu i\pi h\right) - (-1)^\nu \exp\left(-\nu i\pi h\right)\right\} \tag{E.37}$$

puis :

$$C_\nu = -\frac{1}{2\sin\left(\nu\pi h\right)} \left\{\sin\left(\nu\pi h\right) - (-1)^\nu \left[-\sin\left(\nu\pi h\right)\right]\right\} = -\frac{1}{2}\left[1 + (-1)^\nu\right] \tag{E.38}$$

soit finalement

$$C_\nu = \begin{cases} 0 & \text{si } \nu \text{ est impair} \\ -1 & \text{si } \nu \text{ est pair} \end{cases} \tag{E.39}$$

En résumé :

$$C_\nu = \begin{cases} M_{2h} & \text{si } \nu = 0 \text{ (modulo } M_h + 1) \\ 0 & \text{si } \nu \neq 0 \text{ (modulo } M_h + 1) \text{ et } \nu \text{ impair} \\ -1 & \text{si } \nu \neq 0 \text{ (modulo } M_h + 1) \text{ et } \nu \text{ pair} \end{cases} \tag{E.40}$$

Revenons maintenant au calcul de l'élément $\tilde{\sigma}_{m,k}$. Pour cela, on examine les trois cas suivants :

1. $m \neq k$ et $m + k \neq M_h + 1$
2. $m = k$
3. $m + k = M_h + 1$

en supposant toujours que $1 \leq m \leq M_{2h}$ et $1 \leq k \leq M_h$.

Cas 1 : $m \neq k$ et $m + k \neq M_h + 1$ (et $1 \leq m \leq M_{2h}$ et $1 \leq k \leq M_h$).

Dans ce cas, les entiers $m - k$ et $m + k$ sont tous deux non nuls (modulo $M_h + 1$) et de même parité, de sorte que :

$$C_{m-k} = C_{m+k} \quad (= 0 \text{ ou } -1) \tag{E.41}$$

et par conséquent :

$$\tilde{\sigma}_{m,k} = 0 \tag{E.42}$$

Cas 2 : $m = k$ (et $1 \leq m \leq M_{2h}$ et forcément $1 \leq k \leq M_h$).

$$\tilde{\sigma}_{m,m} = 2\cos^2 \frac{m\pi h}{2} \, (C_0 - C_{2m}) \tag{E.43}$$

Or, $2m > 0$; de plus, $2m \leq 2M_{2h} = M_h - 1$; donc $2m \neq 0$ (modulo $M_h + 1$) ce qui implique que $C_{2m} = -1$ car $2m$ est pair. De plus, $C_0 = M_{2h}$ de sorte que :

$$\tilde{\sigma}_{m,m} = 2\cos^2 \frac{m\pi h}{2} \, (M_{2h} + 1) = (M_h + 1)\cos^2 \frac{m\pi h}{2} \tag{E.44}$$

Cas 3 : $m + k = M_h + 1$ (et $1 \leq m \leq M_{2h}$ et $1 \leq k \leq M_h$).

On a $m-k=M_h+1-2k$ de sorte que :

$$\tilde{\sigma}_{m,k} = 2\cos^2\frac{k\pi h}{2}\sum_{j=1}^{M_{2h}}\left\{\cos\left[2\left(M_h+1\right)j\pi h-4kj\pi h\right]-\cos\left[2\left(M_h+1\right)j\pi h\right]\right\}$$
$$= 2\cos^2\frac{k\pi h}{2}\left(C_{2k}-C_0\right) \tag{E.45}$$

Or, d'une part $2k>0$; d'autre part $2k\leq 2M_h$; par conséquent, la seule façon d'avoir $2k=0$ (modulo M_h+1) serait d'avoir $2k=M_h+1$, ce qui impliquerait $m=k=(M_h+1)/2=M_{2h}+1$ ce qui n'est pas admissible étant donné que $m\leq M_{2h}$. On en conclut que $2k\neq 0$ (modulo M_h+1), ce qui implique que $C_{2k}=-1$ car $2k$ est pair. Comme de plus, $C_0=M_{2h}$, on a :

$$\tilde{\sigma}_{m,k} = 2\cos^2\frac{k\pi h}{2}\left(-1-M_{2h}\right) = -\left(M_h+1\right)\cos^2\frac{k\pi h}{2} \tag{E.46}$$

En résumé :

$$\tilde{\sigma}_{m,k} = \begin{cases} \left(M_h+1\right)\cos^2\dfrac{k\pi h}{2} & \text{si } 1\leq m=k\leq M_{2h} \\ -\left(M_h+1\right)\cos^2\dfrac{k\pi h}{2} & \text{si } 1\leq m=M_h+1-k\leq M_{2h} \\ 0 & \text{autrement} \end{cases} \tag{E.47}$$

Expression définitive de la matrice Σ :

On est maintenant en mesure de calculer l'élément générique $\Sigma_{j,k}$, ce que l'on fait en examinant séparément les cinq cas suivants successivement :

1. $1\leq j=k\leq M_{2h}$
2. $j=k=M_{2h}+1$
3. $M_{2h}+2\leq j=k\leq M_h$
4. $1\leq j=M_h+1-k\leq M_{2h}$
5. $j\neq k$ et $j+k\neq M_h+1$

Cas 1 : $1\leq j=k\leq M_{2h}$

$$\Sigma_{j,j} = \frac{1}{\left(\lambda_h\right)_j}-4h^2\sum_{m=1}^{M_{2h}}\frac{\tilde{\sigma}_{m,j}^2}{\left(\lambda_h\right)_{2m}} \tag{E.48}$$

La seule contribution non nulle de $\tilde{\sigma}$ à $\Sigma_{j,j}$ correspond à $m = j$ de sorte que :

$$\Sigma_{j,j} = \frac{1}{4\sin^2\,\frac{j\pi h}{2}} - 4h^2\,\frac{(M_h+1)^2\cos^4\,\frac{j\pi h}{2}}{4\sin^2\,(j\pi h)} = \frac{\cos^2\frac{j\pi h}{2} - \cos^4\frac{j\pi h}{2}}{4\sin^2\frac{j\pi h}{2}\,\cos^2\frac{j\pi h}{2}} = \frac{1}{4} \tag{E.49}$$

Cas 2 : $j = k = M_{2h} + 1$

On a alors

$$\tilde{\sigma}_{m,j} = 0\,,\ \forall\, m \leq M_{2h} \tag{E.50}$$

de sorte que $\Sigma_{j,j}$ se réduit à :

$$\Sigma_{j,j} = \frac{1}{(\lambda_h)_j} = \frac{1}{\sin^2\,\frac{j\pi h}{2}} = \frac{1}{2} \tag{E.51}$$

car $jh = (M_{2h}+1)\,/\,(M_h+1) = 1/2$.

Cas 3 : $M_{2h} + 2 \leq j = k \leq M_h$

La seule contribution non nulle de $\tilde{\sigma}$ à $\Sigma_{j,j}$ correspond à $m = M_h + 1 - j \leq M_h + 1 - (M_{2h}+2) = M_{2h}$ de sorte que :

$$\Sigma_{j,j} = \frac{1}{\sin^2\,\frac{j\pi h}{2}} - 4h^2\,\frac{\left[-\,(M_h+1)\,\cos^2\,\frac{j\pi h}{2}\right]^2}{4\sin^2\,(m\pi h)} \tag{E.52}$$

De plus,

$$\sin\,(m\pi h) = \sin\,\{[(M_h+1) - j]\,\pi h\} = \sin\,(\pi - j\pi h) = \sin\,(j\pi h) \tag{E.53}$$

d'où le résultat :

$$\Sigma_{j,j} = \frac{1}{4} \tag{E.54}$$

Cas 4 : $1 \leq j = M_h + 1 - k \leq M_{2h}$

La seule contribution non nulle de $\tilde{\sigma}$ à $\Sigma_{j,k}$ correspond à $m = j = M_h + 1 - k$; de plus $j \neq k$ de sorte que :

$$\begin{aligned}\Sigma_{j,k} &= -4h^2\,\frac{(M_h+1)\,\cos^2\,\frac{j\pi h}{2}\,\left[-\,(M_h+1)\,\cos^2\,\frac{k\pi h}{2}\right]}{4\sin^2\,(j\pi h)} \\ &= \frac{\cos^2\,\frac{j\pi h}{2}\,\cos^2\,\frac{k\pi h}{2}}{\sin^2\,(j\pi h)}\end{aligned} \tag{E.55}$$

De plus d'une part,

$$\cos\,\frac{k\pi h}{2} = \cos\left[(M_h + 1 - j)\,\frac{\pi h}{2}\right] = \cos\left(\frac{\pi}{2} - \frac{j\pi h}{2}\right) = \sin\,\frac{j\pi h}{2} \qquad \text{(E.56)}$$

et d'autre part,

$$\sin\,(j\pi h) = 2\sin\,\frac{j\pi h}{2}\,\cos\,\frac{j\pi h}{2} \qquad \text{(E.57)}$$

de sorte que finalement :

$$\Sigma_{j,k} = \frac{1}{4} \qquad \text{(E.58)}$$

Cas 5 : $j \neq k$ et $j + k \neq M_h + 1$

Alors on a :

$$\forall\, m \leq M_{2h}\ ,\ \tilde{\sigma}_{m,j}\,\tilde{\sigma}_{m,k} = 0 = \delta_{j,k} \qquad \text{(E.59)}$$

de sorte que :

$$\Sigma_{j,k} = 0 \qquad \text{(E.60)}$$

En rassemblant les résultats relatifs aux cas 1-5 et en tenant compte du fait que la matrice Σ est symétrique, on aboutit à la forme définitive particulièrement simple suivante :

$$\Sigma = \begin{pmatrix} \frac{1}{4} & & & & & & & & \frac{1}{4} \\ & \frac{1}{4} & & & & & & \frac{1}{4} & \\ & & \ddots & & & & \iddots & & \\ & & & \frac{1}{4} & & \frac{1}{4} & & & \\ & & & & \frac{1}{2} & & & & \\ & & & \frac{1}{4} & & \frac{1}{4} & & & \\ & & \iddots & & & & \ddots & & \\ & \frac{1}{4} & & & & & & \frac{1}{4} & \\ \frac{1}{4} & & & & & & & & \frac{1}{4} \end{pmatrix} \qquad \text{(E.61)}$$

où les éléments en dehors des deux diagonales sont des zéros.

Diagonalisation et rayon spectral

On rappelle que la matrice d'amplification est semblable à la matrice suivante :

$$G' = \Sigma\, D^2 \qquad \text{(E.62)}$$

Or,

$$G' = \frac{1}{4}\begin{pmatrix} d_1^2 & & & & & & & & d_{M_h}^2 \\ & d_2^2 & & & & & & d_{M_h-1}^2 & \\ & & \ddots & & & & \iddots & & \\ & & & d_{M_{2h}}^2 & & d_{M_{2h}+2}^2 & & & \\ & & & & 2d_{M_{2h}+1}^2 & & & & \\ & & & d_{M_{2h}}^2 & & d_{M_{2h}+2}^2 & & & \\ & & \iddots & & & & \ddots & & \\ & d_2^2 & & & & & & d_{M_h-1}^2 & \\ d_1^2 & & & & & & & & d_{M_h}^2 \end{pmatrix} \tag{E.63}$$

On note $\{e_k\}$ ($k = 1, ..., M_h$) les vecteurs de la base canonique ($e_{j,k} = \delta_{j,k}$).

On peut alors aisément vérifier que les vecteurs

$$x_k = e_k + e_{M_h+1-k} \quad (k = 1, ..., M_{2h} + 1) \tag{E.64}$$

sont des vecteurs propres de la matrice G' respectivement associés aux valeurs propres $\{\mu_k\}$ suivantes :

$$\mu_k = \frac{d_k^2 + d_{M_h+1-k}^2}{4} \quad (k = 1, ..., M_{2h} + 1) \tag{E.65}$$

On vérifie ensuite que les vecteurs

$$y_k = \frac{e_k}{d_k^2} - \frac{e_{M_h+1-k}}{d_{M_h+1-k}^2} \quad (k = 1, ..., M_{2h}) \tag{E.66}$$

appartiennent au noyau de G'.

On vérifie enfin que ces vecteurs propres,

$$\{x_k\}_{k=1,...,M_{2h}+1} \cup \{y_k\}_{k=1,...,M_{2h}} \tag{E.67}$$

forment une famille libre maximale. On en conclut que l'ensemble :

$$\{\mu_k\}_{k=1,...,M_{2h}+1} \cup \{0\} \tag{E.68}$$

constitue la totalité des valeurs propres des matrices G' et donc G, 0 ayant la multiplicité M_{2h}.

Par conséquent, le rayon spectral satisfait les relations suivantes :

$$\boxed{\rho = \max_{k=1,\ldots,M_{2h}+1} \left(\frac{d_k^2 + d_{M_h+1-k}^2}{4} \right) \leq B_0} \tag{E.69}$$

où la borne B_0 est donnée par :

$$B_0 = \max_{\lambda \in [0,2]} \left[\frac{d(\lambda)^2 + d(4-\lambda)^2}{4} \right] \tag{E.70}$$

On pose :

$$\lambda = 2 - \theta \quad (\theta \in [0,2]) \tag{E.71}$$

de sorte que

$$B_0 = \max_{\theta \in [0,2]} g(\theta) \tag{E.72}$$

où

$$g(\theta) = \frac{1}{4T_3(3)^2} \left[(2-\theta) T_3(1+\theta)^2 + (2+\theta) T_3(1-\theta)^2 \right] \tag{E.73}$$

Puisque T_3 est un polynôme du 3^e degré, $g(\theta)$ est un polynôme du 7^e en θ. Mais de plus c'est un polynôme pair ; il ne contient donc que des monômes en 1, θ^2, θ^4 et θ^6. Par conséquent, $g(\theta)$ se réduit à un polynôme du 3^e degré en $t = \theta^2$ ($t \in [0,4]$). D'ailleurs on obtient facilement :

$$g = \frac{1}{99^2} \left(1 + 96t + 104t^2 - 32t^3 \right) \tag{E.74}$$

Le maximum cherché se localise donc en résolvant l'équation du second degré suivante :

$$\frac{dg}{dt} = \frac{1}{99^2} \left(96 + 208t - 96t^2 \right) = 0 \tag{E.75}$$

On obtient les racines $t = (13 \pm \sqrt{313})/12$ dont la seule positive correspond au signe +, ce qui donne, par substitution dans g, la valeur cherchée de la borne :

$$\boxed{B_0 = \frac{5032 + 313\sqrt{313}}{264627} \approx 0.03994125800} \tag{E.76}$$

Annexe F

Corrigés des exercices

Exercice 1.1 (Erreur d'approximation du modèle discret fondamental)

(1) Le problème continu (1.8) se résout par deux quadratures ; son analogue discret (1.10) par deux sommations. Pour tout $\ell = 2, 3, ..., M+1$:

$$\begin{aligned} h^2 \sum_{k=1}^{\ell-1} f_k &= (2u_1 - u_2) + (-u_1 + 2u_2 - u_3) + (-u_2 + 2u_3 - u_4) \\ &\quad + ... + (-u_{\ell-2} + 2u_{\ell-1} - u_\ell) \\ &= u_1 + u_{\ell-1} - u_\ell \end{aligned} \tag{F.1}$$

Puis, pour $j = 2, 3, ..., M+1$:

$$\begin{aligned} h^2 \sum_{\ell=2}^{j} \sum_{k=1}^{\ell-1} f_k &= \sum_{\ell=2}^{j} (u_1 + u_{\ell-1} - u_\ell) \\ &= (j-1)\, u_1 + (u_1 - u_2) + (u_2 - u_3) + ... + (u_{j-1} - u_j) \\ &= j\, u_1 - u_j \end{aligned} \tag{F.2}$$

D'où :

$$\begin{aligned} u_j &= j\, u_1 - h^2 \sum_{\ell=2}^{j} \sum_{k=1}^{\ell-1} f_k \\ &= j\, u_1 - h^2 \sum_{k=1}^{j-1} (j-k)\, f_k \\ &= j\, u_1 - h^2 \Big((j-1)\, f_1 + (j-2)\, f_2 + ... + f_{j-1} \Big) \end{aligned} \tag{F.3}$$

En particulier,

$$0 = u_{M+1} = (M+1)\,u_1 - h^2\left(M\,f_1 + (M-1)\,f_2 + ... + f_M\right) \tag{F.4}$$

dont on tire

$$u_1 = \frac{h^2}{M+1}\left(M\,f_1 + (M-1)\,f_2 + ... + f_M\right) \tag{F.5}$$

et enfin :

$$\begin{aligned} u_j &= \frac{jh^2}{M+1}\left(M\,f_1 + (M-1)\,f_2 + ... + f_M\right) \\ &\quad - h^2\left((j-1)\,f_1 + (j-2)\,f_2 + ... + f_{j-1}\right) \\ &= \frac{M+1-j}{M+1}\,h^2\left(f_1 + 2f_2 + ... + (j-1)f_{j-1}\right) \\ &\quad + \frac{j\,h^2}{M+1}\left((M+1-j)f_j + (M-j)f_{j+1} + ... + f_M\right) \end{aligned} \tag{F.6}$$

Cette équation est la forme développée de la ligne j de la relation vectorielle suivante :

$$u_h = A_h^{-1}\,f_h \tag{F.7}$$

Elle permet donc d'identifier les coefficients de la ligne j de la matrice A_h^{-1} comme étant les suivants :

$$\begin{gathered} \frac{M+1-j}{M+1}\,h^2,\ 2\,\frac{M+1-j}{M+1}\,h^2,\ ...,\ (j-1)\,\frac{M+1-j}{M+1}\,h^2, \\ (M+1-j)\,\frac{j\,h^2}{M+1},\ (M-j)\,\frac{j\,h^2}{M+1},\ ...,\ \frac{j\,h^2}{M+1} \end{gathered} \tag{F.8}$$

Ces coefficients sont positifs et admettent la somme suivante :

$$\begin{aligned} s_j &= \frac{M+1-j}{M+1}\,h^2\,\frac{(j-1)j}{2} + \frac{j\,h^2}{M+1}\,\frac{(M+1-j)(M+2-j)}{2} \\ &= \frac{h^2}{2}\,j(M+1-j) \end{aligned} \tag{F.9}$$

Par conséquent :

$$\|\,A_h^{-1}\,\|_\infty = \max_j\,s_j = \frac{h^2}{2}\,\max_j\Big(j(M+1-j)\Big) \le \frac{1}{8} \tag{F.10}$$

l'égalité ayant lieu si l'entier M est impair.

(2) L'erreur de troncature au nœud x_j est définie comme suit :

$$E_T(x_j) = \frac{-u(x_{j-1}) + 2u(x_j) - u(x_{j+1})}{h^2} - f(x_j) \tag{F.11}$$

dans laquelle la fonction $u(x)$ est la solution exacte du problème continu (1.8). Or, on a les développements de Taylor suivants :

$$\begin{aligned} u(x_{j+1}) &= u(x_j + h) \\ &= u(x_j) + h\,u'(x_j) + \frac{h^2}{2}\,u''(x_j) + \frac{h^3}{6}\,u'''(x_j) + \frac{h^4}{24}\,u''''(\xi_j) \qquad \text{(F.12)} \\ u(x_{j-1}) &= u(x_j - h) \\ &= u(x_j) - h\,u'(x_j) + \frac{h^2}{2}\,u''(x_j) - \frac{h^3}{6}\,u'''(x_j) + \frac{h^4}{24}\,u''''(\eta_j) \qquad \text{(F.13)} \end{aligned}$$

où $\xi_j \in]x_j, x_{j+1}[$ et $\eta_j \in]x_{j-1}, x_j[$. Par conséquent :

$$E_T(x_j) = -u''(x_j) - f(x_j) + \Big(u''''(\xi_j) + u''''(\eta_j)\Big)\,\frac{h^4}{24} \tag{F.14}$$

En outre, comme $u(x)$ est la solution du problème continu,

$$-u''(x) - f(x) = 0 \text{ et } u''''(x) = f''(x) \quad (\forall\, x) \tag{F.15}$$

de sorte que :

$$E_T(x_j) = \Big(f''(\xi_j) + f''(\eta_j)\Big)\,\frac{h^2}{24} \tag{F.16}$$

Enfin, f'' étant par hypothèse une fonction continue, le théorème de la moyenne permet de simplifier cette expression comme suit

$$E_T(x_j) = f''(c_j)\,\frac{h^2}{12} \tag{F.17}$$

où $c_j \in]x_{j-1}, x_{j+1}[$ et d'en déduire la majoration cherchée :

$$\| E_T \|_\infty \leq \frac{\mu h^2}{12} \tag{F.18}$$

(3) La relation définissant l'erreur de troncature

$$E_T = A_h\,u - f_h = A_h(u - u_h) \quad \text{(en linéaire)} \tag{F.19}$$

s'inverse comme suit :

$$u - u_h = A_h^{-1}\,E_T \tag{F.20}$$

et compte tenu des majorations établies aux questions précédentes, il vient :

$$\begin{aligned} \| u_h - u \|_\infty &= \| A_h^{-1} E_T \|_\infty \\ &\leq \| A_h^{-1} \|_\infty \| E_T \|_\infty \\ &\leq \frac{\mu h^2}{96} \end{aligned} \tag{F.21}$$

Exercice 1.2 (Paramètres optimaux de l'itération de Jacobi)

(1) Chaque itération de Jacobi a pour effet de réduire le mode propre d'indice m d'un facteur égal à $g_m(\tau)$ avec :

$$g_m(\tau) = 1 - \lambda_{h\,m}\,\tau \tag{F.22}$$

Dans le cas présent, les valeurs propres $\{\lambda_{h\,m}\}$ sont des réels positifs ordonnés par valeur croissante. Dans un plan (τ, η), les courbes d'équations $\eta = g_m(\tau)$ forment un faisceau de demi-droites de pentes négatives issues du point $(0, 1)$ (voir figure F.1). Lorsque τ est petit, ces facteurs sont tous positifs et décroissent lorsque τ croît. Le rayon spectral,

$$\rho(\tau) = \max_{m=1,2,\ldots,M} | g_m(\tau) | \tag{F.23}$$

est alors donné par :

$$\rho(\tau) = g_1(\tau) \quad (\tau \text{ petit}) \tag{F.24}$$

Pour une certaine valeur τ^* de τ, on a :

$$g_M(\tau^*) = -g_1(\tau^*) \tag{F.25}$$

Pour $\tau > \tau^*$, on a :

$$| g_M(\tau) | > | g_1(\tau) | \tag{F.26}$$

et par conséquent :

$$\rho(\tau) = | g_M(\tau) | = -g_M(\tau) \quad (\tau > \tau^*) \tag{F.27}$$

Enfin, il existe une certaine valeur $\tau_{\max} > \tau^*$ telle que :

$$g_M(\tau_{\max}) = -1 \tag{F.28}$$

de sorte que

$$\rho(\tau_{\max}) = 1 \tag{F.29}$$

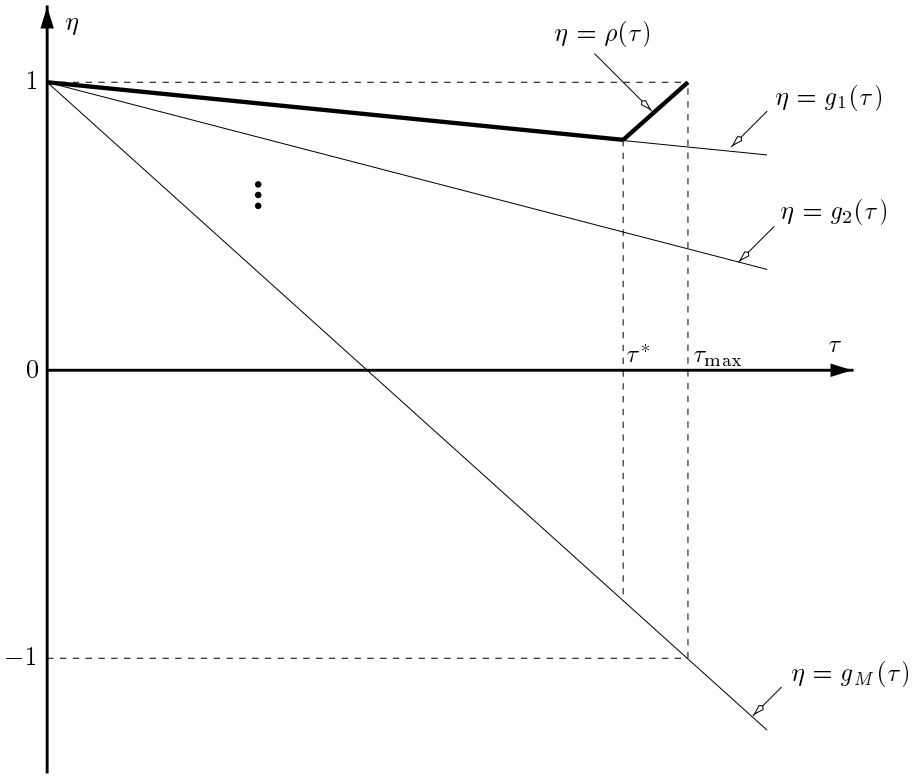


Figure F.1. Facteurs d'atténuation $g_m(\tau)$ (cf. exercice 1.2)

et au-delà de laquelle ($\tau > \tau_{\max}$) l'itération diverge ($\rho(\tau) > 1$).

Il apparaît donc que le rayon spectral $\rho(\tau)$ est minimum pour

$$\tau = \tau^* \tag{F.30}$$

qui est la valeur optimale. En résolvant (F.25), il vient :

$$\tau^* = \left(\frac{\lambda_{h\,1} + \lambda_{h\,M}}{2} \right)^{-1} \tag{F.31}$$

On note que s'il existe un mode d'indice ℓ pour lequel on a exactement

$$\lambda_\ell = \left(\frac{\lambda_{h\,1} + \lambda_{h\,M}}{2} \right) \tag{F.32}$$

et si $\tau = \tau^*$, ce mode est éliminé par une seule application de l'itération de Jacobi. On dit qu'il y a « annihiliation » du mode (voir chapitre 3). Par conséquent, on pourra retenir le résultat précédent comme suit : dans la méthode de Jacobi, la valeur optimale τ^* du paramètre de relaxation τ est celle qui annihile le mode associé à une valeur propre fictive égale à la moyenne arithmétique des bornes de λ, c'est-à-dire l'inverse de cette moyenne. Le rayon spectral correspondant est égal à la quantité suivante :

$$\begin{aligned}\rho^* &= \rho(\tau^*) \\ &= g_1(\tau^*) \\ &= 1 - \lambda_{h\,1} \left(\frac{\lambda_{h\,1} + \lambda_{h\,M}}{2} \right)^{-1} \\ &= \frac{\lambda_{h\,M} - \lambda_{h\,1}}{\lambda_{h\,M} + \lambda_{h\,1}} \\ &= \frac{\kappa - 1}{\kappa + 1}\end{aligned} \tag{F.33}$$

où l'on a posé :

$$\kappa \stackrel{\text{déf}}{=} \frac{\lambda_{h\,M}}{\lambda_{h\,1}} \tag{F.34}$$

(« nombre de conditionnement »). On observe que lorsque M est grand, l'optimum τ^* est peu différent du maximum $\tau_{\max}$ pour lequel l'itération ne converge pas ($\rho = 1$) ; de manière équivalente, le nombre de conditionnement κ est très grand, le rayon spectral ρ^* est très proche de 1, l'itération de Jacobi converge très lentement. On dit que le système est « raide ».

Exercice 1.3 (Propriétés des nombres de conditionnement)

(1) En vertu du Théorème A.4, on a :

$$\| A \| \geq \rho(A) \text{ et } \| A^{-1} \| \geq \rho(A^{-1}) \tag{F.35}$$

où ρ désigne le rayon spectral. Si $\{\lambda_1, \lambda_2, ..., \lambda_M\}$ sont les valeurs propres de la matrice A ordonnées par valeur croissante du module, celles de la matrice A^{-1} sont les nombres $\{1/\lambda_M, 1/\lambda_{M-1}, ..., 1/\lambda_1\}$. Par conséquent :

$$\rho(A) = |\,\lambda_M\,| \text{ et } \rho(A^{-1}) = \frac{1}{|\,\lambda_1\,|} \tag{F.36}$$

de sorte que :

$$\kappa(A) \geq \rho(A)\,\rho(A^{-1}) = \frac{|\,\lambda_M\,|}{|\,\lambda_1\,|} \geq 1 \tag{F.37}$$

Dans le cas d'une matrice scalaire inversible, c'est-à-dire d'un multiple non nul de la matrice identité, $A = \lambda\, I\ (\lambda \neq 0)$, l'égalité est toujours vraie ; en effet dans ce cas, quelle que soit la norme induite choisie, on a :

$$\| A \| = | \lambda | \text{ et } \| A^{-1} \| = \frac{1}{| \lambda |} \tag{F.38}$$

et par conséquent,

$$\kappa(\lambda\, I) = 1 \quad (\lambda \neq 0) \tag{F.39}$$

(2) En vertu de (A.67), on a :

$$\| A \|_2 = \sqrt{\rho(A^*\, A)} \tag{F.40}$$

et

$$\| A^{-1} \|_2 = \sqrt{\rho\Big((A^{-1})^*\, A^{-1}\Big)} = \sqrt{\rho\Big((A\, A^*)^{-1}\Big)} \tag{F.41}$$

Les valeurs propres de la matrice $A^*\, A$ sont des réels positifs $\mu_1, \mu_2, ..., \mu_M$ ordonnés par valeur croissante, de sorte que :

$$\| A \|_2 = \sqrt{\mu_M} \tag{F.42}$$

En outre, la matrice A étant inversible, la relation

$$A\, A^* = A\ (A^*\, A)\, A^{-1} \tag{F.43}$$

prouve que les matrices $A\, A^*$ et $A^*\, A$ sont semblables, ce qui d'ailleurs serait encore vrai si la matrice A était singulière. Les valeurs propres de la matrice $A\, A^*$ sont donc également les nombres $\{\mu_m\}$, et celles de la matrice inverse $1/\mu_M, 1/\mu_{M-1}, ..., 1/\mu_1$. Par conséquent :

$$\| A^{-1} \|_2 = \frac{1}{\sqrt{\mu_1}} \tag{F.44}$$

et finalement

$$\kappa_2(A) = \sqrt{\frac{\mu_M}{\mu_1}} = \sqrt{\frac{\lambda_{\max}(A^*\, A)}{\lambda_{\min}(A^*\, A)}} \tag{F.45}$$

En conséquence, les matrices inversibles A pour lesquelles

$$\kappa_2(A) = 1 \tag{F.46}$$

sont celles qui sont telles que toutes les valeurs propres de la matrice $A^* A$ sont égales à un certain réel $r > 0$. Or cette matrice est hermitienne donc diagonalisable, donc nécessairement scalaire :

$$A^* A = r I \tag{F.47}$$

La matrice $A/\sqrt{r}$ est donc unitaire. La réciproque est immédiate. Il s'agit donc d'une caractérisation.

Si la matrice A est réelle symétrique définie positive, ses valeurs propres sont réelles et positives, et

$$\lambda_{\max}(A^* A) = \lambda_{\max}(A^2) = \lambda^2_{\max}(A) \tag{F.48}$$

et de même pour le minimum. L'expression de $\kappa_2(A)$ se réduit alors à la suivante :

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \tag{F.49}$$

(3) Les valeurs propres de la matrice A_h du modèle discret unidimensionnel (1.10) sont fournies par le Théorème 1.1 :

$$\lambda_{h\,m} = \frac{2 - 2\,\cos\,\theta_m}{h^2} = \frac{4}{h^2}\,\sin^2\,\frac{\theta_m}{2} \tag{F.50}$$

Cette matrice étant réelle symétrique définie positive, le résultat de la question précédente lui est applicable, de sorte que :

$$\kappa_2(A_h) = \frac{\lambda_{\max}(A_h)}{\lambda_{\min}(A_h)} = \frac{\sin^2\,\dfrac{\theta_M}{2}}{\sin^2\,\dfrac{\theta_1}{2}} \tag{F.51}$$

Or,

$$\theta_1 = \frac{\pi}{M+1} \text{ et } \theta_M = \pi - \theta_1 \tag{F.52}$$

de sorte que

$$\kappa_2(A_h) = \frac{1}{\tan^2\,\dfrac{\pi}{M+1}} \sim \frac{(M+1)^2}{\pi^2} \quad (M \to \infty) \tag{F.53}$$

Par ailleurs, par inspection directe de la matrice A_h, il vient

$$\| A_h \|_\infty = \frac{4}{h^2} \tag{F.54}$$

et en vertu d'un résultat de l'exercice 1.1 :

$$\| A_h^{-1} \|_\infty = \frac{1}{8} \tag{F.55}$$

(pour M impair), de sorte que :

$$\kappa_\infty(A_h) = \frac{1}{2h^2} = \frac{(M+1)^2}{2} \tag{F.56}$$

Exercice 1.4 (Encadrement de résidus)

Par leur définition dans le cadre d'un problème linéaire inversible, les vecteurs « résidu » et « erreur » sont liés par les relations :

$$r_h^n = A_h\, e_h^n \text{ et } r_h^0 = A_h\, e_h^0 \tag{F.57}$$

Par inversion, on a donc aussi :

$$e_h^n = A_h^{-1}\, r_h^n \text{ et } e_h^0 = A_h^{-1}\, r_h^0 \tag{F.58}$$

D'où les majorations :

$$\| r_h^n \| \leq \| A_h \| \, \| e_h^n \| \tag{F.59}$$

$$\| r_h^0 \| \leq \| A_h \| \, \| e_h^0 \| \tag{F.60}$$

d'une part, et :

$$\| e_h^n \| \leq \| A_h^{-1} \| \, \| r_h^n \| \tag{F.61}$$

$$\| e_h^0 \| \leq \| A_h^{-1} \| \, \| r_h^0 \| \tag{F.62}$$

d'autre part. Les équations (F.60) et (F.62) se réécrivent comme suit :

$$\frac{1}{\| e_h^0 \|} \leq \| A_h \| \, \frac{1}{\| r_h^0 \|} \tag{F.63}$$

$$\frac{1}{\| r_h^0 \|} \leq \| A_h^{-1} \| \, \frac{1}{\| e_h^0 \|} \tag{F.64}$$

En combinant (F.59) et (F.64), il vient :

$$\frac{\| r_h^n \|}{\| r_h^0 \|} \leq \kappa \, \frac{\| e_h^n \|}{\| e_h^0 \|} \tag{F.65}$$

En combinant (F.61) et (F.63), il vient :

$$\frac{\| e_h^n \|}{\| e_h^0 \|} \leq \kappa \, \frac{\| r_h^n \|}{\| r_h^0 \|} \tag{F.66}$$

ce qui complète le résultat.

La réduction de l'erreur itérative constitue le véritable objectif de l'itération, mais en pratique, le résidu est la seule quantité mesurable. Ces relations établissent le lien entre les réductions relatives de leurs normes respectives, permettant un contrôle rigoureux de l'erreur, dès lors que l'on dispose d'une estimation du nombre de conditionnement κ.

Exercice 1.5 (Estimation de convergence)

(1) En vertu d'un résultat de l'exercice 1.1, la condition portant sur h s'écrit

$$\| u_h - u \| \leq \frac{\mu h^2}{96} \leq 10^{-4} \tag{F.67}$$

où ici

$$\mu = \max_{x \in [0,1]} | f''(x) | = 8 \tag{F.68}$$

Ceci donne :

$$h \leq \sqrt{12}\, 10^{-2} \approx 0.0346 \tag{F.69}$$

ou, de manière équivalente :

$$M \geq 29 \tag{F.70}$$

On fixe donc le nombre d'intervalles de discrétisation à 30 ($M = 29$).

(2) D'après un résultat de l'exercice 1.3,

$$\kappa_\infty(A_h) = \frac{(M+1)^2}{2} = 450 \tag{F.71}$$

(3) Le nombre n d'itérations de Jacobi suffisant à une réduction relative de l'erreur itérative de 10^{-4} est lié au rayon spectral ρ par la condition :

$$\rho^n \leq 10^{-4} \tag{F.72}$$

Ici,

$$\kappa = \frac{\lambda_{h\,29}}{\lambda_{h\,1}} = \frac{\sin^2 \dfrac{29\,\pi}{60}}{\sin^2 \dfrac{\pi}{60}} \approx 364.1 \tag{F.73}$$

d'où le rayon spectral

$$\rho = \frac{\kappa - 1}{\kappa + 1} \approx 0.9945 \tag{F.74}$$

et la condition sur n :

$$n \geq \frac{4}{\log_{10} 1/\rho} \tag{F.75}$$

soit $n \geq 1677$.

(4) Conformément au cours, pour garantir que les erreurs d'arrondi n'altèrent pas la précision du calcul, la condition suivante doit être remplie :

$$\varepsilon \times \kappa \leq 10^{-4} \tag{F.76}$$

ce qui exige ici que $\varepsilon \leq 2.7\ 10^{-7}$. On peut donc estimer qu'il convient de retenir 7 décimales aux représentations en réel flottant pour pouvoir garantir cette précision.

(5) En vertu de l'exercice 1.4,

$$\frac{\| e_h^n \|}{\| e_h^0 \|} \leq \kappa \frac{\| r_h^n \|}{\| r_h^0 \|} \tag{F.77}$$

Par conséquent, il convient d'imposer la condition suivante :

$$\frac{\| r_h^n \|}{\| r_h^0 \|} \leq \frac{1}{\kappa} 10^{-4} \approx 2.7\ 10^{-7} \tag{F.78}$$

pour satisfaire au critère.

Exercice 2.1 (Formes hermitiennes)

Pour la vérification des axiomes voir les définitions 2.1 et 2.2. Dans le cas 1, la quantité

$$(u, u) = \sum_{j=1}^{M} |u_j|^2 = \Big(\|u\|_2 \Big)^2 \tag{F.79}$$

qui est égale au carré de la norme euclidienne, est positive ou nulle quel que soit le vecteur $u \in \mathbb{R}^M$, et nulle seulement si $u = 0$. De même, dans le cas 2, la quantité

$$(u, u) = \int_a^b |u(x)|^2 \, dx \tag{F.80}$$

est positive ou nulle quelle que soit la fonction $u \in L^2\Big([a, b]\Big)$, et nulle seulement si $u = 0$ (c'est-à-dire $u(x) = 0$ presque partout).

Exercice 2.2 (Opérateur de dérivée première)

Soient u et v deux éléments quelconques de l'espace de Hilbert

$$H = \Big\{ u \in L^2\Big([a,b] \to \mathbb{C}\Big) \ / \ u(a) = u(b) \Big\} \tag{F.81}$$

Une intégration par parties donne :

$$\begin{aligned}
(Au, v) &= \int_a^b \frac{du(x)}{dx}\, \overline{v(x)}\, dx \\
&= \Big[u(x)\, \overline{v(x)}\Big]_a^b - \int_a^b u(x)\, \overline{\frac{dv(x)}{dx}}\, dx \\
&= -\int_a^b u(x)\, \overline{\frac{dv(x)}{dx}}\, dx \\
&= (u, -Av)
\end{aligned} \tag{F.82}$$

Par conséquent l'opérateur A est antisymétrique :

$$A^* = -A \tag{F.83}$$

et on en déduit que ses valeurs propres sont purement imaginaires. Plus précisément, les fonctions propres sont les fonctions $\phi \in H$ telles que :

$$A\,\phi = \lambda\,\phi \tag{F.84}$$

pour un certain $\lambda \in \mathbb{C}$, ce qui équivaut à l'équation différentielle suivante :

$$\frac{d\phi}{dx} = \lambda\,\phi \tag{F.85}$$

($\phi \in L^2\Big([a,b]\Big)$) soumise à la condition aux limites

$$\phi(a) = \phi(b) \tag{F.86}$$

Il vient :

$$\phi(x) = C\, \exp\Big(\lambda\,(x-a)\Big) \tag{F.87}$$

dans laquelle C est une constante arbitraire. Les seules solutions non triviales ($C \neq 0$) qui satisfont la condition aux limites sont les fonctions périodiques obtenues lorsque λ est de la forme suivante :

$$\lambda = \lambda_m = \frac{2\pi i m}{b-a} \quad (m \in \mathbb{Z}) \tag{F.88}$$

ce qui fournit, comme prévu, un ensemble dénombrable de modes propres pour lesquels les fonctions propres ont la forme suivante :

$$\phi_m(x) = \exp\Big(2\pi i m \frac{x-a}{b-a}\Big) \tag{F.89}$$

Exercice 2.3 (Equation d'advection pure)

On décompose la condition initiale donnée en série de Fourier :

$$u_0(x) = \sum_{m=-\infty}^{+\infty} c_m^0\, \phi_m(x) \tag{F.90}$$

où ici,

$$\phi_m(x) = \exp\Big(\frac{2\pi\, i\, m\, x}{L}\Big) \tag{F.91}$$

de sorte que pour tout entier $\ell \in \mathbb{Z}$:

$$\begin{aligned}(u_0, \phi_\ell) &= \sum_{m=-\infty}^{+\infty} c_m^0 \int_0^L \exp\Big(\frac{2\pi\, i\, m\, x}{L}\Big)\, \exp\Big(-\frac{2\pi\, i\, \ell\, x}{L}\Big)\, dx \\ &= L\, c_\ell^0 \end{aligned} \tag{F.92}$$

de sorte que :

$$\forall\, m \in \mathbb{Z}\,,\, c_m^0 = \frac{1}{L} \int_0^L u_0(x)\, \exp\Big(-\frac{2\pi\, i\, m\, x}{L}\Big)\, dx \tag{F.93}$$

Notons que dans le cas courant où la condition initiale $u_0(x)$ est réelle, ces coefficients satisfont la condition suivante :

$$c_{-m} = \overline{c_m} \quad (\forall\, m \in \mathbb{Z}) \tag{F.94}$$

de sorte qu'en introduisant les coefficients réels suivants :

$$a_m = c_m + c_{-m} = 2\Re(c_m) \text{ et } b_m = \frac{c_m - c_{-m}}{i} = 2\Im(c_m) \tag{F.95}$$

la série précédente prend la forme habituelle suivante :

$$u_0(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} \Big(a_m\, \cos\, 2\pi m x + b_m\, \sin\, 2\pi m x\Big) \tag{F.96}$$

Par conséquent, on peut chercher la solution du problème sous la forme de la série de Fourier suivante

$$u(x,t) = \sum_{m=-\infty}^{+\infty} c_m(t)\, \phi_m(x) \tag{F.97}$$

dont les coefficients $\{c_m(t)\}$ dépendent du temps et vérifient la condition :

$$c_m(0) = c_m^0 \tag{F.98}$$

équivalente à la condition initiale. On observe que toute fonction représentée par une telle série satisfait automatiquement la condition aux limites. En reportant cette série dans l'EDP, on voit que celle-ci équivaut au système suivant d'EDO :

$$c_m(t)' + c\,\lambda_m\,c_m(t) = 0 \quad (\forall\, m) \tag{F.99}$$

D'où :

$$c_m(t) = c_m^0\ \exp(-\lambda_m ct) \tag{F.100}$$

et finalement :

$$\begin{aligned} u(x,t) &= \sum_{m=-\infty}^{+\infty} c_m^0\ \exp(-\lambda_m ct)\ \exp(\lambda_m x) \\ &= \sum_{m=-\infty}^{+\infty} c_m^0\ \exp\Big(\lambda_m(x-ct)\Big) \\ &= u_0(x-ct) \end{aligned} \tag{F.101}$$

En conclusion, la solution du problème d'advection pure est une série de Fourier dont les coefficients sont obtenus pour $t = 0$ à partir de la condition initiale. L'intégration en temps a pour effet de multiplier le mode de Fourier d'indice m par une exponentielle du temps dont l'exposant est proportionnel à la valeur propre λ_m associée à ce mode et à la célérité c. Comme λ_m est purement imaginaire, ce facteur exponentiel introduit un déphasage proportionnel au temps, l'amplitude étant conservée.

Bien évidemment, l'équation (F.101) peut s'établir directement et simplement.

Exercice 2.4 (Opérateur dérivée seconde)

Soient u et v deux éléments quelconques de l'espace de Hilbert H_0^1. Deux intégrations par parties donnent :

$$\begin{aligned} (Au, v) &= \int_a^b \frac{d^2u(x)}{dx^2}\, v(x)\, dx \\ &= \left[\frac{du(x)}{dx}\, v(x)\right]_a^b - \int_a^b \frac{du(x)}{dx}\, \frac{dv(x)}{dx}\, dx \\ &= -\left[u(x)\, \frac{dv(x)}{dx}\right]_a^b + \int_a^b u(x)\, \frac{d^2v(x)}{dx^2}\, dx \\ &= (u, Av) \end{aligned} \tag{F.102}$$

Par conséquent l'opérateur A est auto-adjoint :

$$A^* = A \tag{F.103}$$

et on en déduit que ses valeurs propres sont réelles. Plus précisément, les fonctions propres sont les fonctions $\psi \in H_0^1$ telles que :

$$A\,\psi = \mu\,\psi \tag{F.104}$$

pour un certain $\mu \in \mathbb{R}$, ce qui équivaut à l'équation différentielle suivante :

$$\frac{d^2\psi}{dx^2} = \mu\,\psi \tag{F.105}$$

soumise aux conditions aux limites suivantes :

$$\psi(a) = \psi(b) = 0 \tag{F.106}$$

Il vient :

$$\psi(x) = \alpha\,\exp\Big(\sqrt{\mu}(x-a)\Big) + \beta\,\exp\Big(-\sqrt{\mu}(x-a)\Big) \tag{F.107}$$

où α et β sont des constantes que l'on détermine en imposant les conditions aux limites :

$$\left\{\begin{array}{l} \alpha + \beta = 0 \\ \exp\Big(\sqrt{\mu}(b-a)\Big)\,\alpha + \exp\Big(-\sqrt{\mu}(b-a)\Big)\,\beta = 0 \end{array}\right. \tag{F.108}$$

Ce système admet une solution non triviale, $(\alpha,\beta) \neq (0,0)$, ssi il est singulier, ce qui équivaut à la condition suivante :

$$\exp\Big(\sqrt{\mu}(b-a)\Big) = \exp\Big(-\sqrt{\mu}(b-a)\Big) \tag{F.109}$$

soit :

$$\exp\Big(2\sqrt{\mu}(b-a)\Big) = 1 \tag{F.110}$$

soit encore :

$$\pm 2\sqrt{\mu}(b-a) = 2m\pi i \quad (m \in \mathbb{N}) \tag{F.111}$$

ce qui fournit, comme prévu, un ensemble dénombrable de modes propres correspondant aux valeurs propres réelles négatives suivantes :

$$\mu_m = -\Big(\frac{m\pi}{b-a}\Big)^2 \tag{F.112}$$

associées aux fonctions propres

$$\begin{aligned}\psi_m(x) &= \frac{1}{2i}\left[\exp\Big(\frac{m\pi i(x-a)}{b-a}\Big) - \exp\Big(-\frac{m\pi i(x-a)}{b-a}\Big)\right] \\ &= \sin\left(m\pi\frac{x-a}{b-a}\right)\end{aligned} \tag{F.113}$$

obtenues en fixant la constante arbitraire α à $1/(2i)$.

Exercice 2.5 (Equation de la chaleur)

On cherche la solution sous la forme d'une série des fonctions propres ; on pose donc ici :

$$\begin{aligned}u(x,t) &= \sum_{m=1}^{\infty} c_m(t)\,\psi_m(x) \\ &= \sum_{m=1}^{\infty} c_m(t)\,\sin\,\frac{m\pi x}{L}\end{aligned} \tag{F.114}$$

($a = 0\,;\ b = L$) en observant qu'une telle fonction satisfait automatiquement les conditions aux limites. Pour déterminer les valeurs initiales des coefficients inconnus $\{c_m(t)\}$, on décompose la fonction donnée $u_0(x)$ en une telle série :

$$\begin{aligned}u_0(x) &= \sum_{m=1}^{\infty} c_m(0)\,\psi_m(x) \\ &= \sum_{m=1}^{\infty} c_m(0)\,\sin\,\frac{m\pi x}{L}\end{aligned} \tag{F.115}$$

Pour inverser cette relation, on utilise à nouveau l'orthogonalité des fonctions de base : pour tout couple d'indices (m,ℓ) distincts ($m \neq \ell$), on a :

$$\begin{aligned}(\psi_m,\psi_\ell) &= \int_0^L \sin\,\frac{m\pi x}{L}\,\sin\,\frac{\ell\pi x}{L}\,dx \\ &= \tfrac{1}{2}\int_0^L \left[\cos\,\frac{(m-\ell)\pi x}{L} - \cos\,\frac{(m+\ell)\pi x}{L}\right]\,dx \\ &= \frac{L}{2\pi}\left[\frac{1}{m-\ell}\,\sin\,\frac{(m-\ell)\pi x}{L} - \frac{1}{m+\ell}\,\sin\,\frac{(m+\ell)\pi x}{L}\right] \\ &= 0\end{aligned} \tag{F.116}$$

Par conséquent :

$$\begin{aligned} c_m(0) &= \frac{(u_0, \psi_m)}{(\psi_m, \psi_m)} \\ &= \frac{2}{L} \int_0^L u_0(x) \, \sin \, \frac{m\pi x}{L} \, dx \end{aligned} \tag{F.117}$$

En reportant dans l'EDP la série exprimant l'inconnue $u(x,t)$, il vient :

$$\sum_{m=1}^{\infty} c'_m(t) \, \sin \, \frac{m\pi x}{L} = -\sigma \sum_{m=1}^{\infty} \left(\frac{m\pi}{L}\right)^2 c_m(t) \, \sin \, \frac{m\pi x}{L} \tag{F.118}$$

Puisque les fonctions propres $\{\psi_m(x)\}$ forment une base, cette équation est satisfaite ssi les coefficients de droite et de gauche sont identiques terme-à-terme, c'est-à-dire ssi les coefficients inconnus $\{c_m(t)\}$ sont solutions du système suivant d'EDO :

$$c'_m(t) = -\sigma \left(\frac{m\pi}{L}\right)^2 c_m(t) \tag{F.119}$$

Il vient :

$$c_m(t) = c_m(0) \, \exp \left[-\sigma t \left(\frac{m\pi}{L}\right)^2\right] \tag{F.120}$$

ce qui complète la détermination de la solution.

En conclusion, pour le problème de la chaleur, la solution s'exprime comme une série de fonctions spatialement sinusoïdales. L'intégration en temps a un effet dissipatif sur chaque terme de la série qui est atténué par une exponentielle du temps dont l'exposant négatif est proportionnel au coefficient de dissipation σ et au carré de la fréquence $m\pi/L$. L'expression de la solution permettrait également de mettre en évidence l'« effet régularisant » de l'intégration en temps : si $u_0(x)$ est une fonction régulière à l'exception d'un ou plusieurs points où la fonction admet un saut, alors quel que soit $t > 0$, $u(x,t)$ est une fonction régulière de x ; les sauts ont été « lissés ».

Exercice 2.6 (Opérateurs de différences finies périodiques usuels)

$$u_j = (u_h)_j \; ; \; v_j = (A_h \, u_h)_j = \sum_{k=-K}^{K} \alpha_k \, u_{j+k}$$

A_h	v_j	...	α_{-2}	α_{-1}	α_0	α_1	α_2	...	K
∇_h	$u_j - u_{j-1}$		0	-1	1	0	0		1
∇_h^2	$u_j - 2u_{j-1} + u_{j-2}$		1	-2	1	0	0		2
Δ_h	$u_{j+1} - u_j$		0	0	-1	1	0		1
Δ_h^2	$u_{j+2} - 2_{j+1} + u_j$		0	0	1	-2	1		2
δ_h	$\frac{1}{2}(u_{j+1} - u_{j-1})$		0	$-\frac{1}{2}$	0	$\frac{1}{2}$	0		1
δ_h^-	$\frac{1}{2}(3u_j - 4u_{j-1} + u_{j-2})$		$\frac{1}{2}$	-2	$\frac{3}{2}$	0	0		2
δ_h^+	$\frac{1}{2}(-3u_j + 4u_{j+1} - u_{j+2})$		0	0	$-\frac{3}{2}$	2	$-\frac{1}{2}$		2

Exercice 2.7 (Structure matricielle d'opérateur périodique)

(1) La relation

$$v_j = (A_h\, u_h)_j = \sum_{k=-K}^{K} \alpha_k\, u_{j+k} \tag{F.121}$$

montre qu'à la ligne j de la matrice A_h le coefficient α_k apparaît à la colonne $j+k$ pour autant que cet indice est compris entre 1 et M. Dans le cas inverse, l'hypothèse de périodicité de la solution permet d'écrire :

$$u_{j+k} = u_{j+k\pm M} \tag{F.122}$$

où le signe est choisi afin que l'indice soit positif ; le coefficient apparaît alors dans la colonne $j + k \pm M$. Par exemple, dans le cas $M = 7$ et $K = 2$, la première ligne s'obtient à partir des relations suivantes :

$$\begin{aligned} v_1 &= \alpha_{-2}\, u_{-1} + \alpha_{-1}\, u_0 + \alpha_0\, u_1 + \alpha_1\, u_2 + \alpha_2\, u_3 \\ &= \alpha_{-2}\, u_6 + \alpha_{-1}\, u_7 + \alpha_0\, u_1 + \alpha_1\, u_2 + \alpha_2\, u_3 \end{aligned} \tag{F.123}$$

d'où la structure matricielle suivante :

$$A_h = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & 0 & 0 & \alpha_{-2} & \alpha_{-1} \\ \alpha_{-1} & \alpha_0 & \alpha_1 & \alpha_2 & 0 & 0 & \alpha_{-2} \\ \alpha_{-2} & \alpha_{-1} & \alpha_0 & \alpha_1 & \alpha_2 & 0 & 0 \\ 0 & \alpha_{-2} & \alpha_{-1} & \alpha_0 & \alpha_1 & \alpha_2 & 0 \\ 0 & 0 & \alpha_{-2} & \alpha_{-1} & \alpha_0 & \alpha_1 & \alpha_2 \\ \alpha_2 & 0 & 0 & \alpha_{-2} & \alpha_{-1} & \alpha_0 & \alpha_1 \\ \alpha_0 & \alpha_2 & 0 & 0 & \alpha_{-2} & \alpha_{-1} & \alpha_0 \end{pmatrix} \tag{F.124}$$

(2) Chaque ligne de la matrice A_h résulte d'une permutation circulaire de la précédente ; la matrice A_h est bien « circulante ». La structure bande de la matrice est légèrement modifiée par les coefficients qui apparaîssent dans les coins supérieur-droit et inférieur-gauche par application des conditions de périodicité.

Exercice 2.8 (Diagonalisation des matrices circulantes)

(1) On calcule le produit $D_h^-\, D_h^+$ et on vérifie aisément que :

$$D_h^-\, D_h^+ = I \tag{F.125}$$

Par conséquent, les opérateurs D_h^+ et D_h^- sont inverses l'un de l'autre, et visiblement transposés (adjoints) l'un de l'autre, donc orthogonaux :

$$\left(D_h^+\right)^{-1} = D_h^- = \left(D_h^+\right)^T \,, \quad \left(D_h^-\right)^{-1} = D_h^+ = \left(D_h^-\right)^T \tag{F.126}$$

Par conséquent, chacun commute avec son adjoint :

$$\left(D_h^+\right)^*\, D_h^+ = D_h^-\, D_h^+ = I = D_h^+\, D_h^- = D_h^+\, \left(D_h^+\right)^* \tag{F.127}$$

$$\left(D_h^-\right)^*\, D_h^- = D_h^+\, D_h^- = I = D_h^-\, D_h^+ = D_h^-\, \left(D_h^-\right)^* \tag{F.128}$$

Chacun d'eux est donc normal et par conséquent diagonalisable par une transformation unitaire :

$$D_h^+ = \Phi_h^+\, \mathcal{D}_h^+\, {\Phi_h^+}^* \,, \quad {\Phi_h^+}^*\, \Phi_h^+ = I \,, \quad \mathcal{D}_h^+ \text{ diagonale} \tag{F.129}$$

$$D_h^- = \Phi_h^-\, \mathcal{D}_h^-\, {\Phi_h^-}^* \,, \quad {\Phi_h^-}^*\, \Phi_h^- = I \,, \quad \mathcal{D}_h^- \text{ diagonale} \tag{F.130}$$

Enfin, ces opérateurs commutent

$$D_h^+\, D_h^- = D_h^-\, D_h^+ \quad (= I) \tag{F.131}$$

Par conséquent :

$$\Phi_h^+ = \Phi_h^- \tag{F.132}$$

(2) Il s'agit de vérifier que l'opérateur Φ_h défini dans l'énoncé est bien unitaire, qu'il diagonalise effectivement les opérateurs D_h^+ et D_h^- et que les valeurs propres ont bien

la forme indiquée.

Pour tout indice $m = 1, 2, ..., M$, on note $\Phi_h^{(m)}$ le m-ème vecteur colonne de la matrice Φ_h dont la j-ème composante est la la quantité $\Phi_{h\,j,m}$. Pour tout couple d'indices $(m.\ell)$ on calcule le produit hermitien :

$$(\Phi_h^{(m)}, \Phi_h^{(\ell)}) = \sum_{j=1}^{M} \Phi_{h\,j,m}\, \overline{\Phi_{h\,j,\ell}} = \frac{1}{M} \sum_{j=1}^{M} z^j \tag{F.133}$$

où l'on a posé :

$$z = \exp\left[i\,(\theta_m - \theta_\ell)\right] = \exp\left[\frac{2i\pi(m-\ell)}{M}\right] \tag{F.134}$$

Par conséquent, si $\ell \neq m$, $z \neq 1$ et :

$$(\Phi_h^{(m)}, \Phi_h^{(\ell)}) = \frac{1}{M}\, z\, \frac{1 - z^M}{1 - z} = 0 \tag{F.135}$$

car $z^M = 1$, ce qui prouve l'orthogonalité des vecteurs colonnes de la matrice Φ_h. Par contre, pour $\ell = m$, $z = 1$ et l'on a :

$$(\Phi_h^{(m)}, \Phi_h^{(m)}) = \left(\|\Phi_h^{(m)}\|_2\right)^2 = 1 \tag{F.136}$$

ce qui prouve que ces vecteurs sont de norme euclidienne égale à 1. La matrice Φ_h est donc bien unitaire :

$$\Phi_h^*\, \Phi_h = I \tag{F.137}$$

Vérifions maintenant que les vecteurs colonnes $\Phi_h^{(m)}$ ($m = 1, 2, ..., M$) sont bien *des* donc *les* vecteurs propres des matrices D_h^+ et D_h^-. Pour tout indice $j = 1, 2, ..., M$, on a :

$$\left(D_h^+\, \Phi_h^{(m)}\right)_j = \Phi_{h\,j+1,m} = d_m^+\, \Phi_{h\,j,m} \tag{F.138}$$

$$\left(D_h^-\, \Phi_h^{(m)}\right)_j = \Phi_{h\,j-1,m} = d_m^-\, \Phi_{h\,j,m} \tag{F.139}$$

où l'on a posé :

$$d_m^+ = \exp(i\theta_m) \tag{F.140}$$

$$d_m^- = \exp(-i\theta_m) \tag{F.141}$$

On a donc bien :

$$D_h^+ = \Phi_h\, \mathcal{D}_h^+\, \Phi_h^* \tag{F.142}$$

$$D_h^- = \Phi_h\, \mathcal{D}_h^-\, \Phi_h^* \tag{F.143}$$

où $\mathcal{D}_h^+$ et $\mathcal{D}_h^-$ sont les matrices diagonales qui contiennent les valeurs propres $\{d_m^+\}$ et $\{d_m^-\}$ ($m = 1, 2, ..., M$).

(3) Dans le cas d'une discrétisation uniforme de l'intervalle [a,b] :

$$x_j = a + j\,\frac{b-a}{M} \quad (j = 0, 1, 2, ..., M) \tag{F.144}$$

on a :

$$\Phi_{h\,j,m} = \frac{1}{\sqrt{M}}\,\phi_m(x_j) \tag{F.145}$$

où $\phi_m(x)$ est une fonction propre de l'opérateur de dérivée première identifiée à l'exercice 2.2 définie en (2.35) ce qui permet de tirer la conclusion suivante : dans le cas périodique, les modes de Fourier discrets sont les discrétisés des modes de Fourier du modèle linéaire périodique continu.

Exercice 2.9 (Spectres d'opérateurs périodiques particuliers)

L'application de la formule (2.75) aux opérateurs de l'exercice 2.6 fournit les expressions suivantes de leurs valeurs propres génériques respectives :

$$\nabla_h \;:\; \lambda_{h\,m} = 1 - \exp(-i\theta_m) \tag{F.146}$$

$$\nabla_h^2 \;:\; \lambda_{h\,m} = 1 - 2\exp(-i\theta_m) + \exp(-2i\theta_m) = [1 - \exp(-i\theta_m)]^2 \tag{F.147}$$

$$\Delta_h \;:\; \lambda_{h\,m} = \exp(i\theta_m) - 1 \tag{F.148}$$

$$\Delta_h^2 \;:\; \lambda_{h\,m} = \exp(2i\theta_m) - 2\exp(i\theta_m) + 1 = [\exp(i\theta_m) - 1]^2 \tag{F.149}$$

$$\delta_h \;:\; \lambda_{h\,m} = \tfrac{1}{2}\,[\exp(i\theta_m) - \exp(-i\theta_m)] = i\,\sin\,\theta_m \tag{F.150}$$

$$\delta_h^- \;:\; \lambda_{h\,m} = \tfrac{3}{2} - 2\exp(-i\theta_m) + \tfrac{1}{2}\exp(-2i\theta_m) \tag{F.151}$$

$$\delta_h^+ \;:\; \lambda_{h\,m} = -\tfrac{1}{2}\exp(2i\theta_m) + 2\exp(i\theta_m) - \tfrac{3}{2} \tag{F.152}$$

Exercice 2.10 (Conservation de la norme par la transformée de Fourier)

Quel que soit le vecteur u_h, on a :

$$(\|\widehat{u_h}\|_2)^2 = \widehat{u_h}^*\,\widehat{u_h} = u_h^*\,\Phi_h\,\Phi_h^*\,u_h = u_h^*\,u_h = (\|u_h\|_2)^2 \tag{F.153}$$

Rien d'étonnant : la transformée de Fourier discrète Φ_h comme son analogue continu est un opérateur unitaire.

Exercice 2.11 (Transformées de Fourier discrète et continue)

Pour une fonction $u \in L^2([a,b])$, $(b-a)$-périodique :

$$\begin{aligned} \widehat{u}_m &= \int_a^b u(x)\, \exp\left(-2\pi i m \frac{x-a}{b-a}\right) \frac{dx}{b-a} \\ u(x) &= \sum_{m=-\infty}^{\infty} \widehat{u}_m\, \exp\left(2\pi i m \frac{x-a}{b-a}\right) \end{aligned} \tag{F.154}$$

Exercice 2.12 (Spectre de la différence centrée)

En vertu des équations (F.88) et (F.150), on a :

$$\frac{\frac{1}{h}\,\lambda_{h\,m}}{\lambda_m} = \frac{i\,\sin\,\frac{2\pi m h}{b-a}}{i\,\frac{2\pi m h}{b-a}} = 1 + O\left(h^2\right) \quad (h \to 0\,,\ \text{m fixé}) \tag{F.155}$$

Exercice 2.13 (Modes propres dans un cas de conditions de Dirichlet-Neumann)

(1) Les fonctions propres du problème continu sont les éléments $\psi \in H$ tels que :

$$\frac{d^2\psi}{dx^2} = \mu\,\psi \tag{F.156}$$

pour un certain $\mu \in \mathbb{C}$. Comme à l'exercice 2.4, cette équation impose que :

$$\psi(x) = \alpha\, \exp\Big(\sqrt{\mu}(x-a)\Big) + \beta\, \exp\Big(-\sqrt{\mu}(x-a)\Big) \tag{F.157}$$

où α et β sont des constantes que l'on détermine en imposant les conditions aux limites dont celle en $x = b$ a changé :

$$\begin{cases} \alpha + \beta = 0 \\ \sqrt{\mu}\exp\Big(\sqrt{\mu}(b-a)\Big)\,\alpha - \sqrt{\mu}\exp\Big(-\sqrt{\mu}(b-a)\Big)\,\beta = 0 \end{cases} \tag{F.158}$$

A nouveau, ce système admet une solution non triviale, $(\alpha,\beta) \neq (0,0)$, ssi il est singulier, ce qui équivaut ici à la condition suivante :

$$\exp\Big(\sqrt{\mu}(b-a)\Big) = -\exp\Big(-\sqrt{\mu}(b-a)\Big) \tag{F.159}$$

soit :

$$\exp\Big(2\sqrt{\mu}(b-a)\Big) = -1 \tag{F.160}$$

soit encore :

$$\pm 2\sqrt{\mu}(b-a) = (2m-1)\pi i \quad (m \in \mathbb{N}) \tag{F.161}$$

ce qui fournit à nouveau, comme prévu, un ensemble dénombrable de modes propres correspondant aux valeurs propres réelles négatives suivantes :

$$\mu_m = -\Big(\frac{(m-\frac{1}{2})\pi}{b-a}\Big)^2 \tag{F.162}$$

associées aux fonctions propres

$$\begin{aligned}\psi_m(x) &= \frac{1}{2i}\ \Big[\exp\Big(\frac{(m-\frac{1}{2})\pi i(x-a)}{b-a}\Big) - \exp\Big(-\frac{(m-\frac{1}{2})\pi i(x-a)}{b-a}\Big)\Big] \\ &= \sin\Big((m-\tfrac{1}{2})\pi\frac{x-a}{b-a}\Big)\end{aligned} \tag{F.163}$$

obtenues en fixant la constante arbitraire α à $1/(2i)$.

(2) Il est naturel de tester, comme vecteurs propres potentiels, les discrétisés des fonctions propres du problème continu. On considère donc les vecteurs $\{\psi_h^{(m)}\}$ ($m = 1, 2, ..., M$) définis comme suit par leurs composantes :

$$\psi_{h\,j,m} = \sqrt{2h}\,\psi_m(x_j) \quad (j = 1, 2, ..., M) \tag{F.164}$$

où la constante $\sqrt{2h}$ n'est pas essentielle mais sert à normaliser le vecteur. On peut également convenir d'appliquer cette formule pour $j = 0$ et $M + 1$, de sorte qu'alors les termes de bords vérifient les « conditions aux limites » au sens suivant : pour tout mode $m = 1, 2, ..., M$:

$$\psi_{h\,0,m} = 0 \tag{F.165}$$

$$\psi_{h\,M+1,m} = \sqrt{2h}\,\psi_m(b+\frac{h}{2}) = \sqrt{2h}\,\psi_m(b-\frac{h}{2}) = \psi_{h\,M-1,m} \tag{F.166}$$

On remarque ensuite que l'on a :

$$\Big(A_h\ \psi_h^{(m)}\Big)_j = \psi_{h\,j-1,m} - 2\psi_{h\,j,m} + \psi_{h\,j+1,m} \tag{F.167}$$

pour tout $j = 1, 2, ..., M$ (limites incluses) en vertu des conditions aux bords satisfaites par le vecteur $\{\psi_h^{(m)}\}$. Il vient donc :

$$\begin{aligned}\Big(A_h\ \psi_h^{(m)}\Big)_j &= -2\psi_{h\,j,m} + \sqrt{2h}\ [\sin(j-1)\theta_m + \sin(j+1)\theta_m] \\ &= \mu_{h\,m}\,\psi_{h\,j,m}\end{aligned} \tag{F.168}$$

où l'on a posé, comme de coutume :

$$\mu_{h\,m} = -2 + 2\,\cos\theta_m \tag{F.169}$$

ce qui prouve que le vecteur $\{\psi_h^{(m)}\}$ est effectivement un vecteur propre de la matrice A_h. L'expression de la valeur propre $\mu_{h\,m}$ en fonction du paramètre de fréquence θ_m reste la même, seule change la forme prise par ce paramètre en fonction de l'indice m et en conséquence l'allure des vecteurs propres. En particulier, pour le mode de plus basse fréquence, $\psi_h^{(1)}$, l'intervalle d'étude $[a, b]$ correspond ici au quart de sa période.

Exercice 3.1 (Propriétés de lissage de l'itération de Gauss-Seidel en périodique)

(1) Dans le cas du modèle discret considéré, l'itération de Gauss-Seidel prend la forme suivante :

$$\begin{aligned} &\text{Pour } j = 1, 2, ..., M \text{ (dans cet ordre) :} \\ &u_j^{n+1} = \tfrac{1}{2} u_{j-1}^{n+1} + \tfrac{1}{2} u_{j+1}^{n} + \tfrac{1}{2} h^2\, f_j \end{aligned} \tag{F.170}$$

ou, de manière équivalente :

$$\begin{aligned} &\text{Pour } j = 1, 2, ..., M \text{ (dans cet ordre) :} \\ &e_j^{n+1} = \tfrac{1}{2} e_{j-1}^{n+1} + \tfrac{1}{2} e_{j+1}^{n} \end{aligned} \tag{F.171}$$

où e_j^n désigne la j-ème composante du vecteur erreur e^n à l'itération n. Le facteur d'amplification (ou d'atténuation) $g(\theta)$ s'obtient en évaluant l'effet de l'itération sur un mode de Fourier exprimé comme suit :

$$e_j^n = C\, e^{ij\theta} \quad (\forall\, j\,;\ \theta \in [-\pi, \pi]) \tag{F.172}$$

dont on sait qu'il s'agit d'un mode propre de l'itération ce qui permet de poser :

$$e_j^{n+1} = C\, g_1(\theta)\, e^{ij\theta} \quad (\forall\, j) \tag{F.173}$$

En reportant ces expressions dans l'équation de l'erreur, il vient :

$$g_1(\theta) = \tfrac{1}{2}\, g_1(\theta)\, e^{-i\theta} + \tfrac{1}{2}\, e^{i\theta} \tag{F.174}$$

Il en résulte :

$$g_1(\theta) = \frac{e^{i\theta}}{2 - e^{-i\theta}} \tag{F.175}$$

et

$$|g_1(\theta)| = \frac{1}{\sqrt{(2 - \cos\,\theta)^2 + \sin^2\,\theta}} = \frac{1}{\sqrt{5 - 4\cos\,\theta}} = \frac{1}{\sqrt{1 + 8\sin^2\,\dfrac{\theta}{2}}} \tag{F.176}$$

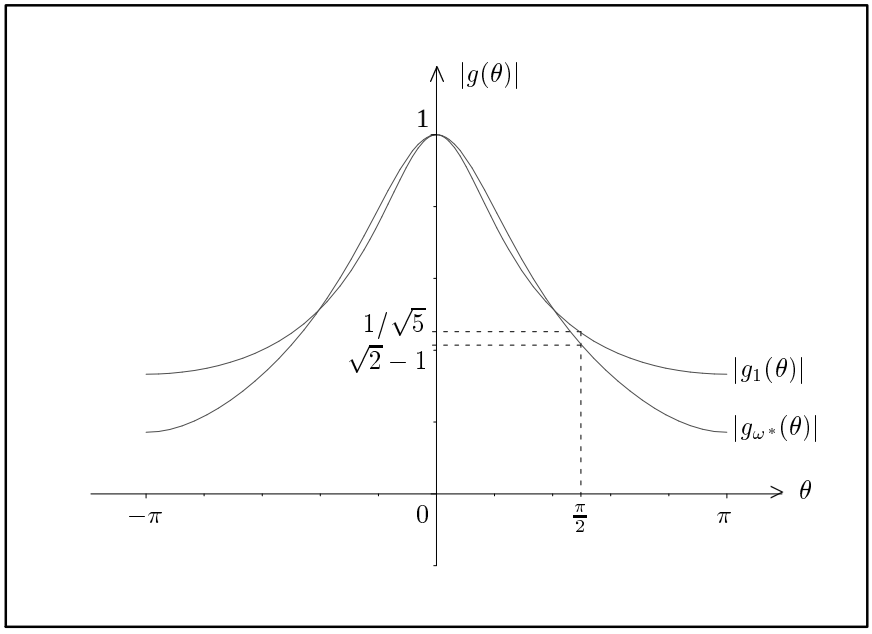


Figure F.2. Facteur d'atténuation de la méthode de Gauss-Seidel en périodique : (a) $|g_1(\theta)|$ ($\omega = 1$) ; (b) avec sous-relaxation, $|g_{w^*}(\theta)|$ ($\omega^* = 2(\sqrt{2}-1)$) ; cf. exercice 3.1.

Cette fonction est représentée à la figure F.2. On y observe que l'atténuation est d'autant meilleure que la fréquence est élevée.

(2) Lorsqu'on inclut une étape de sur ou sous-relaxation, l'itération prend la forme suivante :

Pour $j = 1, 2, \ldots, M$ (dans cet ordre) :

$$v_j^{n+1} = \tfrac{1}{2} u_{j-1}^{n+1} + \tfrac{1}{2} u_{j+1}^{n} + \tfrac{1}{2} h^2 f_j \qquad \text{(F.177)}$$

$$u_j^{n+1} = u_j^n + \omega \left(v_j^{n+1} - u_j^n \right) \qquad \text{(F.178)}$$

de sorte que :

$$u_j^{n+1} = (1-\omega)\, u_j^n + \frac{\omega}{2} \left(u_{j-1}^{n+1} + u_{j+1}^n + h^2 f_j \right) \qquad \text{(F.179)}$$

et pour ce qui est de l'erreur :

$$e_j^{n+1} = (1-\omega)\, e_j^n + \frac{\omega}{2} \left(e_{j-1}^{n+1} + e_{j+1}^n \right) \qquad \text{(F.180)}$$

En injectant un mode de Fourier de fréquence θ dans cette équation, le facteur d'amplification $g_\omega(\theta)$ en résulte :

$$g_\omega(\theta) = \frac{1-\omega+\dfrac{\omega}{2}\,e^{i\theta}}{1-\dfrac{\omega}{2}\,e^{-i\theta}} \tag{F.181}$$

ce qui généralise l'expression de $g_1(\theta)$ de la question précédente. Il vient :

$$\begin{aligned}
|g_\omega(\theta)|^2 &= \frac{\left(1-\omega+\dfrac{\omega}{2}\,\cos\,\theta\right)^2+\dfrac{\omega^2}{4}\,\sin^2\,\theta}{\left(1-\dfrac{\omega}{2}\,\cos\,\theta\right)^2+\dfrac{\omega^2}{4}\,\sin^2\,\theta} \\
&= \frac{(1-\omega)^2+\dfrac{\omega^2}{4}+\omega(1-\omega)\,\cos\,\theta}{1+\dfrac{\omega^2}{4}-\omega\,\cos\,\theta} \\
&= \omega-1+\frac{(1-\omega)^2+\dfrac{\omega^2}{4}+(1-\omega)\left(1+\dfrac{\omega^2}{4}\right)}{1+\dfrac{\omega^2}{4}-\omega\,\cos\,\theta}
\end{aligned} \tag{F.182}$$

On voit que si $\omega \leq 1$, la quantité $|g_\omega(\theta)|^2$ est une fonction décroissante de $\cos\,\theta$, donc croissante de $\theta \in [0,\pi]$. Par conséquent, dans cette hypothèse, dans le but d'optimiser les performances de « lissage », on se limite à la considération des modes de hautes fréquences ($\theta \in [\pi/2,\pi]$), et on minimise le maximum suivant :

$$\phi(\omega) \stackrel{\text{déf}}{=} \min_{\theta\,\in\,[\pi/2,\pi]} |g_\omega(\theta)|^2 = \left|g_\omega(\frac{\pi}{2})\right|^2 \tag{F.183}$$

Il vient :

$$\phi(\omega) = \frac{(1-\omega)^2+\dfrac{\omega^2}{4}}{1+\dfrac{\omega^2}{4}} \tag{F.184}$$

On obtient facilement :

$$\phi'(\omega) = \frac{\omega^2+4\,\omega-4}{2\,\left(1+\dfrac{\omega^2}{4}\right)} \tag{F.185}$$

ce qui permet de dresser le tableau de variation suivant :

ω	0		$2(\sqrt{2}-1)$		1
$\phi'(\omega)$	-2	–	0	+	$\frac{8}{25}$
$\phi(\omega)$	1	$\searrow$	$3-2\sqrt{2}$	$\nearrow$	$\frac{1}{5}$

Par conséquent le minimum de $\phi(\omega)$ est atteint pour :

$$\omega = \omega^* = 2(\sqrt{2}-1) \approx 0.8284 \tag{F.186}$$

La fonction $|g_{\omega^*}(\theta)|$ qui en résulte est représentée à la figure F.2. On a en particulier,

$$\left|g_{\omega^*}(\frac{\pi}{2})\right| = \sqrt{3-2\sqrt{2}} = \sqrt{2}-1 \approx 0.4142 \tag{F.187}$$

au lieu de $1/\sqrt{5} \approx 0.4472$ précédemment obtenu sans sous-relaxation ($\omega = 1$).

Exercice 3.2 (Itération de Gauss-Seidel pour le problème de Dirichlet)

Soit g une valeur propre de l'itération et (u, v) un vecteur propre associé non nul. En reportant les relations indiquées en (3.176) dans (3.174) et (3.175), il vient :

$$g\,v = g\,B\,v + C\,u \tag{F.188}$$
$$g\,u = u + \omega\,(g\,v - u) \tag{F.189}$$

Pour identifier le rayon spectral, on s'intéresse aux valeurs propres non nulles ($g \neq 0$) pour lesquelles :

$$v = \frac{g-1+\omega}{g\omega}\,u \tag{F.190}$$

de sorte que le système aux valeurs propres précédent se réduit à :

$$H\,u = 0 \quad (u \neq 0) \tag{F.191}$$

où l'on a posé :

$$\begin{aligned} H &= (g-1+\omega)\,(I-B) - \omega\,C \\ &= \text{Trid}_{D\,D}\left(a(g,\omega), b(g,\omega), c(g,\omega)\right) \end{aligned} \tag{F.192}$$

et

$$a(g,\omega) = -\frac{g-1+\omega}{2} \tag{F.193}$$

$$b(g,\omega) = g-1+\omega \tag{F.194}$$

$$c(g,\omega) = -\frac{\omega}{2} \tag{F.195}$$

La condition équivaut à dire que la matrice H est singulière ou que l'une de ses valeurs propres est nulle. Or ces valeurs propres, notées $\{h_m\}$, s'expriment par la formule générale (2.191) ou (2.195) qui ici s'écrit comme suit :

$$h_m = g-1+\omega-2\sqrt{\frac{g-1+\omega}{2}.\frac{\omega}{2}}\,\cos\,\theta_m \quad (m=1,2,...,M) \tag{F.196}$$

Les valeurs de g pour lesquelles il existe $h_m = 0$ sont donc celles pour lesquelles on a :

$$g-1+\omega = \omega\,\cos^2\,\theta_m \tag{F.197}$$

(pour un certain indice m), à savoir :

$$g_m = 1-\omega\,\sin^2\,\theta_m \quad (m=1,2,...,M) \tag{F.198}$$

Cette expression est identique à celle des valeurs propres de l'itération de Jacobi appliquée à un système algébrique « équivalent » pour lequel le spectre de la matrice A_h serait constitué des nombres $\{\sin^2\,\theta_m\}$ et dont le paramètre de relaxation τ serait égal à ω. On note que les modes fréquentiels interviennent par le terme $\sin^2\,\theta_m$; par conséquent, le facteur d'atténuation de modes associés à des fréquences complémentaires à π est identique. Les modes de hautes fréquences sont donc globalement traités de la même manière que ceux de basses fréquences, et l'optimisation admet le même résultat dans les cas (a) et (b). Les modes qui limitent le spectre sont ici associés à la plus basse fréquence $\theta_1 = \pi h$ et à la fréquence moyenne $\pi/2$ (en supposant M impair). En application des résultats de l'exercice 1.2, il vient :

$$\omega^* = \left(\frac{\sin^2\,\theta_1+1}{2}\right)^{-1} = \frac{2}{1+\sin^2\,\pi h} \tag{F.199}$$

qui met en évidence que la vitesse de convergence est maximale par sur-relaxation :

$$1 < \omega^* < 2 \tag{F.200}$$

Le rayon spectral en résulte :

$$\rho^* = g_1 = \omega^* - 1 = \frac{1-\sin^2\,\pi h}{1+\sin^2\,\pi h} = 1-2\,\pi^2 h^2+\ldots \tag{F.201}$$

ainsi que la vitesse de convergence :

$$v^* = -\ln\ \rho^* = 2\,\pi^2 h^2 + \ldots \tag{F.202}$$

Pour l'itération de base ($\omega = 1$), le rayon spectral a été évalué à la section 3.2 :

$$\rho = \cos^2\ \pi h = 1 - \pi^2 h^2 + \ldots \tag{F.203}$$

ce qui correspond à la vitesse de convergence

$$v = -\ln\ \rho = \pi^2 h^2 + \ldots \tag{F.204}$$

Par conséquent, la sur-relaxation a permis de réaliser un gain en vitesse de convergence proche de 2 :

$$\frac{v^*}{v} \approx 2 \tag{F.205}$$

Les courbes représentatives du facteur d'amplification,

$$g_\omega(\theta) = 1 - \omega\ \sin^2\ \theta \tag{F.206}$$

en fonction du paramètre de fréquence $\theta \in [0, \pi]$ sont indiquées à la figure F.3 pour $\omega = 1$ (itération de base) et $\omega = \omega^*$ (itération optimisée).

Exercice 4.1 (Approximation de grille grossière)

On se place dans le cas de deux grilles emboîtées unidimensionnelles et uniformes : $\mathcal{M}_h$, la grille fine de dimension M_h, et $\mathcal{M}_{2h}$, la grille grossière de dimension M_{2h} et l'on a : $M_h = 2M_{2h} + 1$. Soit

$$u_{2h} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{M_{2h}} \end{pmatrix} \tag{F.207}$$

un vecteur quelconque défini sur la grille grossière. L'opérateur P de prolongement étant celui qui correspond à l'injection naturelle aux points communs aux deux grilles et à l'interpolation linéaire aux autres, on a :

$$P\,u_{2h} = \begin{pmatrix} \frac{1}{2}(0 + u_1) \\ u_1 \\ \frac{1}{2}(u_1 + u_2) \\ u_2 \\ \frac{1}{2}(u_2 + u_3) \\ u_3 \\ \vdots \end{pmatrix} \tag{F.208}$$

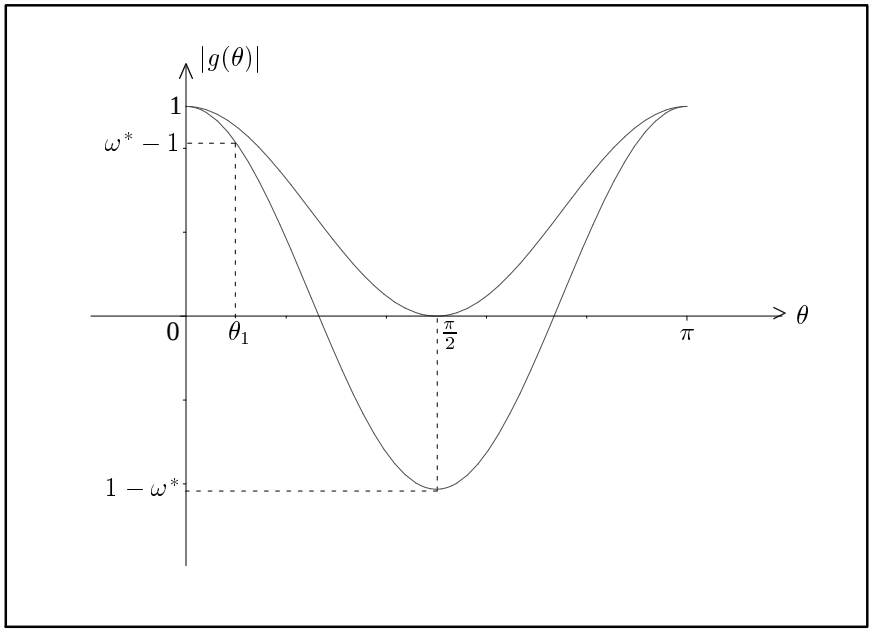


Figure F.3. Facteur d'amplification en fonction du paramètre de fréquence $\theta \in [0, \pi]$ pour l'itération de Gauss-Seidel appliquée au problème modèle de Dirichlet : itération de base, $g_1(\theta)$, et itération optimisée, $g_\omega(\theta)$; figure tracée pour $M = 9$, $\theta_1 = \pi/10$; cf. exercice 3.2

Par conséquent si $A_h = h^{-2}\,\mathrm{Trid}\,(-1, 2, -1)$ (de dimension $M_h \times M_h$), il vient :

$$A_h\, P\, u_{2h} = \frac{1}{h^2} \begin{pmatrix} 2\,\frac{1}{2}u_1 - u_1 \\ -\frac{1}{2}u_1 + 2\,u_1 - \frac{1}{2}(u_1 + u_2) \\ -u_1 + 2\,\frac{1}{2}(u_1 + u_2) - u_2 \\ -\frac{1}{2}(u_1 + u_2) + 2\,u_2 - \frac{1}{2}(u_2 + u_3) \\ -u_2 + 2\,\frac{1}{2}(u_2 + u_3) - u_3 \\ -\frac{1}{2}(u_2 + u_3) + 2\,u_3 - \frac{1}{2}(u_3 + u_4) \\ \vdots \end{pmatrix}$$

$$= \frac{1}{h^2} \begin{pmatrix} 0 \\ u_1 - \frac{1}{2}u_2 \\ 0 \\ -\frac{1}{2}u_1 + u_2 - \frac{1}{2}u_3 \\ 0 \\ -\frac{1}{2}u_2 + u_3 - \frac{1}{2}u_4 \\ \vdots \end{pmatrix} \qquad \text{(F.209)}$$

Le choix $R = \frac{1}{2}P^T$ conduit alors à :

$$R\,A_h\,P\,u_{2h} = \frac{1}{4h^2}\begin{pmatrix} 2u_1 - u_2 \\ -u_1 + 2u_2 - u_3 \\ -u_2 + 2u_3 - u_4 \\ \vdots \end{pmatrix} = \frac{1}{(2h)^2}\,\text{Trid}\,(-1,2,-1)\,u_{2h} \quad \text{(F.210)}$$

où la matrice tridiagonale est de dimension $M_{2h} \times M_{2h}$. Le vecteur u_{2h} étant quelconque, cette relation permet d'identifier les opérateurs :

$$R\,A_h\,P = \frac{1}{(2h)^2}\,\text{Trid}\,(-1,2,-1) = A_{2h} \quad \text{(F.211)}$$

Exercice 4.2 (Expérimentation numérique d'enrichissement progressif de maillage)

(1) On réfère à l'exercice 1.1 pour les notations et les résultats de convergence. On a :

$$\mu = \max_{x\,\in\,[0,1]} \left|-\pi^4 \sin\,\pi x\right| = \pi^4 \quad \text{(F.212)}$$

et :

$$\|\,A_h^{-1}\,\|_\infty = \frac{1}{8} \quad \text{(F.213)}$$

de sorte que si le résidu satisfait le critère proposé, alors on a bien :

$$\|\,e_h^n\,\|_\infty = \|\,A_h^{-1}\,r_h^n\,\|_\infty \leq \|\,A_h^{-1}\,\|_\infty . \|\,r_h^n\,\|_\infty \leq \frac{1}{8}.\frac{\mu h^2}{12} = \frac{\mu h^2}{96} \quad \text{(F.214)}$$

Pour $h = \frac{1}{32}$, on a :

$$\frac{\mu h^2}{96} \approx 0.99\,10^{-3} \quad \text{(F.215)}$$

(2) A la suite d'un calcul monogrille ($M + 1 = 32$), on a tracé la courbe de variation (de la norme) du résidu (en échelle logarithmique) en fonction des itérations à la figure F.4 (a) (ainsi qu'en (b) et à la figure suivante comme calcul de référence). Naturellement, cette courbe est une droite illustrant dans ces échelles une décroissance géométrique. On a également consigné les résultats requis dans le tableau F.1 dans lequel le residu est mesuré en norme-infinie, l'abréviation « UT » désigne les unités de travail dépensées (nombre d'itérations $\times$ nombre de points intérieur M). La quantité UT est un indicateur approximatif de ce que serait la partie principale du coût du calcul, les itérations de relaxation, pour un problème numériquement moins trivial.

$M+1$	résidu initial	résidu final	itérations effectuées	UT
32	.987D+01	.791D-02	1477	45787

Tableau F.1. Résolution itérative monogrille jusqu'à satisfaction du critère d'arrêt

Dans cette expérience, le résidu a été atténué d'un facteur égal au rapport $.791D-02/.987D+01 \approx 0.8\,10^{-3}$. Notons que le nombre de conditionnement du système peut être évalué (cf. exercices 1.2 et 1.5) :

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \text{tg}^{-2}\,\frac{\theta_1}{2} = \text{tg}^{-2}\,\frac{\pi}{2(M+1)} = \text{tg}^{-2}\,\frac{\pi}{64} \approx 414 \tag{F.216}$$

ainsi que le rayon spectral de l'itération de Jacobi lorsque le paramètre de relaxation est optimisé :

$$\rho^* = \frac{\kappa - 1}{\kappa + 1} \approx 0.995185 \tag{F.217}$$

Par conséquent, le nombre d'itérations n nécessaires à la réduction observée du résidu pouvait être estimé *a priori* à :

$$n = \frac{\log\ 0.8\,10^{-3}}{\log\ \rho^*} \approx 1477 \tag{F.218}$$

ce qui est conforme au résultat observé.

(3) Avant toute expérience numérique, évaluons la précision de la formule d'interpolation à 4 points de l'énoncé.

Le polynôme d'interpolation de Lagrange $P_3(x)$ d'une fonction régulière notée ici $u'(x)$ admettant les valeurs

$$u'_{2j-2} = u_{j-1} \tag{F.219}$$

$$u'_{2j} = u_j \tag{F.220}$$

$$u'_{2j+2} = u_{j+1} \tag{F.221}$$

$$u'_{2j+4} = u_{j+2} \tag{F.222}$$

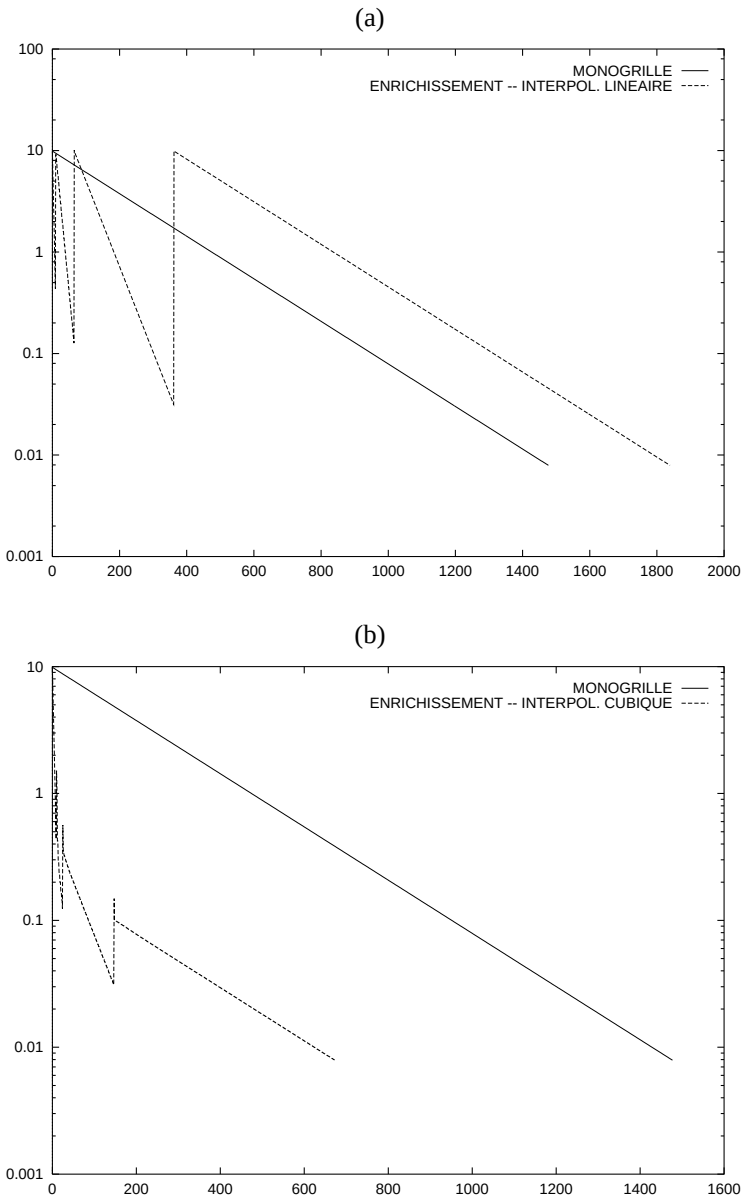


Figure F.4. Comparaison des courbes de convergence du résidu en fonction des itérations de l'algorithme monogrille ($M+1=32$) et de l'algorithme par enrichissement progressif ($M+1=4,8,16,32$) : (a) interpolation linéaire, (b) interpolation cubique.

respectivement aux abscisses

$$x_{2j-2} = (2j-2)h \tag{F.223}$$
$$x_{2j} = 2jh \tag{F.224}$$
$$x_{2j+2} = (2j+2)h \tag{F.225}$$
$$x_{2j+4} = (2j+4)h \tag{F.226}$$

admet l'expresion suivante (« forme de Newton ») :

$$\begin{aligned} P_3(x) &= u'_{2j-2} \\ &+ \frac{u'_{2j} - u'_{2j-2}}{2h}\ [x-(2j-2)h] \\ &+ \frac{u'_{2j+2} - 2u'_{2j} + u'_{2j-2}}{8h^2}\ [x-(2j-2)h]\ [x-2jh] \\ &+ \frac{u'_{2j+4} - 3u'_{2j+2} + 3u'_{2j} - u'_{2j-2}}{48h^3}\ [x-(2j-2)h]\ [x-2jh]\ [x-(2j+2)h] \end{aligned} \tag{F.227}$$
(F.228)

Les formules d'interpolation de l'énoncé correspondent à des valeurs particulières du polynôme $P_3(x)$: (4.32) s'obtient en faisant $j = 1$ et $x = h$; (4.33) en faisant $x = (2j+1)h$; (4.34) en faisant $j = (M+1)/2 - 2$ et $x = Mh$. Par conséquent, il s'agit dans les 3 cas d'une interpolation cubique de ce type. En supposant que la fonction interpolée u' est de classe $C^4[0,1]$, on sait que l'erreur d'interpolation est de la forme suivante :

$$\begin{aligned} \varepsilon(x) &\stackrel{\text{déf}}{=} u'(x) - P_3(x) \\ &= \frac{1}{4!}\ [x-(2j-2)h]\ [x-2jh]\ [x-(2j+2)h]\ [x-(2j+4)h]\ \frac{d^4u'}{dx^4}(\xi) \\ &= O\left(h^4\right) \end{aligned} \tag{F.229}$$

où $\xi \in]0,1[$. Ce résultat pouvait également être atteint en substituant des développements limités aux différents termes des expressions proposées.

Ayant établi la consistance des formules d'interpolation, on procède ensuite à l'expérience d'enrichissement progressif de maillage suivant la procédure décrite dans l'énoncé. Les résultats des opérations effectuées sur un niveau de grille occupent désormais une ligne de tableau, auquel on ajoute une colonne « UTC » des unités de travail cumulées depuis l'initialisation de la solution sur la grille la plus grossière. Les résultats sont consignés au tableau F.2 dans le cas de l'interpolation linéaire et au tableau F.3 pour l'interpolation cubique.

Les courbes de convergence du résidu, qui, quel que soit le niveau de grille a une définition consistante à la quantité $-(u_{xx} + f)$, sont indiquées à la figure F.4 (a) et

$M+1$	résidu initial	résidu final	itérations effectuées	UT	UTC
4	.987D+01	.436	9	27	27
8	.912D+01	.127	54	378	405
16	.101D+02	.316D-01	296	4440	4845
32	.993D+01	.791D-02	1477	45787	50632

Tableau F.2. Résolution itérative par enrichissement progressif du maillage et prolongement par interpolation linéaire

$M+1$	résidu initial	résidu final	itérations effectuées	UT	UTC
4	.987D+01	.436	9	27	27
8	.151	.124	14	98	125
16	.564	.309D-01	121	1815	1940
32	.148	.791D-02	525	16275	18215

Tableau F.3. Résolution itérative par enrichissement progressif du maillage et prolongement par interpolation cubique

(b). Cependant, en abscisse sont portées les itérations dont le coût diffère d'un niveau à l'autre. Ces courbes sont donc indicatrices de l'historique de convergence, mais ne constituent pas exactement un diagramme du résidu, c'est-à-dire du degré de convergence, en fonction du coût.

La technique d'enrichissement progressif se révèle remarquablement inefficace dans le cas du prolongement par interpolation linéaire. Dans ce cas particulier, on constate que chaque prolongement a pour effet de faire remonter le résidu à une valeur proche de 10, indépendamment du niveau de grille, faisant perdre le gain potentiel de l'itération sur la grille précédente. Observons-en le mécanisme : immédiatement après le prolongement, en un nœud d'indice impair, $x = x_{2j-1}$, de la grille fine $\mathcal{M}_h$ on a le

résidu suivant :

$$r_{h\,2j-1} = \frac{-u'_{2j-2} + 2u'_{2j-1} - u'_{2j}}{h^2} - f_{h\,2j-1} = -f_{h\,2j-1} \tag{F.230}$$

alors qu'en un nœud d'indice pair :

$$\begin{aligned} r_{h\,2j} &= \frac{-u'_{2j-1} + 2u'_{2j} - u'_{2j+1}}{h^2} - f_{h\,2j} \\ &= \frac{-\dfrac{u'_{2j-2} + u'_{2j}}{2} + 2u'_{2j} - \dfrac{u'_{2j} + u'_{2j+2}}{2}}{h^2} - f_{h\,2j} \\ &= \frac{-u'_{2j-2} + 2u'_{2j} - u'_{2j+2}}{2h^2} - f_{h\,2j} \\ &= 2\left[r_{H\,j} + f_{H\,j}\right] - f_{h\,2j} \\ &= f_{h\,2j} + O\left(H^2\right) \end{aligned} \tag{F.231}$$

car :

$$f_{H\,j} = f_{h\,2j} \tag{F.232}$$

et $r_{H\,j} = O\left(H^2\right)$ est la valeur du résidu au même point, calculé sur la grille grossière $\mathcal{M}_H$ à l'issue de l'itération sur cette grille, c'est-à-dire à satisfaction du critère d'arrêt. On constate donc qu'à des termes en $O\left(H^2\right)$ près, les composantes du résidu après prolongement de la solution par interpolation linéaire, sont les nombres : $-f_{h\,1},\ f_{h\,2},\ -f_{h\,3},\ f_{h\,4},\ \ldots$. Ces composantes sont donc en général de l'ordre de l'unité et présentent une alternance de signe dans les zones ou le terme source garde un signe constant. Dans l'expérience présente, leur maximum est proche de $\pi^2 \approx 10$, et le signe change à chaque nœud ; le résidu contient donc une composante de la plus haute fréquence proche de 10. En conséquence, sur la grille fine, l'initialisation ne permet de réaliser aucune économie sur le nombre d'itérations nécessaires à la satisfaction du critère d'arrêt. L'interpolation linéaire est donc insuffisamment précise.

A l'inverse, l'interpolation cubique se révèle suffisamment précise. On observe une certaine remontée du résidu par l'effet de l'interpolation, mais par un facteur qui semble borné par une constante inférieure à 5. La technique d'enrichissement permet de réaliser un gain en UTC proche de 2.5.

Il était prévisible qu'une interpolation au moins cubique serait suffisante au succès de la procédure. Pour le comprendre, examinons comment le résidu se transfère de la grille $\mathcal{M}_H$ à la grille $\mathcal{M}_h$ ($H = 2h$). A l'issue de l'itération sur $\mathcal{M}_H$, on a :

$$r_H = A_H\, u_H - f_H \tag{F.233}$$

et :

$$\| r_H \|_\infty \leq \frac{\pi^2 H^2}{12} \tag{F.234}$$

On prolonge alors le vecteur des inconnues :

$$u_h = P\, u_H \tag{F.235}$$

où l'opérateur P correspond à l'injection directe aux points communs aux deux grilles, et à une certaine interpolation aux points qui appartiennent seulement à la grille fine $\mathcal{M}_h$. On calcule ensuite un résidu sur la grille fine :

$$\begin{aligned} r_h &= A_h\, u_h - f_h \\ &= A_h\, P\, A_H^{-1}\, (r_H + f_H) - f_h \\ &= A_h\, P\, A_H^{-1}\, r_H + A_h\, P\, A_H^{-1}\, f_H - f_h \end{aligned} \tag{F.236}$$

Décomposons les vecteurs r_H et f_H dans la base des vecteurs propres $\{S_H^{(m)}\}$ ($m = 1, 2, ..., M_H$) de la matrice A_H :

$$r_H = \sum_{m=1}^{M_H} \widehat{r_H}_m\, S_H^{(m)} \tag{F.237}$$

$$f_H = \sum_{m=1}^{M_H} \widehat{f_H}_m\, S_H^{(m)} \tag{F.238}$$

et posons :

$$\varepsilon_h^{(m)} \overset{\text{déf}}{=} P\, S_H^{(m)} - S_h^{(m)} \tag{F.239}$$

En fait, en raison de l'effet de lissage de l'itération de Jacobi conduite sur la grille $\mathcal{M}_H$, on peut penser que la suite des coefficients (de Fourier) $\{\widehat{r_H}_m\}$ ($m = 1, 2, ..., M_H$) est rapidement décroissante. On peut faire une hypothèse analogue à propos de la suite des coefficients $\{\widehat{f_H}_m\}$ à condition que la grille grossière $\mathcal{M}_H$ soit néanmoins suffisamment dense pour assurer une représentation précise de la fonction f.

Les composantes nodales du vecteur $\varepsilon_h^{(m)}$ sont égales aux erreurs d'interpolation du vecteur propre $S_H^{(m)}$ par prolongement sur la grille fine $\mathcal{M}_h$. Par conséquent, l'ordre de grandeur de la norme de ce vecteur dépend directement de la précision de la formule d'interplation.

Il vient :

$$\begin{aligned} r_h = &\sum_{m=1}^{M_H} \frac{\lambda_{h\,m}}{\lambda_{H\,m}}\, \widehat{r_H}_m\, S_h^{(m)} + A_h \sum_{m=1}^{M_H} \frac{\widehat{r_H}_m}{\lambda_{H\,m}}\, \varepsilon_h^{(m)} \\ &+ \sum_{m=1}^{M_H} \frac{\lambda_{h\,m}}{\lambda_{H\,m}}\, \widehat{f_H}_m\, S_h^{(m)} + A_h \sum_{m=1}^{M_H} \frac{\widehat{f_H}_m}{\lambda_{H\,m}}\, \varepsilon_h^{(m)} \\ &- f_h \end{aligned} \tag{F.240}$$

où les valeurs propres

$$\lambda_{h\,m} = \frac{2 - 2\,\cos\,m\pi h}{h^2} \tag{F.241}$$

$$\lambda_{H\,m} = \frac{2 - 2\,\cos\,m\pi H}{H^2} \tag{F.242}$$

sont des approximations de la même valeur propre du problème continu. On décompose ensuite le vecteur f_h dans la base des vecteurs propres de la matrice A_h :

$$f_h = \sum_{m=1}^{M_h} \widehat{f_{h\,m}}\, S_h^{(m)} \tag{F.243}$$

et on réarrange (F.236) comme suit :

$$r_h = r_H^P + \varepsilon_f + \Delta_f + \Delta_r \tag{F.244}$$

où l'on a posé :

$$r_H^P = \sum_{m=1}^{M_H} \frac{\lambda_{h\,m}}{\lambda_{H\,m}}\, \widehat{r_{H\,m}}\, S_h^{(m)} \tag{F.245}$$

$$\varepsilon_f = \sum_{m=1}^{M_H} \frac{\lambda_{h\,m}}{\lambda_{H\,m}}\, \widehat{f_{H\,m}}\, S_h^{(m)} - \sum_{m=1}^{M_h} \widehat{f_{h\,m}}\, S_h^{(m)} \tag{F.246}$$

$$\Delta_f = \sum_{m=1}^{M_H} \frac{\widehat{f_{H\,m}}}{\lambda_{H\,m}}\, A_h\, \varepsilon_h^{(m)} \tag{F.247}$$

$$\Delta_r = \sum_{m=1}^{M_H} \frac{\widehat{r_{H\,m}}}{\lambda_{H\,m}}\, A_h\, \varepsilon_h^{(m)} \tag{F.248}$$

Le terme r_H^P dans (F.245) a bien l'ordre voulu car $\widehat{r_{H\,m}} = O\left(H^2\right)$. Noter en particulier que :

$$\frac{\lambda_{h\,m}}{\lambda_{H\,m}} = \frac{\dfrac{4}{h^2}\,\sin^2\,\dfrac{m\pi h}{2}}{\dfrac{4}{H^2}\,\sin^2\,\dfrac{m\pi H}{2}} = \frac{1}{\cos^2\,\dfrac{m\pi h}{2}} \leq 2 \tag{F.249}$$

lorsque $m \leq M_H = M_{2h}$.

Le terme ε_f dans (F.246) a bien l'ordre voulu à condition que les maillages $\mathcal{M}_H$ et $\mathcal{M}_h$ soient structurellement proches et que le maillage grossier $\mathcal{M}_H$ soit suffisamment dense pour prendre en compte le contenu fréquentiel du terme source. Dans l'exemple traité, le terme source ne contient que le mode de plus basse fréquence et cette condition est remplie.

Le terme Δ_f dans (F.247) est incertain car $\widehat{f_{H\,m}} = O\,(1)$ en général, $1/\lambda_{H\,m} = O\,(1)$ est la seule majoration uniforme en h possible pour m fixé (basse fréquence), et si $\varepsilon_h^{(m)}$ est bien infiniment petit, l'opérateur A_h n'est pas borné ($h \to 0$) ; en effet :

$$\| A_h \|_\infty = \frac{4}{h^2} \tag{F.250}$$

Le terme Δ_r dans (F.248) n'est pas dimensionnant car d'ordre supérieur au terme précédent.

En définitive, c'est le terme Δ_f qui constitue la quantité critique. Une condition suffisante pour assurer que :

$$\| \Delta_f \|_\infty = O\left(h^2\right) \tag{F.251}$$

est donc la suivante :

$$\frac{\| \varepsilon_h^{(m)} \|_\infty}{h^2} = O\left(h^2\right) \tag{F.252}$$

soit finalement :

$$\| \varepsilon_h^{(m)} \|_\infty = O\left(h^4\right) \tag{F.253}$$

ce qui justifie l'emploi de l'interpolation cubique, alors qu'avec l'interpolation linéaire la quantité $\| \Delta_f \|_\infty$ est de l'ordre de l'unité, conformément au phénomène observé.

Enfin, dans le but d'apporter une démonstration expérimentale supplémentaire de l'intérêt qu'il y a à utiliser une formule d'interpolation précise, on a refait l'expérience d'enrichissement avec cette fois-ci une « interpolation spectrale ». Celle-ci consiste à d'abord appliquer au vecteur de grille grossière u_H la décomposition de Fourier discrète (ici en modes sinusoïdaux) :

$$\widehat{u_H} = S_H^{-1}\, u_H = S_H\, u_H \tag{F.254}$$

puis à compléter ce vecteur de composantes fréquentielles nulles, et enfin à synthétiser un vecteur de grille fine par transformée de Fourier discrète inverse :

$$u_h \overset{\text{déf}}{=} P\, u_H = S_h \begin{pmatrix} \widehat{u_H} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{F.255}$$

De manière explicite :

$$\widehat{u_{H\,m}} = \sqrt{2H} \sum_{j=1}^{M_H} \sin\left(2j \frac{m\pi}{M_h + 1}\right) \underbrace{u_{H\,j}}_{\overset{\text{déf}}{=}\, u_{h\,2j}} \qquad (m = 1, 2, ..., M_H) \tag{F.256}$$

et :

$$u_{h\,\ell} = \sqrt{2H} \sum_{m=1}^{M_H} \sin\left(\ell \frac{m\pi}{M_h+1}\right) \widehat{u_H}_m \quad (\ell = 0, 1, 2, ..., M_h) \tag{F.257}$$

Noter qu'une manière équivalente de définir cet opérateur de prolongement consiste à dire qu'il s'agit de l'opérateur linéaire de $\mathbb{R}^{M_H}$ dans $\mathbb{R}^{M_h}$ pour lequel

$$P\, S_H^{(m)} = S_h^{(m)} \tag{F.258}$$

quel que soit $m = 1, 2, ..., M_H$. Par conséquent, pour cette interpolation spectrale :

$$\varepsilon_h^{(m)} = 0 \quad (m = 1, 2, ..., M_H) \tag{F.259}$$

On a tracé la courbe de convergence correspondante à la figure F.5 (a) et consigné les résultats de l'expérience d'enrichissement progressif de maillage correspondant à cette interpolation au tableau F.4. On y constate qu'après le premier prolongement, le critère d'arrêt est satisfait avant la moindre itération de Jacobi. Après le deuxième, une seule itération suffit. L'effort de calcul est donc presqu'entièrement concentré sur la grille fine où le résidu est réduit du facteur $.555D-01/.789D-02 \approx 7.03$ (au lieu du facteur théorique de $(H/h)^2 = 4$), ce qui s'effectue en 404 itérations de Jacobi, et réalise une économie en coût par rapport à l'itération monogrille proche de 3.6.

$M+1$	résidu initial	résidu final	itérations effectuées	UT	UTC
4	.987D+01	.436	9	27	27
8	.629D-01	.629D-01	0	0	27
16	.322D-01	.316D-01	1	15	42
32	.555D-01	.789D-02	404	12524	12566

Tableau F.4. Résolution itérative par enrichissement progressif du maillage et prolongement par interpolation spectrale

(4) Lissage :

Les expériences précédentes ont révélé que la technique d'enrichissement progressif de maillage permettait de réaliser une économie en coût de calcul seulement si l'interpolation utilisée pour prolonger la solution d'une grille à la suivante était

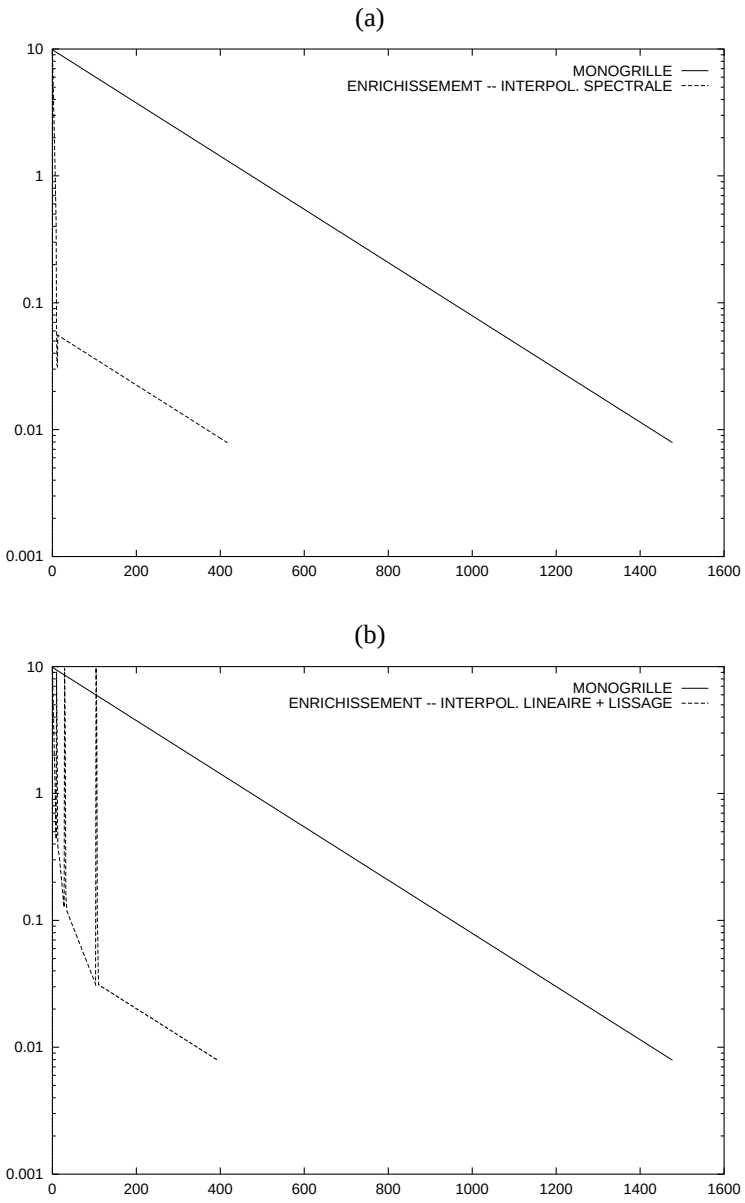


Figure F.5. Comparaison des courbes de convergence du résidu en fonction des itérations de l'algorithme monogrille ($M+1=32$) et de l'algorithme par enrichissement progressif ($M+1=4,8,16,32$) : (a) interpolation spectrale, (b) interpolation linéaire + lissage.

suffisamment précise. De plus, lorsque ce n'est pas le cas, le résidu calculé après interpolation a un contenu de hautes fréquences d'ordre 1. Cette observation conduit à expérimenter comme alternative à une formule d'interpolation très précise, l'application d'une phase de lissage préalablement au démarrage de l'itération de Jacobi précédente.

On a vu à la première question que les erreurs les plus petites qui interviennent sur la grille la plus fine sont de l'ordre de 10^{-3}. On sait d'après les résultats du chapitre 3 que pour ce problème modèle unidimensionnel, un cycle de 3 pseudo-pas de temps de la méthode de Richardson ajustés pour atténuer optimalement les modes de hautes fréquences parvient à les réduire d'un facteur proche de 10^{-2}. Le résidu initial étant de l'ordre de 10, l'application de 2 cycles de ce type devrait ramener les composantes hautes fréquences du résidu à des valeurs de l'ordre de $10 \times (10^{-2})^2 = 10^{-3}$, ce qui est 8 fois moins que la tolérance sur le résidu de grille fine. Il apparaît donc que l'application de ces 2 cycles devrait constituer une phase de lissage amplement suffisante quel que soit le niveau de grille, pour éliminer la contribution des modes de hautes fréquences à la remontée du résidu.

On reprend donc l'expérience d'enrichissement progressif de maillage en faisant suivre le prolongement par interpolation linéaire d'une phase de lissage constituée de 6 itérations de Jacobi correspondant aux 2 cycles définis ci-dessus. L'historique de convergence est représenté à la figure F.5 (b). Par ailleurs, on a consigné les nouveaux résultats au tableau F.5 dans lequel pour $M+1 = 8$, 16, 32, la première ligne correspond à la phase de lissage et la seconde à l'itération de Jacobi sous sa forme usuelle.

$M+1$	résidu initial	résidu final	itérations effectuées	UT	UTC
4	.987D+01	.436	9	27	27
8	.912D+01	.327	6	42	69
8	.327	.127	12	84	153
16	.968D+01	.117	6	90	243
16	.117	.313D-01	68	1020	1263
32	.982D+01	.310D-01	6	186	1449
32	.310D-01	.788D-02	284	8804	10253

Tableau F.5. Résolution itérative par enrichissement progressif du maillage, prolongement par interpolation linéaire, lissage et relaxation

On constate que chaque prolongement fait à nouveau monter le résidu à une valeur proche de 10 ; mais ici on observe que la phase de lissage est suffisante pour réduire le résidu sur la grille fine à une valeur proche de, et même légèrement inférieure à la valeur du résidu à l'issue de l'itération de Jacobi sur la grille précédente. Il en résulte un net gain en efficacité, de l'ordre de 4.5 par rapport à l'itération monogrille, ce qui est satisfaisant lorsqu'on utilise 4 niveaux de grille. En particulier, la contribution principale au coût global correspond aux 284 itérations de Jacobi effectuées sur la grille fine ; il est notoire que cette phase agit pour réduire le résidu du facteur $.310D-01/.788D-02 \approx 3.93$ qui est très proche du facteur théorique de 4, ce qui apporte un certain support expérimental aux estimations théoriques des coûts.

En conclusion, pour une EDP d'ordre β, approchée discrètement à l'ordre α, il apparaît que la technique d'enrichissement progressif de maillage pour être effective doit reposer ou bien sur un prolongement par interpolation d'ordre au moins égal à $\alpha+\beta-1$, ou bien sur l'introduction d'un lisseur adéquat. Dans ce cas, on peut escompter un gain en efficacité de l'ordre du nombre de niveaux de grille utilisés.

Exercice 5.1 (Complexité de la méthode multigrille complète)

Par la méthode multigrille complète, le coût est proportionnel au nombre d'inconnues. Si la deuxième campagne de calculs était initialisée de la même manière que la première, son coût serait donc triple (3C). L'initialisation différente permet l'économie du coût de la première campagne (C). En définitive, le coût de la deuxième campagne seule est de 2C.

Exercice 5.2 (Variante de la méthode bigrille idéale)

(1) La forme du vecteur $s^{(m)}$ indique qu'il s'agit du discrétisé de la fonction :

$$\psi^{(m)}(x) = C\ \sin(m\pi x) \tag{F.260}$$

sur la grille fine $\mathcal{M}_h$. Par injection directe sur la grille grossière $\mathcal{M}_{3h}$, on obtient le discrétisé de cette fonction sur cette nouvelle grille. Ce discrétisé est précisément le vecteur propre de même fréquence spatiale associé à cette grille pour autant que :

$$m \leq M_{3h} = \frac{M_h+1}{3} - 1 \tag{F.261}$$

Cette condition caractérise donc les modes de basses fréquences.

On a :

$$\theta_{M_{3h}+1} = \frac{(M_{3h}+1)\pi}{M_h+1} = \frac{\pi}{3} \tag{F.262}$$

par conséquent :

$$h^2 \lambda_{h\,M_{3h}+1} = 1 \tag{F.263}$$

L'intervalle de variation du paramètre $h^2 \lambda_{h\,m}$ est donc approximativement [0,1] pour les modes de « basses fréquences », et [1,4] pour ceux de « hautes fréquences ».

(2) Le facteur d'atténuation qui résulte de l'application de la méthode de Richardson avec k pseudo-pas de temps est donné par :

$$\rho = 1/A_k \tag{F.264}$$

où :

$$A_k = T_k(c) \tag{F.265}$$

et :

$$c = \frac{b+a}{b-a} \tag{F.266}$$

Ici, pour construire un lisseur, on choisit « d'attaquer » seulement les hautes fréquences, ce qui correspond à :

$$\left\{ \begin{array}{l} a = 1 \\ b = 4 \end{array} \right. \tag{F.267}$$

Par conséquent :

$$c = 5/3 \tag{F.268}$$

La condition sur k est donc :

$$T_k(5/3) \geq 100 \tag{F.269}$$

On peut essayer à tâtons, ou se rappeler que pour $x > 1$,

$$T_k(x) = \text{ch}\ (k\,\text{Argch}\,x) \tag{F.270}$$

La condition équivaut donc à :

$$k \geq \frac{\text{Argch}\,(100)}{\text{Argch}\,(5/3)} \approx 4.822 \tag{F.271}$$

Il faut donc prendre :

$$k \geq 5 \tag{F.272}$$

Les pseudo-pas de temps $\{\tau_\ell\}$ ($\ell = 1, ..., 5$) sont donnés par :

$$h^2\,\tau_\ell^{-1} = \frac{4+1}{2} + \frac{4-1}{2}\,\cos\,\frac{(2\ell-1)\pi}{10} \tag{F.273}$$

et l'atténuation des hautes fréquences effectivement réalisée correspond à :

$$\rho = 1/T_5(5/3) = 1/\mathrm{ch}\,(5\,\mathrm{Argch}\,(5/3)) \approx 1/121.502 \approx 0.00823 \tag{F.274}$$

Exercice 5.3 (Etude d'un problème aux limites anisotrope)

(1) L'injection directe préserve la structure tensorielle « à variables séparées » des modes propres. Par conséquent le phénomène d'*aliasing* apparaît ssi il se manifeste séparément dans la direction de x, ou dans la direction de y, ou les deux. Mais ce n'est jamais le cas dans la direction de x car le maillage grossier est ici basé sur la même discrétisation que le maillage fin dans cette direction. En définitive, le phénomène s'observe pour les modes (que l'on définit comme les modes de HF) pour lesquels la discrétisation en y est insuffisante, c'est-à-dire ceux pour lesquels $\ell > L$.

$$\lambda^h_{m,\ell} = 4\,\sin^2\,\frac{\theta_{x\,m}}{2} + 4\,\sin^2\,\frac{\theta_{y\,\ell}}{2} \tag{F.275}$$

$$\begin{aligned}\lambda^h_{\mathrm{min}} &= 4\,\sin^2\,\frac{\theta_{x\,1}}{2} + 4\,\sin^2\,\frac{\theta_{y_1}}{2}\\ &= 4\,\sin^2\,\frac{\pi}{20} + 4\,\sin^2\,\frac{\pi}{200}\\ &\approx 0.0989\end{aligned} \tag{F.276}$$

$$\begin{aligned}\lambda^h_{\mathrm{max}} &= 4\,\sin^2\,\frac{\theta_{x\,9}}{2} + 4\,\sin^2\,\frac{\theta_{y_{99}}}{2}\\ &= 4\,\cos^2\,\frac{\pi}{20} + 4\,\cos^2\,\frac{\pi}{200}\\ &= 8 - \lambda^h_{\mathrm{min}}\\ &\approx 7.9011\end{aligned} \tag{F.277}$$

$$\begin{aligned}a &= 4\,\sin^2\,\frac{\theta_{x\,1}}{2} + 4\,\sin^2\,\frac{\theta_{y_{10}}}{2}\\ &= 4\,\sin^2\,\frac{\pi}{20} + 4\,\sin^2\,\frac{\pi}{20}\\ &= 2\,\lambda^h_{\mathrm{min}}\\ &\approx 0.1958\end{aligned} \tag{F.278}$$

$$\begin{aligned} b &= \lambda^h_{\max} \\ &\approx 7.9011 \end{aligned} \tag{F.279}$$

En conclusion, l'enveloppe des hautes fréquences $[a, b]$ occupe près de 99%, c'est-à-dire une très grande proportion, du spectre complet $[\lambda^h_{\min}, \lambda^h_{\max}]$, car le maillage grossier, très peu dense par rapport au maillage fin, ne sous-tend qu'un faible nombre de basses fréquences. En conséquence, l'efficacité du cycle bigrille exige que le lisseur soit performant sur une plage de fréquences relativement grande.

(2) L'expression des pseudo-pas de temps est la suivante :

$$\left\{ \begin{array}{l} (\tau^*_\ell)^{-1} = \dfrac{b+a}{2} + \dfrac{b-a}{2}\,\xi_\ell \\ \\ \xi_\ell = \cos\,\frac{(2\ell-1)\,\pi}{2K}\,,\ \ell = 1, 2, ..., k. \end{array} \right. \tag{F.280}$$

Les modes de HF les moins atténués le sont par le facteur $1/A_k$ où :

$$A_k = T_k(c) = \text{ch}\ (k\,\text{Argch}\,c) \tag{F.281}$$

où :

$$c = \frac{b+a}{b-a} \approx 1.0508 \tag{F.282}$$

La condition s'écrit :

$$\text{ch}\ (k\,\text{Argch}\,c) \geq 10 \tag{F.283}$$

soit :

$$k \geq \frac{\text{Argch}\,10}{\text{Argch}\,c} \approx 9.43 \tag{F.284}$$

Il faudrait donc prendre $k = 10$ au moins, ou un autre lisseur.

Exercice 6.1 (Nature mathématique des équations d'Euler stationnaires)

Le système des équations d'Euler stationnaires peut s'écrire sous la forme quasi-linéaire suivante :

$$A(W)\,W_x + B(W)\,W_y = 0 \tag{F.285}$$

L'étude locale s'effectue en gelant les matrices jacobiennes A et B.

En supposant par exemple que la matrice jacobienne A n'est pas singulière, on voit que ce système local est hyperbolique ssi la matrice $A^{-1}\,B$ est diagonalisable sur $\mathbb{R}$. On est donc amené à chercher les solutions μ de l'équation caractéristique suivante :

$$\det\left(B-\mu\,A\right)=0 \tag{F.286}$$

Cette condition équivaut à dire que 0 est une valeur propre de la matrice $k_1\,A+k_2\,B$ où l'on a posé :

$$k_1=-\mu\,,\quad k_2=1 \tag{F.287}$$

Or, les valeurs propres de cette matrice sont fournies par (6.9)-(6.10) :

$$\lambda_1=\lambda_2=-\mu\,u+v \tag{F.288}$$

$$\lambda_{3,4}=-\mu\,u+v\pm\sqrt{\mu^2+1}\,c \tag{F.289}$$

La condition :

$$\lambda_m=0 \tag{F.290}$$

fournit d'abord la solution double :

$$\mu=\frac{v}{u}=\mathrm{tg}\,\theta \tag{F.291}$$

où θ est l'angle polaire de la vitesse avec l'axe des x :

$$u=V\,\cos\,\theta=Mc\,\cos\,\theta \tag{F.292}$$

$$v=V\,\sin\,\theta=Mc\,\sin\,\theta \tag{F.293}$$

($V=\sqrt{u^2+v^2}$). Cette solution est associée au transport des grandeurs par mouvement du fluide, puisque les ondes simples correspondantes ont la forme :

$$\widehat{W}(x,y)=\widehat{W}_0(y-\mu x)=\widehat{W}_0(y-x\,\mathrm{tg}\,\theta) \tag{F.294}$$

et sont associées à des droites caractéristiques portées par le vecteur vitesse. Les deux autres valeurs propres s'obtiennent en résolvant l'équation suivante :

$$(\mu\,u-v)^2=(\mu^2+1)\,c^2 \tag{F.295}$$

qui se développe comme suit :

$$(u^2-c^2)\,\mu^2-2\,uv\,\mu+v^2-c^2=0 \tag{F.296}$$

et dont les solutions sont réelles ssi le discriminant :

$$\delta=u^2v^2-(u^2-c^2)(v^2-c^2)=c^2(u^2+v^2-c^2)=c^4(M^2-1) \tag{F.297}$$

est positif, c'est-à-dire ssi l'écoulement est localement supersonique ($M > 1$), ce qui constitue donc bien la condition d'hyperbolicité du système des équations d'Euler stationnaires.

Plus précisément, lorsque $M > 1$, les deux autres valeurs propres sont les suivantes :

$$\mu = \frac{uv \pm c^2\sqrt{M^2-1}}{u^2-c^2} \tag{F.298}$$

Il est habituel de simplifier cette expression en introduisant l'« angle de Mach » :

$$\alpha = \text{Arctg}\,\frac{1}{\sqrt{M^2-1}} \tag{F.299}$$

de sorte que :

$$\begin{aligned}
\mu &= \frac{M^2\,\cos\,\theta\,\sin\,\theta \pm \sqrt{M^2-1}}{M^2\,\cos\,\theta - 1} \\
&= \frac{\text{tg}\,\theta \pm \dfrac{\sqrt{M^2-1}}{M^2}\,(1+\text{tg}^2\,\theta)}{1-\dfrac{1}{M^2}\,(1+\text{tg}^2\,\theta)} \\
&= \frac{\text{tg}\,\theta \pm \dfrac{\text{tg}\,\alpha}{1+\text{tg}^2\,\alpha}\,(1+\text{tg}^2\,\theta)}{1-\dfrac{\text{tg}^2\,\alpha}{1+\text{tg}^2\,\alpha}\,(1+\text{tg}^2\,\theta)} \\
&= \frac{(\text{tg}\,\theta \pm \text{tg}\,\alpha)\,(1 \pm \text{tg}\,\alpha\,\text{tg}\,\theta)}{(1+\text{tg}\,\alpha\,\text{tg}\,\theta)(1-\text{tg}\,\alpha\,\text{tg}\,\theta)} \\
&= \frac{\text{tg}\,\theta \pm \text{tg}\,\alpha}{1 \mp \text{tg}\,\alpha\,\text{tg}\,\theta} \\
&= \text{tg}\,(\theta \pm \alpha)
\end{aligned} \tag{F.300}$$

Ces valeurs propres sont associées aux « ondes simples acoustiques » suivantes :

$$\widehat{W}(x,y) = \widehat{W}_1\Big(y - x\,\text{tg}\,(\theta \pm \alpha)\Big) \tag{F.301}$$

En définitive, les droites caractéristiques issues d'un point courant (x, y) ont donc pour pentes tg θ (double), tg $(\theta+\alpha)$ et tg $(\theta-\alpha)$; elles délimitent en 2D un secteur angulaire de $\pm\alpha$ de part et d'autre du support de la vitesse, et en 3D un « cône de Mach ». En régime supersonique, les particules matérielles se déplacent plus vite que les ondes acoustiques. L'écoulement au point courant n'est pas modifié par une perturbation (infinitésimale) apportée à l'écoulement en un point extérieur au cône dont ce point est le sommet ; réciproquement, lorsqu'une particule matérielle se déplace, sa présence

n'est pas perçue à l'extérieur du cône. Lorsque le nombre de Mach $M \to 1$, l'angle de Mach $\alpha \to \frac{\pi}{2}$, on approche le régime d'écoulement subsonique, pour lequel les domaines d'influence et de dépendance approchent des demi-plans complémentaires.

Exercice 6.2 (Invariance des équations d'Euler par rotation)

Soit γ l'angle polaire du vecteur $\vec{\eta}$:

$$\eta_x = \| \vec{\eta} \| \cos \gamma \tag{F.302}$$

$$\eta_y = \| \vec{\eta} \| \sin \gamma \tag{F.303}$$

Il vient :

$$\eta_x F(W) + \eta_y G(W) = \| \vec{\eta} \| \begin{pmatrix} \rho U \\ \rho u U + p \cos \gamma \\ \rho v U + p \sin \gamma \\ (E+p) U \end{pmatrix} \tag{F.304}$$

où l'on a posé :

$$U = u \cos \gamma + v \sin \gamma \tag{F.305}$$

Soit $\mathcal{R}_\gamma$ la matrice suivante :

$$\mathcal{R}_\gamma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \gamma & \sin \gamma & 0 \\ 0 & -\sin \gamma & \cos \gamma & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{F.306}$$

de sorte que :

$$\mathcal{R}_\gamma W = \begin{pmatrix} \rho \\ \rho U \\ \rho V \\ E \end{pmatrix} \tag{F.307}$$

où l'on a posé :

$$V = -u \sin \gamma + v \cos \gamma \tag{F.308}$$

Par conséquent :

$$F(\mathcal{R}_\gamma W) = \begin{pmatrix} \rho U \\ \rho U^2 + P \\ \rho U V \\ U(E+P) \end{pmatrix} \tag{F.309}$$

où la pression P s'obtient par l'équation d'état :

$$\begin{aligned} P &= (\gamma - 1)\ \left(E - \frac{U^2 + V^2}{2} \right) \\ &= (\gamma - 1)\ \left(E - \frac{u^2 + v^2}{2} \right) \\ &= p \end{aligned} \tag{F.310}$$

Finalement :

$$\begin{aligned} \mathcal{R}_\gamma^{-1}\, F\,(\mathcal{R}_\gamma\, W) &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\,\gamma & -\sin\,\gamma & 0 \\ 0 & \sin\,\gamma & \cos\,\gamma & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho\, U \\ \rho\, U^2 + p \\ \rho\, U\, V \\ U\,(E+p) \end{pmatrix} \\ &= \begin{pmatrix} \rho\, U \\ \rho u\, U + p\,\cos\,\theta \\ \rho v\, U + p\,\sin\,\theta \\ (E+p)\, U \end{pmatrix} \end{aligned} \tag{F.311}$$

d'où le résultat. □

Exercice 7.1 (Application de l'algorithme de Schwarz)

(1) Les résultats de l'exercice 1.1 fournissent :

$$\| u_h - u \|_\infty \leq \frac{\mu h^2}{96} \tag{F.312}$$

où

$$\mu = \max_{x\,\in\,[0,1]} |f"(x)| = \max_{x\,\in\,[0,1]} \left[16\, x(1-x)\, e^{-2(x-\frac{1}{2})^2} \right] = 4 \tag{F.313}$$

(le maximum étant atteint pour $x = \frac{1}{2}$). Il vient :

$$\| u_h - u \|_\infty \leq \frac{4}{96}(\frac{1}{10})^2 \approx 4.17 \times 10^{-4} \tag{F.314}$$

(2) En vertu de (7.25), la condition équivaut à :

$$\rho^n \leq \varepsilon \tag{F.315}$$

où :

$$\rho = \left(\frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta} \right)^2 \tag{F.316}$$

où δ représente la demi-largeur du recouvrement ; ici $\delta = \frac{1}{10}$. Il vient :

$$n \geq \frac{-\log \varepsilon}{-\log \rho} \approx \frac{-\log 4.17 \times 10^{-4}}{-2 \log \frac{4}{6}} \approx 9.6 \tag{F.317}$$

Il faut donc une dizaine d'itérations environ.

Exercice 7.2 (Advection pure et conditions aux limites)

(1) Les dérivées partielles de la fonction proposée sont les suivantes :

$$u_x = -a \frac{\partial u_0}{\partial x}(x - at, y - bt) \tag{F.318}$$

$$u_y = -b \frac{\partial u_0}{\partial x}(x - at, y - bt) \tag{F.319}$$

En reportant ces expressions dans l'EDP on vérifie qu'elle est bien satisfaite ainsi que la condition initiale. □

Par conséquent, l'advection pure a pour effet de translater la condition intiale à la vitesse constante $\vec{V}$. Les droites de vecteur directeur $\vec{V}$ sont nommées « caractéristiques ». Elles indiquent la provenance de l'information à partir de laquelle on peut calculer l'inconnue (voir figure F.6).

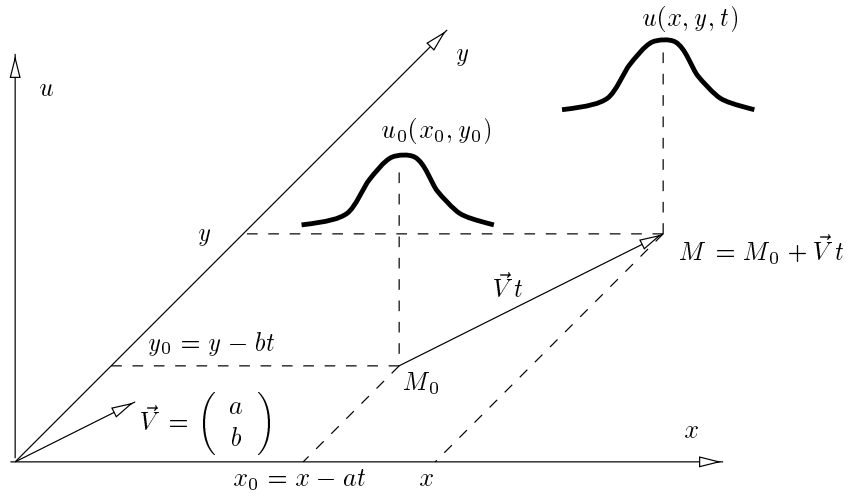


Figure F.6. Schéma de translation de la solution par advection pure

(2) D'après les résultats de la question précédente, on détermine l'inconnue $u(x, y, t)$

en traçant la caractéristique issue du point $M(x, y)$ qui intersecte la frontière du domaine en un point $M_0(x_0, y_0)$ tel que :

$$x_0 = x - a\tau \tag{F.320}$$

$$y_0 = y - b\tau \tag{F.321}$$

pour un certain $\tau > 0$ de sorte que :

$$M_0 \in \Gamma^- \tag{F.322}$$

(voir figure F.7). Dans ce cas, on a :

$$u(x, y, t) = u(M, t) = u(M_0, t - \tau) \tag{F.323}$$

C'est donc sur Γ^- qu'il convient de spécifier la condition de Dirichlet. Il serait impropre de le faire sur Γ^+.

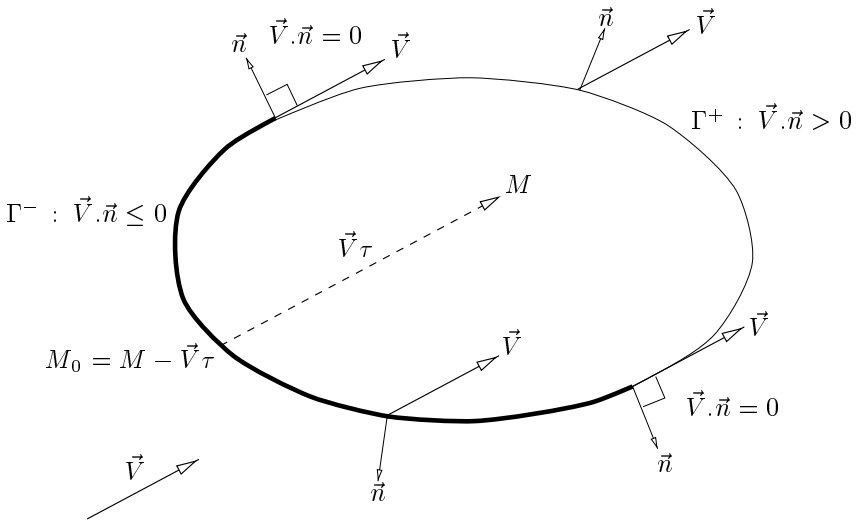


Figure F.7. Schéma d'advection pure en domaine borné

Exercice 7.3 (Algorithme de Dirichlet-Neumann)

(1) On obtient successivement :

$$\frac{u_{xx}}{u_x} = r \tag{F.324}$$

$$\ln |u_x| = rx + c_1 \tag{F.325}$$

$$u_x = c_2\, e^{rx} \quad (c_2 = e^{c_1}) \tag{F.326}$$

$$u(x) = A\, e^{rx} + B \quad (A = c_2/r) \tag{F.327}$$

Les conditions aux limites donnent :

$$\begin{cases} u(0) = 1 = A + B \\ u(1) = 0 = A\, e^{r} + B \end{cases} \tag{F.328}$$

d'où A et B puis :

$$u(x) = \frac{1 - e^{r(x-1)}}{1 - e^{-r}} \tag{F.329}$$

et :

$$u'(1) = \frac{-r}{1 - e^{-r}} \approx -r \tag{F.330}$$

Si r est grand, il convient donc de densifier le maillage près du bord $x = 1$.

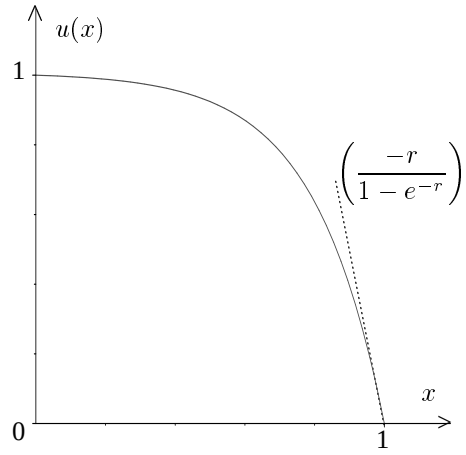


Figure F.8. Profil de « couche limite » solution de l'équation d'advection-diffusion ; courbe tracée pour $r = 5$; cf. exercice 7.3

(2) La résolution du sous-problème associé au sous-domaine Ω_1 donne :

$$u(x) = A_1\, e^{rx} + B_1 \quad (0 \leq x \leq \delta) \tag{F.331}$$

où les constantes A_1 et B_1 sont soumises aux conditions suivantes :

$$\begin{cases} u(0) = 1 = A_1 + B_1 \\ u_x(\delta) = p = A_1\, r\, e^{r\delta} \end{cases} \tag{F.332}$$

d'où :

$$A_1 = \frac{p}{r}\, e^{-r\delta}\ ,\ B_1 = 1 - \frac{p}{r}\, e^{-r\delta} \tag{F.333}$$

ce qui fournit :

$$v' = u(\delta) = A_1\, e^{r\delta} + B_1 = \frac{p}{r} + 1 - \frac{p}{r}\, e^{-r\delta} = g_{12}\, p + 1 \tag{F.334}$$

où :

$$g_{12} = \frac{1 - e^{-r\delta}}{r} \tag{F.335}$$

Symétriquement, la résolution du sous-problème associé au sous-domaine Ω_2 donne :

$$u(x) = A_2\, e^{rx} + B_2 \quad (\delta \leq x \leq 1) \tag{F.336}$$

où les constantes A_2 et B_2 sont soumises aux conditions suivantes :

$$\left\{ \begin{array}{l} u(\delta) = v = A_2\, e^{r\delta} + B_2 \\ u(1) = 0 = A_2\, e^{r} + B_2 \end{array} \right. \tag{F.337}$$

d'où :

$$A_2 = \frac{-v}{e^r - e^{r\delta}}\,,\ B_2 = \frac{v\, e^r}{e^r - e^{r\delta}} \tag{F.338}$$

ce qui fournit :

$$p' = u_x(\delta) = A_2\, r\, e^{r\delta} = g_{21}\, v \tag{F.339}$$

où :

$$g_{21} = \frac{-re^{r\delta}}{e^r - e^{r\delta}} \tag{F.340}$$

Par conséquent l'itération en (v, p) a bien la forme donnée dans l'énoncé. Les valeurs propres de la matrice G sont les racines de l'équation :

$$\lambda^2 = g_{12}\, g_{21} = -\frac{\left(1 - e^{-r\delta}\right)\, e^{r\delta}}{e^r - e^{r\delta}} = -\frac{1 - e^{-r\delta}}{e^{r(1-\delta)} - 1} < 0 \tag{F.341}$$

Les valeurs propres sont donc les nombres complexes conjugués $\pm i\rho$, où ρ, le rayon spectral, est donné par :

$$\rho = \sqrt{\frac{1 - e^{-r\delta}}{e^{r(1-\delta)} - 1}} \approx e^{\frac{-r(1-\delta)}{2}} \ll 1 \tag{F.342}$$

pour $r \gg 1$. Par conséquent l'itération converge (et même très rapidement).

C'est sur le domaine Ω_2 où la diffusion domine qu'il convient d'utiliser un maillage fin. Par contre sur Ω_1 où la convection domine le maillage peut être moins dense.

L'algorithme proposé permet notamment de coordonner deux discrétisations correspondant à des densités de points très différentes sans avoir à densifier la discrétisation uniformément.

(3) Dans l'algorithme multiplicatif, on a encore :

$$v' = g_{12}\, p + b_1 \quad (b_1 = 1) \tag{F.343}$$

mais désormais :

$$p' = g_{21}\, v' + b_2 = g_{21} g_{12}\, p + b'_2 \tag{F.344}$$

Par conséquent la matrice G est remplacée par la suivante :

$$G' = \begin{pmatrix} 0 & g_{12} \\ 0 & g_{12} g_{21} \end{pmatrix} \tag{F.345}$$

Le rayon spectral devient :

$$\rho' = |g_{12} g_{21}| = \rho^2 \tag{F.346}$$

c'est-à-dire le carré du précédent.

Cet algorithme de coordination converge deux fois plus vite mais il ne permet pas une résolution simultanée "parallèle" des sous-problèmes.

(4) Les conditions portant sur les constantes A_1 et B_1 sont ici :

$$\begin{cases} A_1 + B_1 = 1 \\ A_1\, e^{r\delta} + B_1 = v \end{cases} \tag{F.347}$$

d'où :

$$A_1 = \frac{v-1}{e^{r\delta} - 1} \tag{F.348}$$

et :

$$p' = A_1\, r\, e^{r\delta} = \frac{(v-1)\, r\, e^{r\delta}}{e^{r\delta} - 1} = g_{21}\, v + b_1 \tag{F.349}$$

où ici :

$$g_{21} = \frac{r\, e^{r\delta}}{e^{r\delta} - 1} = \frac{r}{1 - e^{-r\delta}} \tag{F.350}$$

Les conditions portant sur les constantes A_2 et B_2 sont ici :

$$\begin{cases} A_2\, r\, e^{r\delta} = p \\ A_2\, e^{r} + B_2 = 0 \end{cases} \tag{F.351}$$

d'où :

$$A_2 = \frac{p}{r}\, e^{-r\delta}\ ,\ B_2 = -A_2\, e^{r} = -\frac{p}{r}\, e^{r(1-\delta)} \tag{F.352}$$

et :

$$v' = A_2\, e^{r\delta} + B_2 = \frac{p}{r} - \frac{p}{r}\, e^{r(1-\delta)} = g_{12}\, p \tag{F.353}$$

où :

$$g_{12} = \frac{1 - e^{r(1-\delta)}}{r} \tag{F.354}$$

On peut à nouveau définir une matrice G ayant la même structure et un vecteur constant $\vec{b}$ mais les valeurs des coefficients ont changé. Par conséquent le rayon spectral ρ'' est ici donné par :

$$\rho'' = \sqrt{|g_{12} g_{21}|} = \sqrt{\frac{e^{r(1-\delta)} - 1}{1 - e^{-r\delta}}} \approx e^{+\frac{r(1-\delta)}{2}} \gg 1 \tag{F.355}$$

(On pouvait également obtenir ces résultats par le changement de variable $x' = 1 - x$ et certaines substitutions de symboles.)

En conclusion $\rho'' \gg 1$ (car $r \gg 1$) ce qui implique que cet algorithme de coordination diverge. En effet, dans ce problème à convection dominante, l'information se propage principalement dans le sens des x croissants (car $c > 0$) et la condition de Dirichlet à droite de Ω_1 ainsi que la condition de Neumann à gauche de Ω_2 sont impropres.

Exercice 7.4 (Contrôle de la continuité par la dérivée normale)

Les sous-problèmes non linéaires se formulent respectivement comme suit :

sur Ω_1 :

$$\begin{gathered} \alpha u^{(1)} + a u^{(1)} u_x^{(1)} + b u^{(1)} u_y^{(1)} - \varepsilon (u_{xx}^{(1)} + u_{yy}^{(1)}) = f\,, \\ (x,y) \in]-1,0[\times]-1,1[\end{gathered}$$

$$\begin{aligned} &u^{(1)}(-1,y) = g(y)\,,\ \forall y \in]-1,1[\\ &u^{(1)}(x,-1) = h(x)\,,\ \forall x \in]-1,0[\\ &u_y^{(1)}(x,1) = 0\,,\ \forall x \in]-1,0[\\ &u_x^{(1)}(0,y) = v(y)\,,\ \forall y \in]-1,1[\end{aligned} \tag{F.356}$$

et sur Ω_2 :

$$\begin{gathered}\alpha u^{(2)} + a u^{(2)} u_x^{(2)} + b u^{(2)} u_y^{(2)} - \varepsilon (u_{xx}^{(2)} + u_{yy}^{(2)}) = f\,, \\ (x,y) \in]0,1[\times]-1,1[\end{gathered}$$

$$\begin{aligned} & u_x^{(2)}(0,y) = v(y)\,,\ \forall y \in]-1,1[\\ & u^{(2)}(x,-1) = h(x)\,,\ \forall x \in]0,1[\\ & u_y^{(2)}(x,1) = 0\,,\ \forall x \in]0,1[\\ & u_x^{(2)}(1,y) = 0\,,\ \forall y \in]-1,1[\end{aligned} \tag{F.357}$$

Dans cet exemple, le sous-problème sur Ω_1 est de type Dirichlet-Neumann en x et en y, alors que le sous-problème sur Ω_2 est de type Neumann-Neumann en x et Dirichlet-Neumann en y.

Fonctionnelle de coût : on définit le « critère » suivant :

$$J(v) = \frac{1}{2} \int_{-1}^{1} \left(u^{(1)} - u^{(2)} \right)^2 (0,y)\, \omega(y)\, dy \tag{F.358}$$

qui mesure la violation de continuité à l'interface.

Sous-problèmes linéarisés :

sur Ω_1 :

$$\begin{gathered}\left(\alpha + a u_x^{(1)} + b u_y^{(1)}\right) \delta u^{(1)} + a u^{(1)} \delta u_x^{(1)} + b u^{(1)} \delta u_y^{(1)} \\ = \varepsilon (\delta u_{xx}^{(1)} + \delta u_{yy}^{(1)})\,,\ (x,y) \in]-1,0[\times]-1,1[\end{gathered}$$

$$\begin{aligned} & \delta u^{(1)}(-1,y) = 0\,,\ \forall y \in]-1,1[\\ & \delta u^{(1)}(x,-1) = 0\,,\ \forall x \in]-1,0[\\ & \delta u_y^{(1)}(x,1) = 0\,,\ \forall x \in]-1,0[\\ & \delta u_x^{(1)}(0,y) = \delta v(y)\,,\ \forall y \in]-1,1[\end{aligned} \tag{F.359}$$

et sur Ω_2 :

$$\begin{gathered}\left(\alpha + a u_x^{(2)} + b u_y^{(2)}\right) \delta u^{(2)} + a u^{(2)} \delta u_x^{(2)} + b u^{(2)} \delta u_y^{(2)} \\ = \varepsilon (\delta u_{xx}^{(2)} + \delta u_{yy}^{(2)})\,,\ (x,y) \in]0,1[\times]-1,1[\end{gathered}$$

$$\begin{aligned} & \delta u_x^{(2)}(0,y) = \delta v(y)\,,\ \forall y \in]-1,1[\\ & \delta u^{(2)}(x,-1) = 0\,,\ \forall x \in]0,1[\\ & \delta u_y^{(2)}(x,1) = 0\,,\ \forall x \in]0,1[\\ & \delta u_x^{(2)}(1,y) = 0\,,\ \forall y \in]-1,1[\end{aligned} \tag{F.360}$$

Première variation de la fonctionnelle coût :

$$\delta J = \int_{-1}^{1} \left(u^{(1)} - u^{(2)}\right) \left(\delta u^{(1)} - \delta u^{(2)}\right) (0,y)\, \omega(y)\, dy \tag{F.361}$$

Equation adjointe sur Ω_1 :

$$\begin{array}{l} (\alpha + a u_x + b u_y)\, \lambda^{(1)} - \left(a u \lambda^{(1)}\right)_x - \left(b u \lambda^{(1)}\right)_y \\ \quad = \varepsilon(\lambda^{(1)}_{xx} + \lambda^{(1)}_{yy})\,,\ (x,y) \in]-1,0[\times]-1,1[\\ \\ \lambda^{(1)}(-1,y) = 0\,,\ \forall y \in]-1,1[\\ \lambda(x,-1) = 0\,,\ \forall x \in]-1,0[\\ \left[b u^{(1)} \lambda^{(1)} + \varepsilon \lambda^{(1)}_y\right](x,1) = 0\,,\ \forall x \in]-1,0[\\ \left[a u^{(1)} \lambda^{(1)} + \varepsilon \lambda^{(1)}_x\right](0,y) = \varepsilon \left(u^{(1)} - u^{(2)}\right)(0,y)\, \omega(y)\,, \\ \qquad\qquad \forall y \in]-1,1[\end{array} \tag{F.362}$$

ce qui fournit le résultat partiel suivant :

$$\int_{-1}^{1} \left(u^{(1)} - u^{(2)}\right) \delta u^{(1)}(0,y)\, \omega(y)\, dy = \int_{-1}^{1} \lambda^{(1)}(0,y)\, \delta v(y)\, dy \tag{F.363}$$

Equation adjointe sur Ω_2 :

$$\begin{array}{l} (\alpha + a u_x + b u_y)\, \lambda^{(2)} - \left(a u \lambda^{(2)}\right)_x - \left(b u \lambda^{(2)}\right)_y \\ \quad = \varepsilon(\lambda^{(2)}_{xx} + \lambda^{(2)}_{yy})\,,\ (x,y) \in]0,1[\times]-1,1[\\ \\ \left[a u^{(2)} \lambda^{(2)} + \varepsilon \lambda^{(2)}_x\right](1,y) = 0\,,\ \forall y \in]-1,1[\\ \lambda^{(2)}(x,-1) = 0\,,\ \forall x \in]0,1[\\ \left[b u^{(2)} \lambda^{(2)} + \varepsilon \lambda^{(2)}_y\right](x,1) = 0\,,\ \forall x \in]0,1[\\ \lambda^{(2)}_x(0,y) = \left(u^{(1)} - u^{(2)}\right)(0,y)\, \omega(y)\,,\ \forall y \in]-1,1[\end{array} \tag{F.364}$$

ce qui fournit le résultat partiel suivant :

$$\int_{-1}^{1} \left(u^{(1)} - u^{(2)}\right) \delta u^{(2)}(0,y)\, \omega(y)\, dy = \int_{-1}^{1} \lambda^{(2)}(0,y)\, \delta v(y)\, dy \tag{F.365}$$

Finalement, en combinant les résultats partiels (F.363) et (F.365), on aboutit au gradient recherché :

$$\delta J = \int_{-1}^{1} K(y)\, \delta v(y)\, dy \tag{F.366}$$

où ici on a posé :

$$K(y) = \left(\lambda^{(1)} - \lambda^{(2)}\right)(0, y) \tag{F.367}$$

Exercice 7.5 (Moindres carrés)

On reprend et on adapte les calculs de la section 7.4 en tenant compte des définitions légèrement modifiées des domaines Ω_1 et Ω_2 qui désormais se recouvrent partiellement :

$$\Omega_1 = [-1, \delta] \times [-1, 1] \tag{F.368}$$
$$\Omega_2 = [-\delta, 1] \times [-1, 1] \tag{F.369}$$

En particulier, les sous-problèmes non linéaires se formulent ici respectivement comme suit :

sur Ω_1 :

$$\begin{aligned} &\alpha u^{(1)} + a u^{(1)} u_x^{(1)} + b u^{(1)} u_y^{(1)} - \varepsilon(u_{xx}^{(1)} + u_{yy}^{(1)}) = f\,, \\ &\qquad (x, y) \in]-1, \delta[\times]-1, 1[\\ &\\ &u^{(1)}(-1, y) = g(y)\,,\ \forall y \in]-1, 1[\\ &u_y^{(1)}(x, -1) = u_y^{(1)}(x, 1) = 0\,,\ \forall x \in]-1, \delta[\\ &u^{(1)}(\delta, y) = v_1(y)\,,\ \forall y \in]-1, 1[\end{aligned} \tag{F.370}$$

et sur Ω_2 :

$$\begin{aligned} &\alpha u^{(2)} + a u^{(2)} u_x^{(2)} + b u^{(2)} u_y^{(2)} - \varepsilon(u_{xx}^{(2)} + u_{yy}^{(2)}) = f\,, \\ &\qquad (x, y) \in]-\delta, 1[\times]-1, 1[\\ &\\ &u^{(2)}(-\delta, y) = v_2(y)\,,\ \forall y \in]-1, 1[\\ &u_y^{(2)}(x, -1) = u_y^{(2)}(x, 1) = 0\,,\ \forall x \in]-\delta, 1[\\ &u_x^{(2)}(1, y) = 0\,,\ \forall y \in]-1, 1[\end{aligned} \tag{F.371}$$

Les sous-problèmes linéarisés suivants en résultent :

sur Ω_1 :

$$\begin{aligned} &\left(\alpha + a u_x^{(1)} + b u_y^{(1)}\right) \delta u^{(1)} + a u^{(1)} \delta u_x^{(1)} + b u^{(1)} \delta u_y^{(1)} \\ &\quad = \varepsilon\left(\delta u_{xx}^{(1)} + \delta u_{yy}^{(1)}\right), (x, y) \in]-1, \delta[\times]-1, 1[\\ &\\ &\delta u^{(1)}(-1, y) = 0\,,\ \forall y \in]-1, 1[\\ &\delta u_y^{(1)}(x, -1) = \delta u_y^{(1)}(x, 1) = 0\,,\ \forall x \in]-1, \delta[\\ &\delta u^{(1)}(\delta, y) = \delta v_1(y)\,,\ \forall y \in]-1, 1[\end{aligned} \tag{F.372}$$

et sur Ω_2 :

$$\begin{array}{l}\left(\alpha + a u_x^{(2)} + b u_y^{(2)}\right) \delta u^{(2)} + a u^{(2)} \delta u_x^{(2)} + b u^{(2)} \delta u_y^{(2)} \\ \quad = \varepsilon \left(\delta u_{xx}^{(2)} + \delta u_{yy}^{(2)}\right) , (x,y) \in]-\delta, 1[\times]-1, 1[\\ \\ \delta u^{(2)}(-\delta, y) = \delta v_2(y) \, , \; \forall y \in]-1, 1[\\ \delta u_y^{(2)}(x, -1) = \delta u_y^{(2)}(x, 1) = 0 \, , \; \forall x \in]-\delta, 1[\\ \delta u_x^{(2)}(1, y) = 0 \, , \; \forall y \in]-1, 1[\end{array} \tag{F.373}$$

En conséquence, la première variation de la fonctionnelle de coût prend la forme suivante :

$$\delta J = \iint_{\Omega_1 \cap \Omega_2} \left(u^{(1)} - u^{(2)}\right) \left(\delta u^{(1)} - \delta u^{(2)}\right) \omega(x,y) \, dx \, dy \tag{F.374}$$

On peut supposer sans perte de généralité que la fonction de pondération $\omega(x,y)$ admet le domaine $\Omega_1 \cap \Omega_2$ comme support :

$$\forall \, (x,y) \notin \Omega_1 \cap \Omega_2 \; : \; \omega(x,y) = 0 \tag{F.375}$$

On définit des variables adjointes, $\lambda^{(1)}$ sur Ω_1 et $\lambda^{(2)}$ sur Ω_2. On aboutit à l'équation suivante analogue de (7.140) :

$$\begin{array}{c} \displaystyle\iint_{\Omega_1} \left[(\alpha + a u_x + b u_y) \, \lambda - (a u \lambda)_x - (b u \lambda)_y - \varepsilon \left(\lambda_{xx} + \lambda_{yy}\right)\right] \delta u \\ \displaystyle + \int_{-1}^{1} \left(a u \lambda + \varepsilon \lambda_x\right)(\delta, y) \, \delta v_1(y) \, dy \\ \displaystyle + \int_{-1}^{\delta} \left[\left(b u \lambda + \varepsilon \lambda_y\right) \delta u(x,1) - \left(b u \lambda + \varepsilon \lambda_y\right) \delta u(x,-1)\right] \, dx \\ \displaystyle + \int_{-1}^{1} \left[-\varepsilon \lambda \delta u_x(\delta, y) + \varepsilon \lambda \delta u_x(-1, y)\right] \, dy \\ = 0 \, , \; \forall \, \lambda \end{array} \tag{F.376}$$

On est donc amené à poser l'équation adjointe suivante :

$$\begin{array}{l} (\alpha + a u_x + b u_y) \, \lambda^{(1)} - \left(a u \lambda^{(1)}\right)_x - \left(b u \lambda^{(1)}\right)_y - \varepsilon \left(\lambda_{xx}^{(1)} + \lambda_{yy}^{(1)}\right) \\ \quad = \left(u^{(2)} - u^{(1)}\right) \omega(x,y) \, , (x,y) \in]-1, \delta[\times]-1, 1[\\ \\ \lambda^{(1)}(\delta, y) = \lambda^{(1)}(-1, y) = 0 \, , \; \forall y \in]-1, 1[\\ \left[b u^{(1)} \lambda^{(1)} + \varepsilon \lambda_y^{(1)}\right](x, -1) = 0 \, , \; \forall x \in]-1, 0[\\ \left[b u^{(1)} \lambda^{(1)} + \varepsilon \lambda_y^{(1)}\right](x, 1) = 0 \, , \; \forall x \in]-1, 0[\end{array} \tag{F.377}$$

(où le terme source est nul à l'extérieur du support $\Omega_1 \cap \Omega_2$ de la fonction $\omega(x,y)$), ce qui fournit le résultat partiel suivant :

$$\iint_{\Omega_1 \cap \Omega_2} \left(u^{(1)} - u^{(2)}\right) \delta u^{(1)}\, \omega(x,y)\, dx\, dy = \int_{-1}^{1} \kappa_1(y)\, \delta v_1(y)\, dy \qquad \text{(F.378)}$$

où l'on a posé :

$$\kappa_1(y) = \left(a u \lambda^{(1)} + \varepsilon\, \lambda_x^{(1)}\right)(\delta, y) \qquad \text{(F.379)}$$

Symétriquement, l'équation suivante représente l'analogue de (7.147) :

$$\begin{aligned} \iint_{\Omega_2} \left[(\alpha + a u_x + b u_y)\, \lambda - (a u \lambda)_x - (b u \lambda)_y - \varepsilon\, (\lambda_{xx} + \lambda_{yy})\right] \delta u \\ - \int_{-1}^{1} (a u \lambda + \varepsilon \lambda_x)\, (-\delta, y)\, \delta v_2(y)\, dy \\ + \int_{-\delta}^{1} \left[(b u \lambda + \varepsilon \lambda_y)\, \delta u(x,1) - (b u \lambda + \varepsilon \lambda_y)\, \delta u(x,-1)\right] dx \\ + \int_{-1}^{1} \left[(a u \lambda + \varepsilon \lambda_x)\, \delta u(1,y) + \varepsilon \lambda \delta u_x(-\delta, y)\right] dy \\ = 0\,,\ \forall\, \lambda \end{aligned} \qquad \text{(F.380)}$$

ce qui conduit à poser l'équation adjointe suivante :

$$\begin{aligned} (\alpha + a u_x + b u_y)\, \lambda^{(2)} - \left(a u \lambda^{(2)}\right)_x - \left(b u \lambda^{(2)}\right)_y - \varepsilon \left(\lambda_{xx}^{(2)} + \lambda_{yy}^{(2)}\right) \\ = \left(u^{(2)} - u^{(1)}\right) \omega(x,y)\,,\ (x,y) \in]0,1[\times]-1,1[\end{aligned}$$

$$\begin{aligned} &\left[a u^{(2)} \lambda^{(2)} + \varepsilon \lambda_x^{(2)}\right](1,y) = 0\,,\ \forall y \in]-1,1[\\ &\left[b u^{(2)} \lambda^{(2)} + \varepsilon \lambda_y^{(2)}\right](x,-1) = 0\,,\ \forall x \in]0,1[\\ &\left[b u^{(2)} \lambda^{(2)} + \varepsilon \lambda_y^{(2)}\right](x,1) = 0\,,\ \forall x \in]0,1[\\ &\lambda^{(2)}(-\delta, y) = 0\,,\ \forall y \in]-1,1[\end{aligned} \qquad \text{(F.381)}$$

ce qui fournit le résultat partiel suivant :

$$-\iint_{\Omega_1 \cap \Omega_2} \left(u^{(1)} - u^{(2)}\right) \delta u^{(2)}\, \omega(x,y)\, dx\, dy = \int_{-1}^{1} \kappa_2(y)\, \delta v_2(y)\, dy \qquad \text{(F.382)}$$

où l'on a posé :

$$\kappa_2(y) = \left(a u \lambda^{(2)} + \varepsilon\, \lambda_x^{(2)}\right)(-\delta, y) \qquad \text{(F.383)}$$

Finalement, en combinant (F.378) et (F.382), on aboutit au gradient recherché :

$$\delta J = \int_{-1}^{1} \left(\kappa_1(y)\, \delta v_1(y) + \kappa_2(y)\, \delta v_2(y)\right) dy \qquad \text{(F.384)}$$

Exercice A.1 (Majoration de la norme-p par la norme infinie)

Par exemple, $u = (1, 1, ..., 1)^T$.

Exercice A.2 (Majoration de la norme infinie par la norme-p)

Par exemple, $u = (1, 0, ..., 0)^T$.

Exercice A.3 (Identification de sphères et de boules)

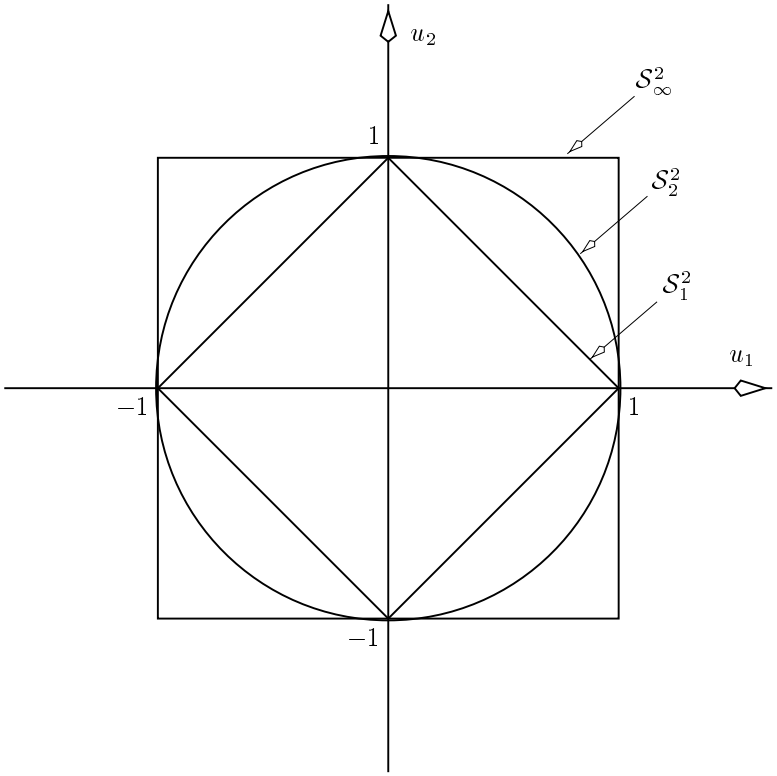


Figure F.9. Sphères (cercles) et boules (disques) de $\mathbb{R}^2$ (cf. exercice A.3)

Les boules $\mathcal{B}^2_p$ ($p = 1, 2$ et ∞) correspondent aux domaines intérieurs (bords compris) aux sphères $\mathcal{S}^2_p$ de même indices p (voir figure F.9).

Exercice A.4 (Norme euclidienne d'une matrice)

$$\begin{aligned}
\text{Trace}\,(A^*\,A) &= \sum_{k=1}^{M} (A^*\,A)_{k,k} \\
&= \sum_{k=1}^{M} \sum_{j=1}^{M} (A^*)_{k,j}\; A_{j,k} \\
&= \sum_{k=1}^{M} \sum_{j=1}^{M} \overline{a_{j,k}}\, a_{j,k} \\
&= \sum_{k=1}^{M} \sum_{j=1}^{M} |a_{j,k}|^2 \\
&= (\| A \|_E)^2
\end{aligned} \tag{F.385}$$

Exercice A.5 (Norme induite d'un produit de matrices)

Quels que soient les vecteurs non nuls u, v et w, on a :

$$\frac{\| A\,u \|}{\|u\|} \leq \| A \|\,, \quad \frac{\| B\,v \|}{\|v\|} \leq \| B \|\,, \quad \frac{\| A\,B\,w \|}{\|w\|} \leq \| A\,B \| \tag{F.386}$$

Par conséquent, w étant un vecteur non nul quelconque, ou bien $B\,w \neq 0$ et

$$\begin{aligned}
\frac{\| A\,B\,w \|}{\|w\|} &= \frac{\| A\,B\,w \|}{\|B\,w\|} \frac{\| B\,w \|}{\|w\|} \\
&\leq \| A \|\, \| B \|
\end{aligned} \tag{F.387}$$

ou bien $B\,w = 0$ et le résultat est vrai a fortiori. Par conséquent, toute valeur particulière de ce quotient satisfait cette majoration, et le maximum absolu du quotient, c'est-à-dire $\| A\,B \|$, aussi.

Exercice A.6 (Norme induite et rayon spectral d'une matrice)

Pour tout vecteur non nul u posons :

$$q(u) = \frac{\| A\,u \|}{\|u\|} \tag{F.388}$$

Soit λ_m une valeur propre de la matrice A et u_m un vecteur propre associé non nul ; on a :

$$q(u_m) = |\lambda_m| \leq \max_{u \neq 0} q(u) = \| A \| \tag{F.389}$$

Par conséquent, le rayon spectral,

$$\rho(A) = \max_m \, |\lambda_m| \tag{F.390}$$

vérifie également cette majoration :

$$\rho(A) \leq \| \, A \, \| \tag{F.391}$$

Les matrices diagonales fournissent un exemple de cas où il y a égalité, mais cette propriété ne leur est pas exclusive. Par exemple, pour

$$A = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix} \tag{F.392}$$

on a :

$$\| \, A \, \|_\infty = \rho(A) \quad (= 2) \tag{F.393}$$

Exercice A.7 (Norme 2 induite et norme euclidienne d'une matrice)

Les valeurs propres $\{\lambda_m\}$ ($m = 1, 2, ..., M$) de la matrice hermitienne semi-définie positive $A^* \, A$ sont des réels positifs (ou nuls), et l'on a :

$$\text{Trace} \, (A^* \, A) = \sum_m \, \lambda_m \tag{F.394}$$

$$\rho \, (A^* \, A) = \max_m \, |\lambda_m| = \max_m \, \lambda_m \tag{F.395}$$

Il est donc évident que :

$$\text{Trace} \, (A^* \, A) \leq M \, \rho \, (A^* \, A) \tag{F.396}$$

$$\rho \, (A^* \, A) \leq \text{Trace} \, (A^* \, A) \tag{F.397}$$

d'où le résultat.

Exercice A.8 (Propriétés de l'application $p \rightarrow \|A\|_p$)

(1) On pose :

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_M \end{pmatrix} \, , \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_M \end{pmatrix} \tag{F.398}$$

on calcule :

$$A = u \, v^T = \begin{pmatrix} u_1 \, v_1 & u_1 \, v_2 & \dots & u_1 \, v_M \\ u_2 \, v_1 & u_2 \, v_2 & \dots & u_2 \, v_M \\ \vdots & \vdots & \dots & \vdots \\ u_M \, v_1 & u_M \, v_2 & \dots & u_M \, v_M \end{pmatrix} \tag{F.399}$$

Il apparaît alors que :

$$\| A \|_1 = \max_k \left| v_k \sum_j |u_j| \right| = \|u\|_1 \, \|v\|_\infty \tag{F.400}$$

Par ailleurs :

$$\| A \|_2 = \sqrt{\rho\,(A^* A)} = \sqrt{\rho \left(v \underbrace{u^T u}_{(\|u\|_2)^2} v^T \right)} = \|u\|_2 \sqrt{\rho(v\, v^T)} \tag{F.401}$$

Pour évaluer ce rayon spectral, considérons une valeur propre quelconque λ de la matrice $v\, v^T$ et un vecteur propre associé x non nul :

$$v\, v^T x = \lambda\, x \tag{F.402}$$

Ou bien $\lambda = 0$, ou bien le produit scalaire $v^T x \neq 0$; dans ce dernier cas, multipliant l'équation précédente par v^T, il vient après simplification par $v^T x \neq 0$:

$$\lambda = v^T v = (\|v\|_2)^2 \tag{F.403}$$

Par conséquent, toute valeur propre non nulle de la matrice $v\, v^T$ est égale à $(\|v\|_2)^2$. D'autre part, la trace d'une matrice étant un invariant par changement de base :

$$\text{Trace}\left(v\, v^T\right) = \sum_j v_j^2 = (\|v\|_2)^2 = \sum_m \lambda_m \tag{F.404}$$

Le spectre est donc constitué de $(\|v\|_2)^2$ (valeur propre simple) et 0 de multiplicité $M - 1$. D'où le rayon spectral,

$$\rho(v\, v^T) = (\|v\|_2)^2 \tag{F.405}$$

et le résultat cherché :

$$\| A \|_2 = \|u\|_2 \, \|v\|_2 \tag{F.406}$$

Enfin, il apparaît que :

$$\| A \|_\infty = \max_j \left| u_j \sum_k |v_k| \right| = \|v\|_1 \, \|u\|_\infty \tag{F.407}$$

(2) L'application $p \to \| A \|_p$ n'est pas monotone. En effet, dans le cas particulier :

$$u = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad v = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{F.408}$$

on a :

$$\|v\|_1 = \|v\|_2 = \|v\|_\infty = 1 \tag{F.409}$$

et :

$$\| A \|_1 = \|u\|_1 = M > \| A \|_2 = \|u\|_2 = \sqrt{M} > \| A \|_\infty = \|u\|_\infty = 1 \tag{F.410}$$

alors que dans le cas symétrique :

$$u = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} , \quad v = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{F.411}$$

on a :

$$\|u\|_1 = \|u\|_2 = \|u\|_\infty = 1 \tag{F.412}$$

et l'ordre inverse s'observe :

$$\| A \|_1 = \|v\|_\infty = 1 < \| A \|_2 = \|v\|_2 = \sqrt{M} < \| A \|_\infty = \|v\|_1 = M \tag{F.413}$$

(3) Des applications particulières des majorations indiquées en (A.13) fournissent les résultas suivants :

$$\forall\, w \neq 0 \ : \frac{\|w\|_1}{\|w\|_2} \leq \sqrt{M} , \quad \frac{\|w\|_2}{\|w\|_1} \leq 1 \tag{F.414}$$

$$\frac{\|w\|_2}{\|w\|_\infty} \leq \sqrt{M} , \quad \frac{\|w\|_\infty}{\|w\|_2} \leq 1 \tag{F.415}$$

$$\frac{\|w\|_\infty}{\|w\|_1} \leq 1 , \quad \frac{\|w\|_1}{\|w\|_\infty} \leq M \tag{F.416}$$

Ces résultats permettent de construire le tableau suivant des majorations du rapport

$\| A \|_p / \| A \|_q$:

$\frac{\| A \|_p}{\| A \|_q}$	$q = 1$	$q = 2$	$q = \infty$
$p = 1$	1	$\frac{\|u\|_1}{\|u\|_2}.\frac{\|v\|_\infty}{\|v\|_2} \leq \sqrt{M}.1$	$\frac{\|u\|_1}{\|u\|_\infty}.\frac{\|v\|_\infty}{\|v\|_1} \leq M.1$
$p = 2$	$\frac{\|u\|_2}{\|u\|_1}.\frac{\|v\|_2}{\|v\|_\infty} \leq 1.\sqrt{M}$	1	$\frac{\|u\|_2}{\|u\|_\infty}.\frac{\|v\|_2}{\|v\|_1} \leq \sqrt{M}.1$
$p = \infty$	$\frac{\|u\|_\infty}{\|u\|_1}.\frac{\|v\|_1}{\|v\|_\infty} \leq 1.M$	$\frac{\|u\|_\infty}{\|u\|_2}.\frac{\|v\|_1}{\|v\|_2} \leq 1.\sqrt{M}$	1

(F.417)

Exercice B.1 (Approximation centrée d'un opérateur elliptique en 2D)

$$r_{j,k} = \sigma_x^2 \frac{-u_{j-1,k} + 2\,u_{j,k} - u_{j+1,k}}{h_x^2} + \sigma_y^2 \frac{-u_{j,k-1} + 2\,u_{j,k} - u_{j,k+1}}{h_y^2} \quad \text{(F.418)}$$

Exercice B.2 (Somme directe d'opérateurs, spectre)

(1) Explicitons le vecteur résidu :

$$r_h = r_{h\,x} + r_{h\,y} \quad \text{(F.419)}$$

où :

$$r_{h\,x} = \frac{\sigma_x}{h_x^2} \begin{pmatrix} & 2u_{1,1} & -u_{2,1} \\ & 2u_{1,2} & -u_{2,2} \\ & \vdots & \\ & 2u_{1,L} & -u_{2,L} \\ & - - - & \\ -u_{1,1}+ & 2u_{2,1} & -u_{3,1} \\ -u_{1,2}+ & 2u_{2,2} & -u_{3,2} \\ & \vdots & \\ -u_{1,L}+ & 2u_{2,L} & -u_{3,L} \\ & - - - & \\ & \vdots & \\ & - - - & \\ -u_{M-1,1}+ & 2u_{M,1} & \\ -u_{M-1,2}+ & 2u_{M,2} & \\ & \vdots & \\ -u_{M-1,L}+ & 2u_{M,L} & \end{pmatrix} \tag{F.420}$$

et :

$$r_{h\,y} = \frac{\sigma_y}{h_y^2} \begin{pmatrix} & 2u_{1,1} & -u_{1,2} \\ -u_{1,1}+ & 2u_{1,2} & -u_{1,3} \\ & \vdots & \\ -u_{1,L-1}+ & 2u_{1,L} & \\ & - - - & \\ & 2u_{2,1} & -u_{2,2} \\ -u_{2,1}+ & 2u_{2,2} & -u_{2,3} \\ & \vdots & \\ -u_{2,L-1}+ & 2u_{2,L} & \\ & - - - & \\ & \vdots & \\ & - - - & \\ & 2u_{M,1} & -u_{M,2} \\ -u_{M,1}+ & 2u_{M,2} & -u_{M,3} \\ & \vdots & \\ -u_{M,L-1}+ & 2u_{M,L} & \end{pmatrix} \tag{F.421}$$

Pour $j = 1, 2, \ldots, M$, on définit le vecteur de $\mathbb{R}^L$ suivant :

$$u_{j,\cdot} \overset{\text{déf}}{=} \begin{pmatrix} u_{j,1} \\ u_{j,2} \\ \vdots \\ u_{j,L} \end{pmatrix} \tag{F.422}$$

de sorte que :

$$u_h = \begin{pmatrix} u_{1,\cdot} \\ u_{2,\cdot} \\ \vdots \\ u_{M,\cdot} \end{pmatrix} \tag{F.423}$$

(vecteur de $\mathbb{R}^{M\,L}$). On introduit les opérateurs suivants :

$$A_{h\,x} = \frac{\sigma_x}{h_x^2}\,\text{Trid}_{D\,D\,M\times M}\,(-1, 2, -1) \tag{F.424}$$

$$A_{h\,y} = \frac{\sigma_y}{h_y^2}\,\text{Trid}_{D\,D\,L\times L}\,(-1, 2, -1) \tag{F.425}$$

où ces matrices sont de dimensions respectives $M \times M$ et $L \times L$ et ont la structure habituelle associée à un problème unidimensionnel où les conditions aux limites sont de type Dirichlet (cf. (1.11)). Il vient d'une part :

$$\begin{aligned} r_{h\,x} &= \frac{\sigma_x}{h_x^2} \begin{pmatrix} & & 2u_{1,\cdot} & -u_{2,\cdot} \\ & & --- & \\ & -u_{1,\cdot}+ & 2u_{2,\cdot} & -u_{3,\cdot} \\ & & --- & \\ & & \vdots & \\ & & --- & \\ & -u_{M-1,\cdot}+ & 2u_{M,\cdot} & \end{pmatrix} \\ &= \frac{\sigma_x}{h_x^2} \begin{pmatrix} 2I_L & -I_L & & & \\ -I_L & 2I_L & -I_L & & \\ & \ddots & \ddots & \ddots & \\ & & -I_L & 2I_L & -I_L \\ & & & -I_L & 2I_L \end{pmatrix} \begin{pmatrix} u_{1,\cdot} \\ u_{2,\cdot} \\ \vdots \\ u_{M,\cdot} \end{pmatrix} \\ &= A_{h\,x} \otimes I_L\, u_h \end{aligned} \tag{F.426}$$

(où I_L est la matrice identité de dimension $L \times L$) ; d'autre part :

$$\begin{aligned} r_{h\,y} &= \begin{pmatrix} A_{h\,y}\,u_{1,.} \\ - - - \\ A_{h\,y}\,u_{2,.} \\ - - - \\ \vdots \\ - - - \\ A_{h\,y}\,u_{M,.} \end{pmatrix} \\ &= \begin{pmatrix} A_{h\,y} & & & \\ & A_{h\,y} & & \\ & & \ddots & \\ & & & A_{h\,y} \end{pmatrix} \begin{pmatrix} u_{1,.} \\ u_{2,.} \\ \vdots \\ u_{M,.} \end{pmatrix} \\ &= I_M \otimes A_{h\,y}\,u_h \end{aligned} \tag{F.427}$$

(où I_M est la matrice identité de dimension $M \times M$). Par identification, il vient finalement :

$$A_h = A_{h\,x} \otimes I_L + I_M \otimes A_{h\,y} = A_{h\,x} \oplus A_{h\,y} \tag{F.428}$$

(2) En adaptant les notations, le Théorème 1.1 fournit :

$$\alpha_m = \underline{\lambda}_m^{h_x} = \sigma_x \, \frac{2 - 2\,\cos\,\theta_{x\,m}}{h_x^2}\,,\ \theta_{x\,m} = m\pi h_x = \frac{m\pi}{M+1}\,,\ m = 1, 2, ..., M \tag{F.429}$$

$$\beta_\ell = \underline{\lambda}_\ell^{h_y} = \sigma_y \, \frac{2 - 2\,\cos\,\theta_{y_\ell}}{h_y^2}\,,\ \theta_{y_\ell} = \ell\pi h_y = \frac{\ell\pi}{L+1}\,,\ \ell = 1, 2, ..., L \tag{F.430}$$

et il résulte directement du Corollaire B.1 que le spectre de la matrice A_h est consitué des nombres suivants :

$$\begin{aligned} \underline{\lambda}_{m,\ell}^{h} &= \underline{\lambda}_m^{h_x} + \underline{\lambda}_\ell^{h_y} \\ &= \sigma_x \, \frac{2 - 2\,\cos\,\theta_{x\,m}}{h_x^2} + \sigma_y \, \frac{2 - 2\,\cos\,\theta_{y_\ell}}{h_y^2} \quad (m = 1, 2, ..., M\ ;\ \ell = 1, 2, ..., L) \end{aligned} \tag{F.431}$$

Exercice C.1 (Propriétés des polynômes de Tchebychev)

(1) L'équation

$$\boxed{T_k(x) \stackrel{\text{déf}}{=} \sum_{m=0}^{E(k/2)} C_k^{2m}\, x^{k-2m}\, (x^2 - 1)^m} \tag{F.432}$$

montre que le polynôme $T_k(x)$ est une somme de polynômes de degré k de la parité de k. Le coefficient de x^k est le suivant :

$$\alpha_k = \sum_{m=0}^{E(k/2)} C_k^{2m} = 2^{k-1} \neq 0 \tag{F.433}$$

Par conséquent le polynôme $T_k(x)$ est de degré k exactement, et de la parité de k.

(2) Soit $x \in [-1, 1]$; en posant :

$$\theta = \text{Arccos}\, x \tag{F.434}$$

il vient :

$$T_k(x) = T_k(\cos\,\theta) = \cos\,k\theta = \cos\,(k\,\text{Arccos}\, x) \tag{F.435}$$

Si maintenant, $x > 1$, en posant :

$$\theta = \text{Argch}\, x \tag{F.436}$$

il vient :

$$\begin{aligned} T_k(x) &= \sum_{m=0}^{E(k/2)} C_k^{2m}\, \text{ch}^{k-2m}\,\theta \underbrace{\left(\text{ch}^2\,\theta - 1\right)^m}_{\text{sh}^{2m}\,\theta} \\ &= \frac{\left(\text{ch}\,\theta + \text{sh}\,\theta\right)^k + \left(\text{ch}\,\theta - \text{sh}\,\theta\right)^k}{2} \\ &= \text{ch}\,k\theta \\ &= \text{ch}\ (k\,\text{Argch}\ \ \text{x}) \end{aligned} \tag{F.437}$$

Enfin, si $x < -1$ la parité du polynôme permet d'écrire :

$$T_k(x) = (-1)^k\, T_k(-x) = (-1)^k\ \text{ch}\ [k\,\text{Argch}\ \ \text{(-x)}] \tag{F.438}$$

En particulier :

$$T_k(1) = \cos\,(k \times 0) = 1 \tag{F.439}$$

et :

$$T_k(-1) = (-1)^k\, T_k(1) = (-1)^k \tag{F.440}$$

(3) On considère d'abord le cas de $x \in [-1, 1]$ pour lequel :

$$\begin{aligned} T_{k+1}(x) + T_{k-1}(x) &= \cos\,(k+1)\theta + \cos\,(k-1)\theta \\ &= 2\,\cos\,\theta\,\cos\,k\theta \\ &= 2\,x\,T_k(x) \end{aligned} \tag{F.441}$$

$$
\begin{aligned}
&T_0(x) = 1\\
&T_1(x) = x\\
&T_2(x) = 2\,x^2 - 1\\
&T_3(x) = 4\,x^3 - 3\,x\\
&T_4(x) = 8\,x^4 - 8\,x^2 + 1\\
&T_5(x) = 16\,x^5 - 20\,x^3 + 5\,x\\
&T_6(x) = 32\,x^6 - 48\,x^4 + 18\,x^2 - 1\\
&T_7(x) = 64\,x^7 - 112\,x^5 + 56\,x^3 - 7\,x\\
&T_8(x) = 128\,x^8 - 256\,x^6 + 160\,x^4 - 32\,x^2 + 1\\
&T_9(x) = 256\,x^9 - 576\,x^7 + 432\,x^5 - 120\,x^3 + 9\,x\\
&T_{10}(x) = 512\,x^{10} - 1280\,x^8 + 1120\,x^6 - 400\,x^4 + 50\,x^2 - 1\\
&T_{11}(x) = 1024\,x^{11} - 2816\,x^9 + 2816\,x^7 - 1232\,x^5 + 220\,x^3 - 11\,x
\end{aligned}
$$

Tableau F.6. Douze premiers polynômes de Tchebychev

Comme cette égalité entre polynômes est vraie uniformément sur [-1,1], c'est une identité sur $\mathbb{R}$ entier.

(4) La relation de récurrence précédente permet de générer les premiers éléments de la suite des polynômes de Tchebychev récursivement (voir tableau F.6).

A l'extérieur de l'intervalle [-1,1], le polynôme $T_k(x)$ a une variation monotone et ne s'annule pas. Ses zéros et ses extrêmas appartiennent donc à cet intervalle.

Zéros du polynôme $T_k(x)$:

Ce sont les abscisses ξ_j telles que :

$$k\,\text{Arccos}\,\xi_j = (j - \tfrac{1}{2})\,\pi \quad (j = 1, 2, ...) \tag{F.442}$$

On en trouve k distincts dans l'intervalle ouvert]-1,1[:

$$\xi_j = \cos\,\frac{(2j-1)\,\pi}{2k} \quad (j = 1, 2, ..., k) \tag{F.443}$$

Extrêmas du polynôme $T_k(x)$:

Ces extrêmas sont égaux à 1 et -1 alternativement, et sont localisés aux abscisses η_j telles que :

$$k\,\text{Arccos}\,\eta_j = j\,\pi \quad (j = 0, 1, ...) \tag{F.444}$$

On en trouve $k + 1$ distincts dans l'intervalle fermé [-1,1] :

$$\eta_j = \cos\,\frac{j\,\pi}{k} \quad (j = 0, 1, ..., k) \tag{F.445}$$

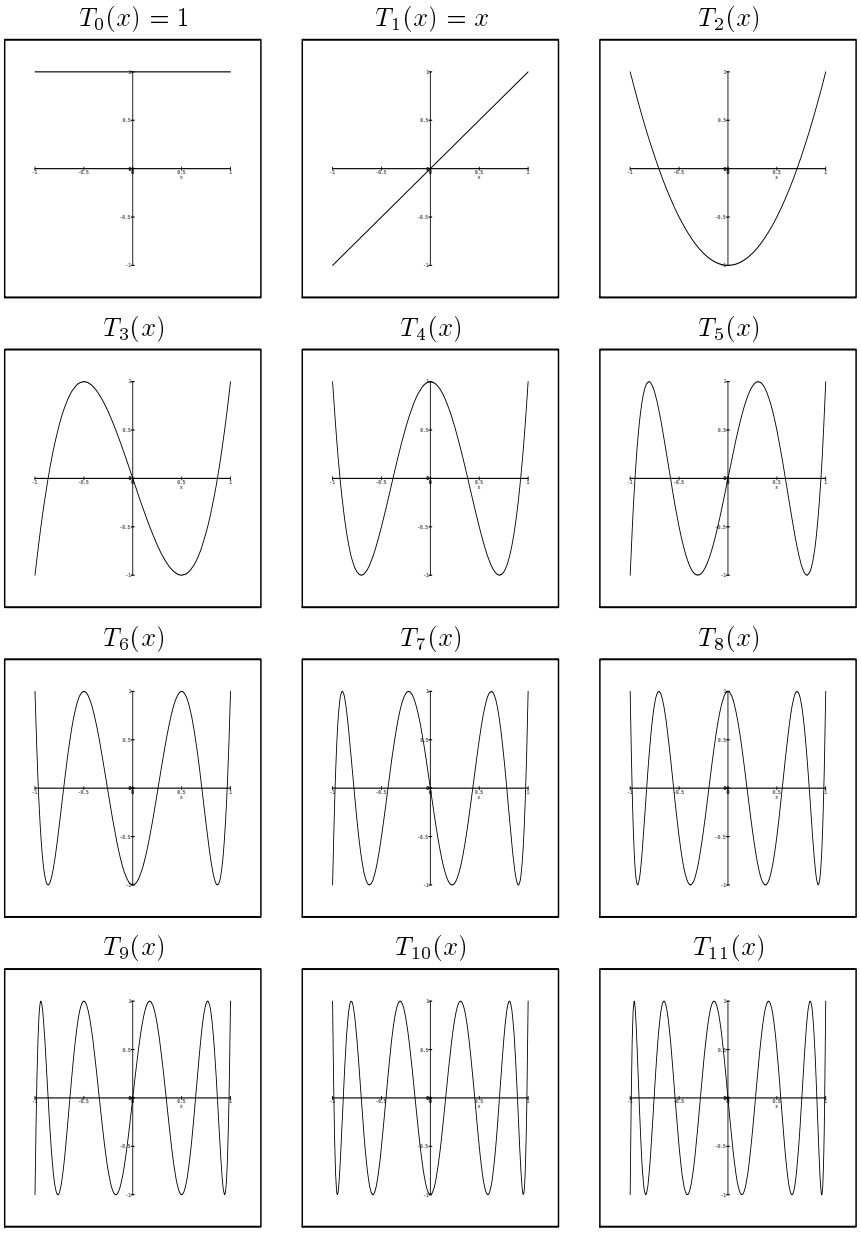


Figure F.10. Courbes représentatives des variations des douze premiers polynômes de Tchebychev sur l'intervalle [-1,1]

Les courbes représentatives des 12 premiers polynômes de Tchebychev sont données à la figure F.10.

(5) Orthogonalité : Soient k et ℓ deux entiers positifs distincts. Calculons le produit scalaire :

$$\begin{aligned}(T_k, T_\ell) &= \int_{-1}^{1} T_k(x)\, T_\ell(x)\, \frac{dx}{\sqrt{1-x^2}} \\ &= \int_{\pi}^{0} \cos\, k\theta\, \cos\, \ell\theta \frac{-\sin\theta\, d\theta}{\sin\, \theta} \\ &= \tfrac{1}{2} \int_{0}^{\pi} \left[\cos\,(k+\ell)\theta + \cos\,(k-\ell)\theta\right] d\theta \\ &= \tfrac{1}{2} \left[\frac{1}{k+\ell}\, \sin\,(k+\ell)\theta + \frac{1}{k-\ell}\, \sin\,(k-\ell)\theta\right]_0^{\pi} \\ &= 0 \end{aligned} \tag{F.446}$$

Ce résultat établit que les polynômes de Tchebychev forment une famille de polynômes orthogonaux vis à vis de ce produit scalaire.

Bibliographie

[1] Achdou, Y., Maday, Y. & Widlund, O. (1996) : « Méthode itérative de sous-structuration pour les éléments avec joints », *C.R. Acad. Sci., Paris*, Ser. **I** 322 (2) 185-190.

[2] Achdou, Y., Kuznetsov, Yu. A. & Pironneau, O. (1996) : « A mortar element method for an approximate Navier-Stokes solver », *Experimentation, modelling and computation in flow, turbulence and combustion, Volume 1*, Désidéri, J.A. (ed.) *et al.* , *Proc. of the 2nd French-Russian workshop on fluid dynamics, Sophia-Antipolis, 1993*, John Wiley & Sons, *Computational Methods in Applied Sciences*, 1-13.

[3] Ahlfors, Lars V. (1966) : *Complex Analysis, an introduction to the theory of analytic functions of one complex variables*, Second Edition, International Student Edition, McGraw-Hill Kogakusha, Ltd..

[4] Anderson, D.A., Tannehill, J.C. & Pletcher, R.H. (1984) : *Computational Fluid Mechanics and Heat Transfer*, McGraw-Hill, New York.

[5] Angrand, F. & Leyland, P. (1987) : « Schéma multigrille dynamique pour la simulation d'écoulements de fluides visqueux compressibles », *Rapport de Recherche INRIA* N° 659.

[6] Angrand, F. & Leyland, P. (1989) : « Compressible Viscous Flow Simulation by Multigrid Methods », *Computer Methods in Applied Mechanics and Engineering* **75**, 167-183.

[7] Arminjon, P. & Dervieux, A. (1993) : « Construction of TVD-like Artificial Viscosities on Two-Dimensional Arbitrary FEM Grids », *J. Comput. Phys.* **106**, N° 1, 176-198.

[8] Atelier ERCIM (1992) : « Méthodes multigrilles en mécanique des fluides numérique », *INRIA Sophia-Antipolis, 18-21 février, 1992*.

[9] Bank, R. & Xu, J. (1994) : « The hierarchical basis multigrid method and incomplete LU decomposition », *Domain decomposition methods in scientific and*

engineering computing, Proc. Seventh International Conference on Domain Decomposition, Pennsylvania State University, 1993, Keyes, David E. (ed.) *et al.*, *AMS*, Providence, *Contemp. Math.* **180**, 163-173.

[10] Bellman, R. (1960) : *Introduction to Matrix Analysis*, McGraw-Hill, New York.

[11] Bernardi, C., Maday, Y. & Patera, A.T. (1994) : « A new nonconforming approach to domain decomposition : The mortar element method », *Nonlinear partial differential equations and their applications, Collège de France Seminar*, Volume XI, Paris, 1989-1991, Brezis, H. (ed.) *et al.*, Pitman Research Notes in Mathematics Series, **299** 13-51.

[12] Beux, F. & Dervieux, A. (1994) : « A Hierarchical Approach for Shape Optimization », *Engineering Computations* **11**, N° 1, 25-48.

[13] Böhmer, K., Hemker, P. & Stetter, H. (1984) : « The defect correction approach », *Comput. Suppl.* **5** 1-32.

[14] Bourgat, J.F., Le Tallec, P., Perthame, B. & Qiu, Y. (1994) : « Coupling Boltzmann and Euler equations without overlapping », *Domain decomposition methods in science and engineering, Proc. Sixth International Conference on Domain Decomposition*, Como, 1992, Quarteroni, Alfio (ed.) *et al.*, *AMS*, Providence, *Contemp. Math.* **157**, 377-398.

[15] Bourouchaki, H. & Frey, P. (1997) : « Maillage géométrique de surfaces. Partie I : enrichissement », *Rapport de Recherche INRIA* N° 3236.

[16] Bourouchaki, H. & Frey, P. (1997) : « Maillage géométrique de surfaces. Partie II : appauvrissement », *Rapport de Recherche INRIA* N° 3237.

[17] Boyer, R., Martinet, B. & Saikouk, K. (1990) : « Algorithme de type FAC pour la résolution d'un problème d'écoulement diphasique en milieu poreux », *Rapport interne Maths. Appl., Université de Provence*.

[18] Bramble, J.H. (1993) : *Multigrid Methods*, Pitman Research Notes in Mathematics Series, Longman Scientific & Technical, Harlow UK, copublished with John Wiley & Sons, New York.

[19] Brandt, A. (1977) : « Multi-level adaptive solutions to boundary value problems », *Math. Comput.* **31**, 333-390.

[20] Brandt, A. (1981) : « Guide to Multigrid Development » – *Multigrid Methods, Lecture Notes in Mathematics*, N° 960, 221-312, Springer-Verlag, Berlin.

[21] Brandt, A., McCormick, S. & Ruge, J. : « Algebraic Multigrid (AMG) for automatic multigrid solution with application to geodetic computations », *SIAM J. on Scientific Computing*.

[22] Brandt, A., McCormick, S. & Ruge, J. (1984) : « Algebraic Multigrid (AMG) for sparse matrix equations », *Sparsity and its applications, D.J. Evans (Ed.), Cambridge University Press*.

[23] Brandt, A. & Diskin, B. (1994) : « Multigrid solvers on decomposed domains », *Domain decomposition methods in science and engineering, Proc. Sixth International Conference on Domain Decomposition*, Como, 1992, Quarteroni, Alfio (ed.) *et al.*, *AMS*, Providence, *Contemp. Math.* **157**, 135-155.

[24] Brezis, H. (1983) : *Analyse fonctionnelle, Théorie et applications*, Collection Mathématiques Appliquées pour la Maîtrise sous la direction de P.G. Ciarlet et J.-L. Lions, Masson, Paris.

[25] Briggs, William L. (1991) : *A Multigrid Tutorial*, SIAM Publication, Philadelphia Penn., Third Printing.

[26] Bristeau, M.O., Etgen, G., Fitzgibbon, W., Lions, J.L., Périaux, J. & Wheeler M.F. Eds. (1997) : *Computational Science for the 21st Century*, John Wiley & Sons, Chichester.

[27] Byron, F.W., Jr. & Fuller, R.W. (1969) : *Mathematics of Quantum Physics, Volume One*, Addison-Wesley Publishing Company, Reading Mass., London.

[28] Carré, G. (1995) : « Simulation numérique d'écoulements turbulents compressibles et stationnaires par méthodes multigrilles », Thèse de Doctorat, Université de Nice–Sophia-Antipolis.

[29] Carré, G. (1997) : « An Implicit Multigrid Method by Agglomeration Applied to Turbulent Flows », *Computers & Fluids* **26**, N° 3, 299-320.

[30] Cocquebert, C. (1997) : « Méthodes de type "Waveform-FAC" pour la résolution numérique de problèmes paraboliques », Thèse de doctorat, Université de Provence (C.M.I. Technopôle de Château-Gombert).

[31] Conte, S.D. & de Boor, C. (1972) : *Elementary Numerical Analysis - An algorithmic Approach*, McGraw-Hill New York, Second Edition.

[32] Cowsar, L.C., Dean, E.J., Glowinski, R., Le Tallec, P., Li, C.H., Périaux, J. & Wheeler, M.F. (1992) : « Decomposition principles and their applications in scientific computing », *Proc. of the fifth SIAM conference on parallel processing for scientific computing*, Houston, USA, 1991, Dongarra, Jack (ed.) *et al.*, *SIAM*, Philadelphia, 213-237.

[33] Dautray, R. & Lions, J.L. (1984) : *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*, Collection du Commissariat à l'Energie Atomique, Série Scientifique, Tome 1, Masson (Paris – New York), 1984 ; Chap. II : L'opérateur de Laplace.

[34] Dautray, R. & Lions, J.L. (1984) : *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*, Collection du Commissariat à l'Energie Atomique, Série Scientifique, Tome 2, Masson (Paris – New York), 1984 ; Chap. VIII : Théorie spectrale.

[35] Désidéri, J.A. & Dervieux, A. (1988) : « Compressible Flow Solvers using Unstructured Grids », *Von Karman Institute for Fluid Dynamics, Lecture Series 1988-5, Computational Fluid Dynamics*, March 7-11, 1988, également *Rapport de Recherche INRIA* N° 1732, Juin 1992.

[36] Désidéri, J.-A. & Hemker, P.W. (1990) : « Analysis of the Convergence of Iterative Implicit and Defect-Correction Algorithms for Hyperbolic Problems », *Rapport de Recherche INRIA* N° 1200, Mars 1990.

[37] Désidéri, J.-A. (1993) : « La technique d'annihilation de modes propres et applications », *Rapport de Recherche INRIA* N° 1875, Avril 1993.

[38] Désidéri, J.-A. & Hemker, P.W. (1995) : « Convergence Analysis of the Defect-Correction Iteration for Hyperbolic Problems », *SIAM J. Scientific Computing* **16**, N° 1, 88-118, January 1995.

[39] Dryja, M., Smith, B.F. & Widlund, O.B. (1994) : « Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions », *SIAM J. Numer. Anal.* **31** (6) 1662-1694.

[40] Francescatto, J. (1998) : « Méthodes multigrilles par agglomération directionnelle pour le calcul d'écoulements turbulents », Thèse de doctorat, Université de Nice–Sophia-Antipolis.

[41] Francescatto, J. & Dervieux, A. (1998) : « A Semi-Coarsening Strategy for Unstructured Multigrid Method Based on Agglomeration », *Int. J. Numer. Meth. Fluids* **26**, 927-957.

[42] Gastaldi, F., Gastaldi, L. & Quarteroni, A. (1996) : « Adaptive domain decomposition methods for advection dominated equations », *East-West J. Numer. Math.* **4**, N° 3, 165-206.

[43] George, P.L. & Bourouchaki, H. (1997) : *Triangulation de Delaunay et maillage*, Editions Hermès, Paris.

[44] Guillard, H. (1993) : « Node-Nested Multi-Grid with Delaunay Coarsening », *Rapport de Recherche INRIA* N° 1898.

[45] Guillard, H. (à paraître) : « Une présentation des méthodes multigrilles ».

[46] Hackbuch, W. (1985) : *Multi-Grid Methods and Applications*, Springer Series in Computational Mathematics, Heidelberg.

[47] Hemker, P. & Koren, B. (1988) : « Defect Correction and Nonlinear Multigrid for the Steady Euler Equations », *CWI Report, NM-N8801, Amsterdam.*

[48] Householder, A.S. (1964) : *The Theory of Matrices in Numerical Analysis*, Dover Publications, Inc. New York.

[49] Jespersen, D. : communication privée.

[50] Khadra, K. , Angot, Ph., Caltagirone, J.P. & Morel, P. (1996) : « Concept de zoom adaptatif en architecture multigrille locale; étude comparative de méthodes L.D.C., F.A.C. et F.I.C. (Adaptive zoom concept in local multigrid architecture; comparative study of the methods L.D.C., F.A.C. and F.I.C.) », *RAIRO, Modelisation Math. Anal. Numer.* **30**, N° 1, 39-82.

[51] Kolmogorov, A. & Fomine, S. (1977) : *Eléments de la théorie des fonctions et de l'analyse fonctionnelle*, Editions Mir Moscou, chapitre II.

[52] Koobus, B. (1994) : « Algorithmes multigrille et Algorithmes Implicites pour les Ecoulements Compressibles Turbulents », Thèse de doctorat, Université de Nice–Sophia-Antipolis.

[53] Koobus, B., Lallemand, M.H. & Dervieux, A. (1994) : « Unstructured Volume Agglomeration MG : Solution of Poisson Equation », *Int. J. Numer. Methods Fluids* **18**, 27-42.

[54] Koren, B. (1988) : « Defect Correction and Multigrid for an Efficient and Accurate Computation of Airfoil Flows » – *J. Comput. Phys.* **77** 183-206.

[55] Kronsjö, L. & Dahlquist, G. (1971-2) : « On the design of nested iterations for elliptic difference equations », *BIT* **12**, 63-71, 1972. (Le tiré-à-part porte la référence *BIT* **11** (1971).)

[56] Kuznetsov, Yu. A. (à paraître) : « Overlapping domain decomposition with nonmatching grids », *Proc. Ninth conference on domain decomposition methods*, Bergen, 1996, Bjørstad, P.E. (ed.) *et al.*, John Wiley & Sons.

[57] Lallemand, M.H. (1988) : « Schémas décentrés multigrilles pour la résolution des équations d'Euler en éléments finis », Thèse de doctorat, Université de Provence.

[58] Lallemand, M.H., Stève, H. & Dervieux, A. (1992) : « Unstructured Multigridding by Volume Agglomeration : Current Status », *Computers & Fluids* **21**, 397-433.

[59] Leclercq, M.P. (1990) : « Résolution des équations d'Euler par des méthodes multigrilles – Conditions aux limites en régime hypersonique », Thèse de doctorat, Université de Saint-Etienne.

[60] Leclercq, M.P. & Stoufflet, B. (1993) : « Characteristic Multigrid Method Application to Solve the Euler Equations with Unstructured and Unnested Grids », *J. Comput. Phys.* **104** (2) 329-46.

[61] Le Tallec, P. (1993) : « Méthodes de décomposition de domaines en calcul des structures », *Colloque national en calcul des structures*, Giens, 1993, Editions Hermès, Paris, **1**, 33-49.

[62] Le Tallec, P. (1994) : « Domain decomposition methods in computational mechanics », *J. Comput. Mech. Adv.* **1** (2) 121-220.

[63] Le Tallec, P., Mandel, J. & Vidrascu, M. (1994) : « Balancing domain decomposition for plates », *Domain decomposition methods in scientific and engineering computing, Proc. Seventh International Conference on Domain Decomposition*, Pennsylvania State University, 1993, Keyes, David E. (ed.) *et al.*, *AMS*, Providence, *Contemp. Math.* **180**, 515-524.

[64] Lions, P.L. (1988) : « On the Schwarz alternating method », *Proc. of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations*, *SIAM*, R. Glowinski, G.H. Golub, G.A. Meurant and J. Périaux, eds., 1-42.

[65] Lomax, H. : communication privée.

[66] Maître, J.F. & Musy, F. (1984) : « Multigrid methods : convergence theory in a variational framework », *SIAM J. Numer. Anal.* **4**, 657-671.

[67] Manteufel, T.A. (1977) : « The Tchebychev Iteration for Nonsymmetric Linear Systems », *Numerical Mathematics* **28**, 307-327.

[68] Marco, N. (1995) : « Optimisation de formes aérodynamiques 2D et 3D par une méthode multi-niveau en maillages non structurés », Thèse de doctorat, Université de Nice–Sophia-Antipolis.

[69] Martin, R. & Guillard, H. (1996) : « A Second-Order Defect Correction Scheme for Unsteady Problems », *Computers & Fluids* **25**, N° 1, 9-27.

[70] McCormick, S. & Thomas, J. (1986) : « The fast adaptive composite grid (FAC) methods for elliptic equations », *Math. Comp.* **46** (174) 439-456.

[71] McCormick, S. (1992) : « Multilevel Projection Methods for Partial Differential Equations », *CBMS-NSF Regional Conference Series in Applied Mathematics*, *SIAM*, Philadelphia.

[72] Mitchell, W.F. (1988) : *Unified multilevel adaptive finite element methods for elliptic problems*, Ph.D. thesis, Technical Report UIUCDCS-R-88-1436, Department of Computer Science, University of Illinois, Urbana, IL, 1988. Disponible par ftp anonyme à partir de casper.cs.yale.edu dans mgnet/papers/Mitchell/thesis.ps

[73] Morano, E. (1992) : « Résolution des équations d'Euler par une méthode multigrille stationnaire », Thèse de doctorat, Université de Nice–Sophia-Antipolis.

[74] Morano, E. & Dervieux, A. (1995) : « Steady Relaxation Methods for Unstructured Multigrid Euler and Navier-Stokes Solutions », *Comp. Fluid Dyn.* **5**, 137-167.

[75] Nepomnyaschikh, S.V. (1997) : « Domain decomposition and multilevel techniques for preconditioning operators », *Domain decomposition methods in sciences and engineering, Proc. Eighth Conference on Domain Decomposition*, Pekin, P.R. China, 1995, Glowinski, R. (ed.) *et al.*, *John Wiley & Sons*, 193-203.

[76] N3S-Natur V1.2, *Manuel d'utilisation et manuel théorique*, Simulog, 1998.

[77] Ortega, J.M. & Rheinboldt, W.C. (1970) : *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, London, chapitre 5.

[78] Pignol, D. (1995) : « Etude de quelques méthodes numériques de raffinement local. Application à la dynamique des fronts de flamme », Thèse de doctorat, Université de Provence (Aix-Marseille I).

[79] Quarteroni, A. (1994) : « Mathematical Aspects of Domain Decomposition Methods », *Joseph, A. (ed.) et al., First European Congress of Mathematics (ECM)*, Paris, 1992. *Volume II : Invited Lectures (Part 2). Basel : Birkhaeuser, Prog. Math.* 120, 355-379.

[80] Queysanne, M. (1964) : *Algèbre, M.P. et Spéciales AA'*, Collection U, Cinquième Edition, Librairie Armand Colin, Paris.

[81] Ruge, J. & Stüben, K. (1987) : « Algebraic Multigrid (AMG) », *Frontiers in App. Math. Vol. 3 : Multigrid Methods, SIAM, S. McCormick (Ed.)*, Philadelphia.

[82] Saad, Y. (1996) : *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, Paris.

[83] Smith, B., Bjørstad, P.E. & Gropp, W. (1996) : *Domain Decomposition : Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, New York.

[84] Stève, H. (1988) : « Schémas Implicites Linéarisés Décentrés pour la Résolution des Equations d'Euler en Plusieurs Dimensions », Thèse de Doctorat, Université de Provence Aix-Marseille I.

[85] Strang, G. (1980) : *Linear Algebra and Its Applications*, Second Edition, Academic Press, New York, 1980.

[86] Swartz, C. (1992) : *An introduction to functional analysis*, in Pure and Applied Mathematics, A series of monographs and textbooks, Marcel Dekker, New York.

[87] Ta'asan, S. (1993) : « Optimal multigrid method for inviscid flows », *Multigrid methods IV, CWI Tract, Proc. Fourth European Multigrid Conference*, P.W. Hemker, P. Wesseling (Eds.), Amsterdam. *Int. Ser. Numer. Math.* **116**, 309-320 (1994).

[88] Van Albada, G.D., van Leer, B. & Roberts, W.W. (1982) : « A Comparative Study of Computational Methods in Cosmic Gas Dynamics », *Astron. Astrophys.* **108**, 76-84.

[89] van Leer, B. (1979) : « Towards the ultimate conservative difference scheme V – A second order sequel to Godunov's method », *J. Comput. Phys.* **32**, 101-136.

[90] van Leer, B. (1982) : « Flux-Vector Splitting for the Euler Equations », 8th International Conference on Numerical Methods in Fluid Dynamics, Krause Ed., pp. 507-512, *Lecture Notes in Physics* **170**, Springer-Verlag.

[91] van Leer, B. (1983) : « Computational Methods for Ideal Compressible Flow », Von Karman Institute for Fluid Dynamics, Rhode-Saint-Genèse, *Lecture Series* 1983-04.

[92] Varga, R.S. (1962) : *Matrix Iterative Analysis*, Prentice Hall, Inc., Englewood Cliffs, New Jersey.

[93] Warming, R.F. & Hyett, B.J. (1974) : « The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods », *J. Comput. Phys.* **14**, 159-179.

[94] Warming, R.F., Beam, R.M. & Hyett, B.J. (1975) : « Diagonalization and Simultaneously Symmetrization of the Gas-Dyanmic Matrices », *Mathematics of Computation* **29** (132) 1037-1045.

[95] Wesseling, P. (1991) : *An Introduction to Multigrid Methods*, John Wiley & Sons, Chichester.

[96] Wilkinson, J.H. (1965) : *The Algebraic Eigenvalue Problem*, Oxford University Press, London and New York.

Index