La Technique d'annihilation de modes propres et applications

Jean-Antoine Desideri

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

2004 route des Lucioles
B.P. 93
06902 Sophia-Antipolis
France

Rapports de Recherche

N°1875

Programme 6

Calcul scientifique, Modélisation
et Logiciel numérique

LA TECHNIQUE D'ANNIHILATION DE MODES PROPRES ET APPLICATIONS

Jean-Antoine DESIDERI

Mars 1993

La Technique d'Annihilation de Modes Propres et Applications

Jean-Antoine DESIDERI

Résumé

Ce rapport vise trois objectifs principaux. Le premier est de rappeler dans un but pédagogique certaines notions classiques sur les méthodes itératives linéaires. Le second est de présenter une technique d'annihilation de modes propres assez générale, d'en discuter la réalisation algorithmique, et de démontrer qu'elle équivaut à brancher des boucles simples ou imbriquées de sur(sous)-relaxation sur l'algorithme de base. Enfin, on examine diverses applications spécifiques, et notamment en hyperbolique, la construction d'un "lisseur" explicite de type Runge-Kutta, et l'accélération de la méthode implicite du Résidu Corrigé. Dans ce dernier cas, le gain théorique en efficacité est de l'ordre de 2 pour le problème modèle bi-dimensionnel.

The Eigenmode Annihilation Technique and Applications

Jean-Antoine DESIDERI

Abstract

This report aims at three main objectives. The first is to recall for the purpose of pedagogy some classical notions on linear iterative methods. The second is to present a rather general technique of eigenmode annihilation, to discuss its algorithmic realization, and to prove that it is equivalent to branching simple or nested over(under)-relaxation loops on the base algorithm. Finally, several specific applications are examined, including for the hyperbolic case, the construction of an explicit Runge-Kutta-type "smoother", and the acceleration of the Defect-Correction implicit algorithm. In the latter case, the theoretical gain in efficiency is of the order of 2 for the two-dimensional model problem.

Table des matières

"La Technique d'Annihilation de Modes Propres"

Dis-moi quel spectre te hante,
Je te dirai qui tuer!

1 Généralités

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

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

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

$$G = T\,\Gamma\,T^{-1} \tag{2}$$

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

$$\Gamma = \begin{pmatrix} g_1 & & & \\ & g_2 & & \\ & & \ddots & \\ & & & g_M \end{pmatrix} \tag{3}$$

et les valeurs propres $\{g_m\}$ satisfont la condition de convergence :

$$\rho(G) = \rho(\Gamma) = \max_m\left(|g_m|\right) < 1 \tag{4}$$

où ρ désigne le rayon spectral. Dans ce cas, aucune des valeurs propres de G n'est égale à 1 et la matrice $I - G$ est inversible de sorte que (1) admet un point fixe unique

$$u^\infty = (I - G)^{-1}\,b\,. \tag{5}$$

Il en résulte que le vecteur "erreur" défini par

$$e^n = u^n - u^\infty\,, \tag{6}$$

vérifie l'équation récurrente linéaire homogène suivante :

$$e^{n+1} = G\,e^n\,. \tag{7}$$

En conséquence,

$$e^n = G^n\,e^0 = T\,\Gamma^n\,T^{-1}\,e^0\,, \tag{8}$$

et on pose :

$$\epsilon^n = T^{-1}\,e^n = \Gamma^n\,\epsilon^0\,. \tag{9}$$

Il vient :

$$\boxed{\begin{array}{l} \epsilon^n = \epsilon_1^0\, g_1^n\, \hat{I}_1 + \epsilon_2^0\, g_2^n\, \hat{I}_2 + \ldots + \epsilon_M^0\, g_M^n\, \hat{I}_M\,, \\ e^n = \epsilon_1^0\, g_1^n\, \hat{T}_1 + \epsilon_2^0\, g_2^n\, \hat{T}_2 + \ldots + \epsilon_M^0\, g_M^n\, \hat{T}_M\,, \end{array}} \tag{10}$$

où $\hat{I}_1$, $\hat{I}_2$, ..., $\hat{I}_M$ sont les vecteurs colonnes de la matrice identité (i.e. 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 munit $\mathbb{R}^M$ (ou $\mathbb{C}^M$) d'une p-norme (pour un réel $p > 0$),

$$\|u\| = \max \left(\sum_{j=1,\ldots,M} |u_j|^p \right)^{\frac{1}{p}} \quad (u \in \mathbb{R}^M)\,, \tag{11}$$

et l'ensemble des matrices $M \times M$ à coefficients réels (ou complexes) (noté ici $\mathcal{M}_{M\times M}$) de la norme induite :

$$\|A\| = \max_{u \in \mathbb{R}^M - \{0\}} \frac{\|A\,u\|}{\|u\|} \quad (A \in \mathcal{M}_{M\times M})\,. \tag{12}$$

On rappelle que

$$\forall u \in \mathbb{R}^M\,, \quad \forall A \in \mathcal{M}_{M\times M}\,, \quad \|A\,u\| \leq \|A\|\,.\,\|u\|\,, \tag{13}$$

et de plus, pour toute matrice diagonale Δ :

$$\|\Delta\| = \rho\,(\Delta)\,, \tag{14}$$

et en particulier :

$$\|\Gamma^n\| = \rho^n\,, \tag{15}$$

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 (8) :

$$\boxed{\|e^n\| \leq K\,\rho^n \|e^0\|\,,} \tag{16}$$

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

$$\rho = |g_1| \geq |g_2| \geq \ldots \geq |g_M|\,, \tag{17}$$

et que $|\epsilon_1^0| \neq 0$. Alors, en vertu de (10), lorsque $n \to \infty$:

$$\|\epsilon^n\| \sim C_1\, \rho^n\,, \quad \|e^n\| \sim C_2\, \rho^n\,, \tag{18}$$

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

$$\boxed{v = -\ln\, \rho\,,} \tag{19}$$

appelée **vitesse de convergence asymptotique**. La quantité $1/v$ 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 $\|e^n\|/\|e^0\|$ en échelle logarithmique en fonction de n; dans cette représentation, la suite des points admet une droite de pente $-v$ 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 (16) suggère que la convergence est conforme au résultat asymptotique seulement pour n très grand. Un tel comportement peut se produire et on réfère à [9] pour une discussion détaillée du cas défectif et des illustrations numériques.

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 connait 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 1.)

On pose

$$\boxed{A = I - G\,,} \tag{20}$$

de sorte que l'itération prend la forme suivante :

$$u^{n+1} = (I - A)\, u^n + b \tag{21}$$

que l'on identifie à un "pas en temps" de la méthode d'Euler sur le système d'équations différentielles ordinaires

$$\dot{u} = b - Au \tag{22}$$

avec

$$\Delta t = 1\,. \tag{23}$$

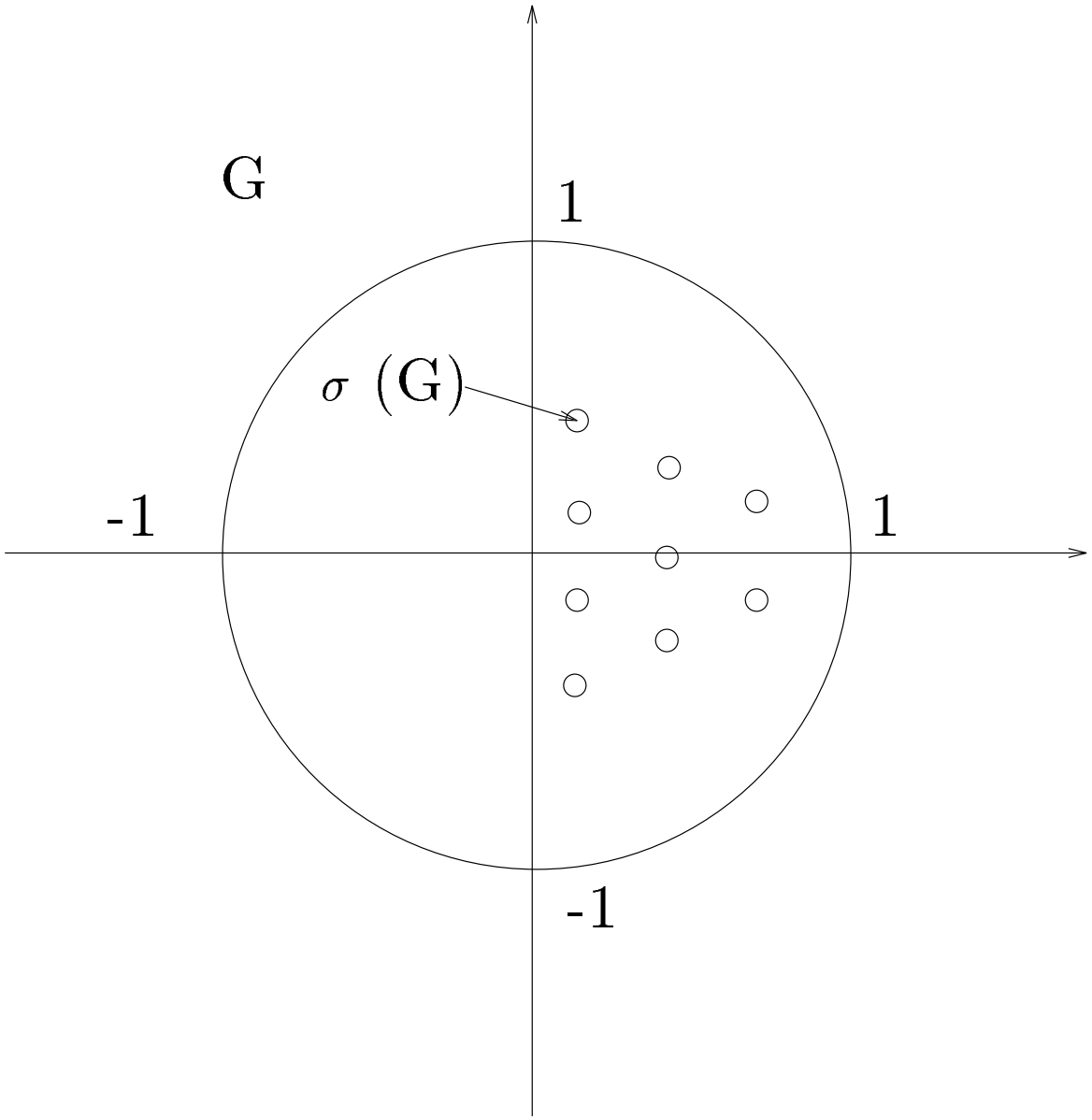

Figure 1: Spectre de G schématisé

Les valeurs propres de la matrice A sont les suivantes :

$$\lambda_m = 1 - g_m \tag{24}$$

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 2.)

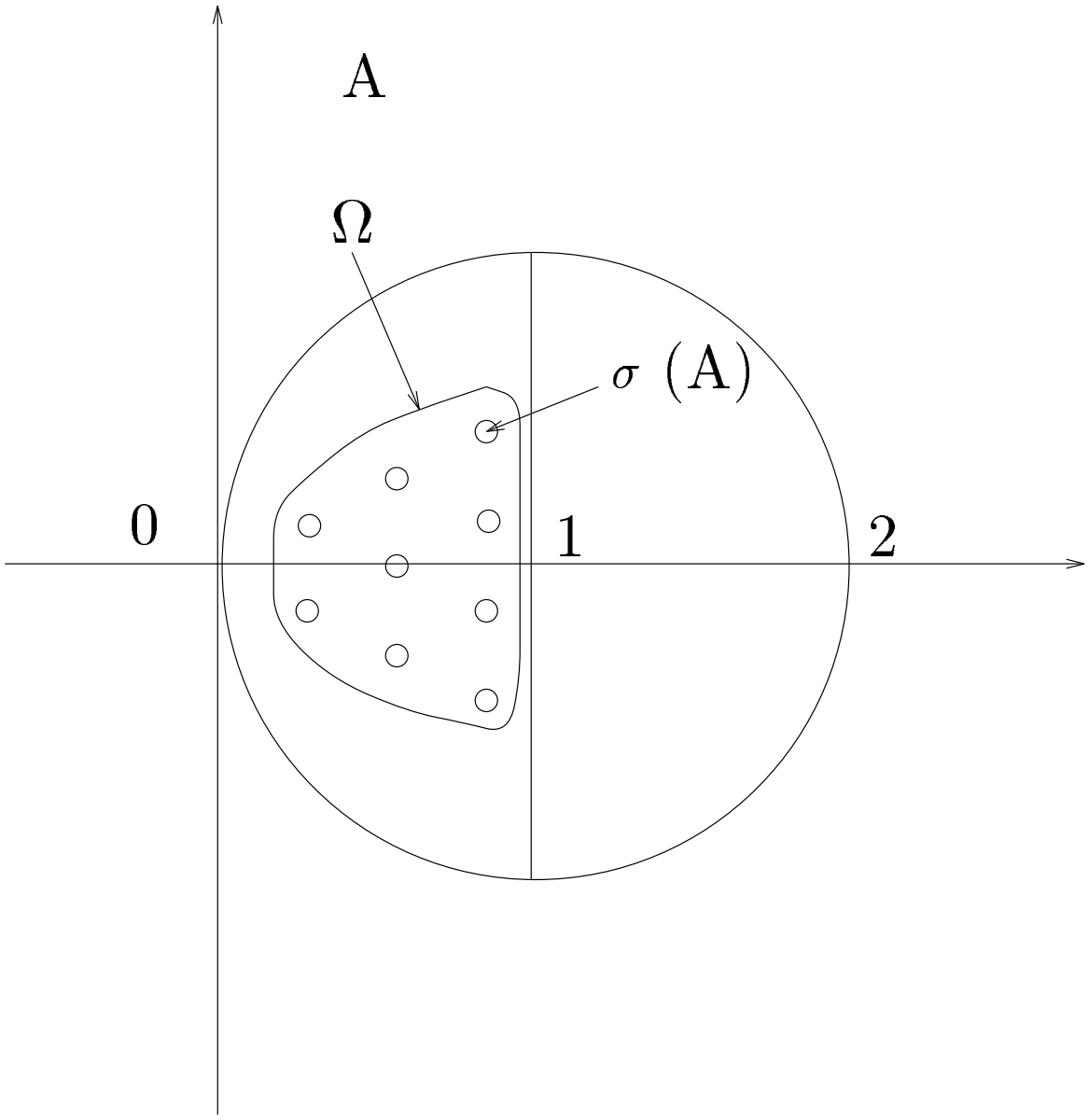

Figure 2: Spectre de A schématisé

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 :

$$u^{\infty} = A^{-1}\, b = (I - G)^{-1}\, b \tag{25}$$

qui est le point fixe de l'itération de départ; de plus, la solution exacte du système instationnaire

$$u(t) = u^{\infty} + (u(0) - u^{\infty})\; exp\,(-At) \tag{26}$$

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 opérateur 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 [2]. Cette propriété résulte directement de l'expression du $(n+1)$ème itéré de l'erreur (dans la base des vecteurs propres), ϵ^{n+1}, en fonction des composantes ϵ_m^n du précédent :

$$\epsilon^{n+1} = \sum_{m=1}^{M} (1 - \lambda_m \tau)\ \epsilon_m^n \,. \tag{27}$$

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 :

$$\Delta t = \tau = \frac{1}{\lambda_m} \,. \tag{28}$$

Lorsque $\lambda_m \in I\!R$ la réalisation de l'algorithme est évidente :

$$u^{n+1} = (I - \tau A) u^n + \tau b \tag{29}$$

ce qui s'interprète comme un schéma de sur(sous)-relaxation appliquée à l'itération de base :

$$\boxed{\begin{aligned} v^{n+1} &= (I - A) u^n + b = \mathbf{g}(u^n)\,, \\ u^{n+1} &= (1 - \omega)\, u^n + \omega\, v^{n+1}\,, \end{aligned}} \tag{30}$$

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 $\mathbf{g}$. (Voir Figure 3.)

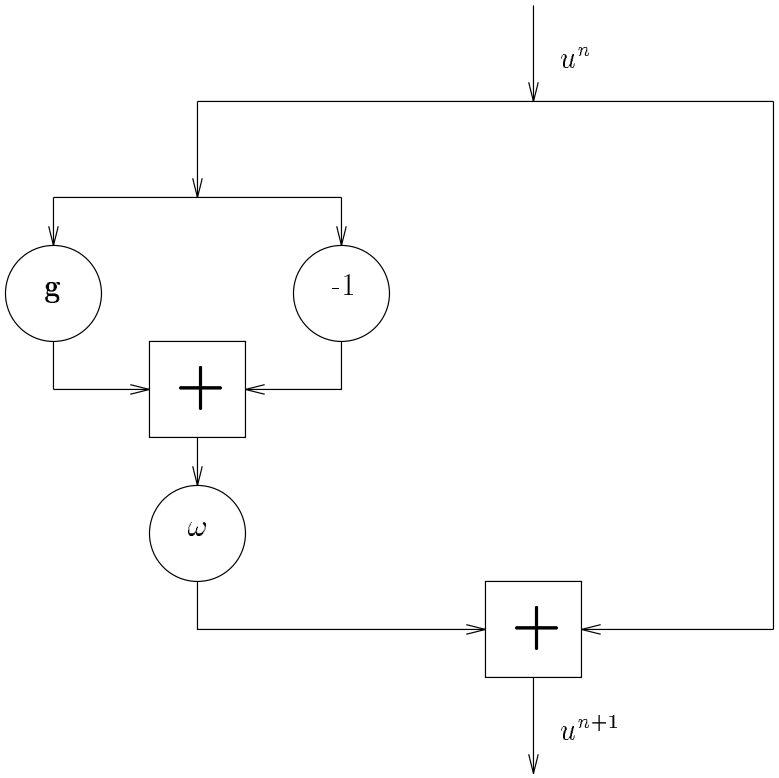

Figure 3: Schéma de sur(sous)-relaxation simple.

Revenons maintenant au cas général où λ_m est complexe dans (28). Si les matrices G et A sont elles-mêmes complexes, l'utilisation d'un pas de temps complexe 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 $\Delta t = \frac{1}{\lambda_m} = \tau$, le second avec $\Delta t = \bar{\tau}$, ce qui donne :

$$\begin{aligned} u^{n+1} &= (I - \tau A)u^n + \tau b \\ u^{n+2} &= (I - \bar{\tau} A)u^{n+1} + \bar{\tau} b\,. \end{aligned} \tag{31}$$

Ce cycle est naturel car $\bar{\lambda}_m$ est aussi une valeur propre; il équivaut à :

$$u^{n+2} = (I - \bar{\tau} A)(I - \tau A)u^n + b' \tag{32}$$

où le vecteur

$$\begin{aligned} b' &= (I - \bar{\tau} A)\tau b + \bar{\tau} b \\ &= 2\Re(\tau) b - |\tau|^2 A b \end{aligned} \tag{33}$$

est réel, et comme

$$\begin{aligned} (I - \bar{\tau} A)(I - \tau A) &= I - 2\Re(\tau) A + |\tau|^2 A^2 \\ &= I - 2\Re(\tau) A \left(I - \frac{|\tau|^2}{2\Re(\tau)} A \right) , \end{aligned} \tag{34}$$

le cycle se réalise en arithmétique réelle par la séquence "prédicteur-correcteur" suivante [3] :

Prédicteur :

$$v^{n+1} = u^n - \frac{|\tau|^2}{2\Re(\tau)} \left(Au^n - b\right) , \tag{35}$$

Correcteur :

$$u^{n+2} = u^n - 2\Re(\tau) \left(Av^{n+1} - b\right) . \tag{36}$$

Notons que l'implémentation de ce cycle nécessite donc la mise en mémoire d'un vecteur supplémentaire, v^{n+1}. Enfin posons,

$$\begin{aligned} \omega_1 &= \frac{|\tau|^2}{2\Re(\tau)} \\ \omega_2 &= 2\Re(\tau) \end{aligned} \tag{37}$$

et remplaçons la matrice A par $I - G$, afin d'obtenir la nouvelle formulation du schéma prédicteur-correcteur suivante :

Prédicteur :

$$\boxed{\begin{aligned} v^{n+1} &= u^n + \omega_1 \left[G\, u^n + b - u^n \right] , \\ &= u^n + \omega_1 \left[\mathbf{g}\left(u^n\right) - u^n \right] , \end{aligned}} \tag{38}$$

Correcteur :

$$\boxed{\begin{aligned} u^{n+2} &= u^n + \omega_2 \left[G\, v^{n+1} + b - v^{n+1} \right] , \\ &= u^n + \omega_2 \left[\mathbf{g}\left(v^{n+1}\right) - v^{n+1} \right] . \end{aligned}} \tag{39}$$

Sous cette forme, le cycle s'interprète comme deux étapes de sur(sous)-relaxation appliquée 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 4.)

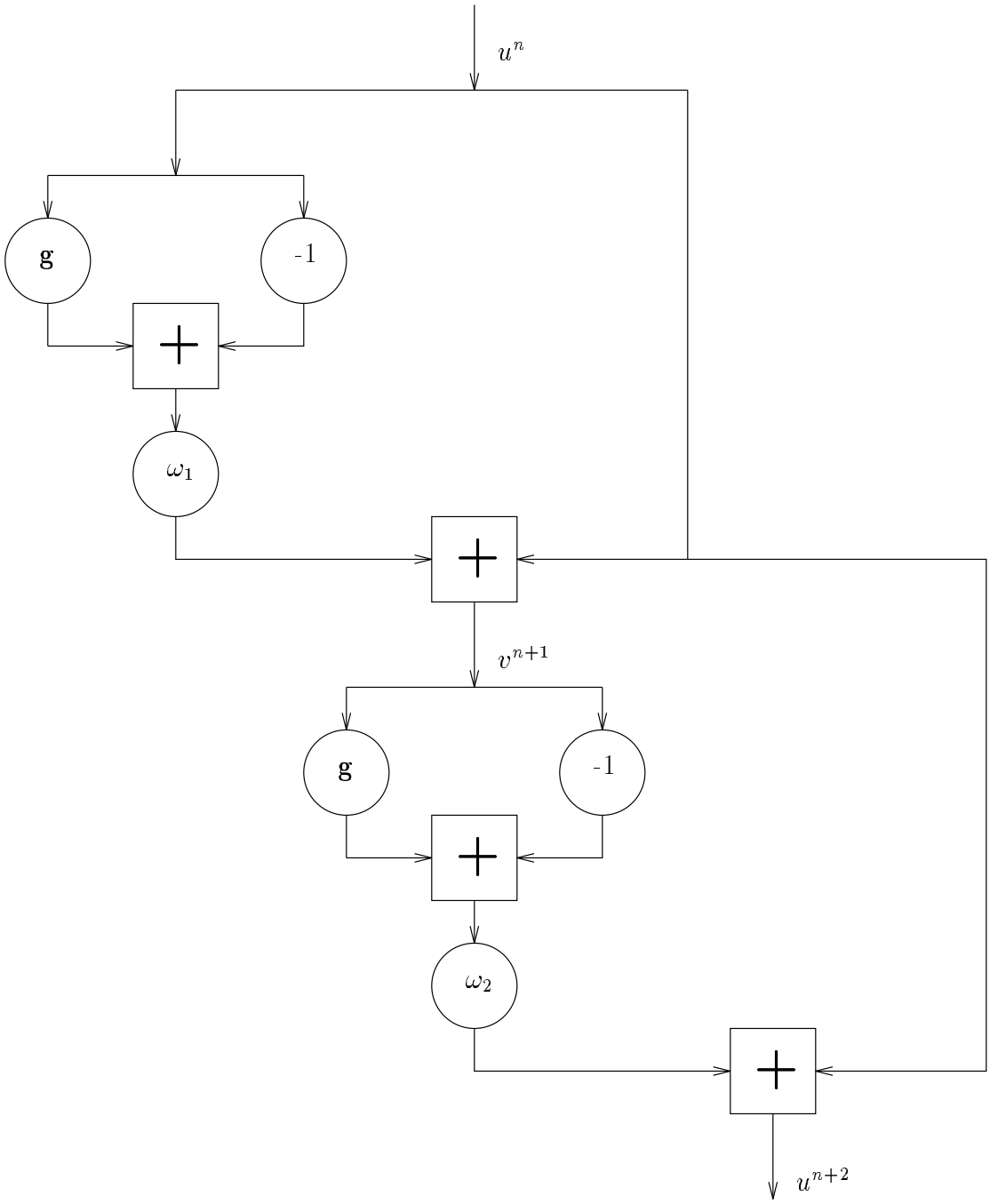

Figure 4: Schéma de sur(sous)-relaxation double.

Notons, enfin que si $\tau = \tau_r + i\,\tau_i$, on a :

$$\begin{cases} \omega_1 = \dfrac{\tau_r^2 + \tau_i^2}{2\tau_r}\,, \\ \\ \omega_2 = 2\tau_r\,, \end{cases} \tag{40}$$

ou inversement,

$$\begin{cases} \tau_r = \dfrac{\omega_2}{2}\,, \\ \\ \tau_i = \pm\sqrt{\omega_2\left(\omega_1 - \dfrac{\omega_2}{4}\right)}\,. \end{cases} \tag{41}$$

Ces équations montrent que lorsque $(\tau_r, \tau_i) \in I\!R_+^* \times I\!R$, (ω_1, ω_2) appartient au secteur angulaire de $I\!R_+^* \times I\!R_+^*$ correspondant à $\omega_2 \leq 4\omega_1$. Dans cette zone, ω_1 et ω_2 peuvent être supérieurs ou inférieurs à 1 indépendamment (sur ou sous-relaxation). (Voir Figure 5.)

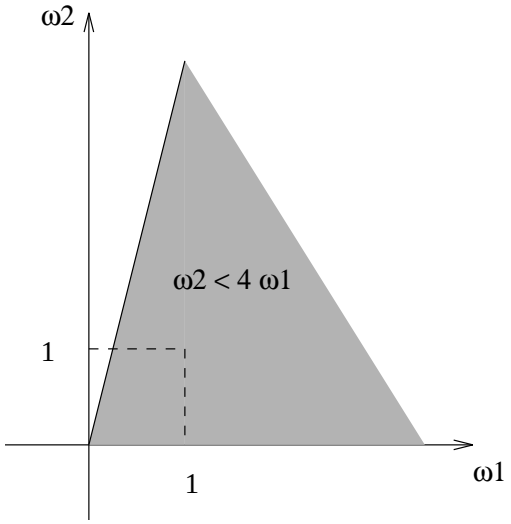

Figure 5: Domaine possible pour le couple (ω_1, ω_2).

3 Le problème classique du min-max

Plus généralement, lorsqu'on effectue un cycle de pas de temps complexes $\Delta t = \tau_1,\ \tau_2,\ ..., \tau_k$, l'optimisation de ces paramètres est réalisée lorsque le rayon spectral du cycle est minimisé. Ceci conduit à poser le problème suivant :

$$\min_{\tau_1,\ \tau_2,\ \cdots\ \tau_k}\ \max_{\lambda\,\in\,\sigma(A)} \left| \prod_{j=1}^{k} (1 - \lambda\tau_j) \right|^{1/k} . \tag{42}$$

(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 terme de rayon spectral équivalent par évaluation de la fonction $\mathbf{g}$.) 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, $k < M$. En pratique 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 :

$$\min_{\tau_1,\ \tau_2,\ \cdots\ \tau_k}\ \max_{\lambda\,\in\,\Omega} \left| \prod_{j=1}^{k} (1 - \lambda\tau_j) \right|^{1/k} . \tag{43}$$

Ce problème admet suivant la forme du domaine Ω quelques solutions exactes, et d'autres approchées, que nous allons maintenant examiner.

4 Solutions exactes connues et solutions approchées

4.1 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.

Posons

$$P(\lambda) = \prod_{j=1}^{k} (1 - \lambda\tau_j) \; . \tag{44}$$

Le polynôme $P(\lambda)$ est un polynôme à coefficients réels, de degré k exactement, dont les zéros sont réels et dont la valeur est égale à 1 pour $\lambda = 0$. Soit $\mathcal{P}$ la classe de tous les polynômes ayant ces propriétés. Réciproquement, tout polynôme de $\mathcal{P}$ peut se mettre sous la forme de $P(\lambda)$, de sorte que le problème de minimisation dans (43) est une optimisation dans $\mathcal{P}$ en entier; clairement, il s'agit de la minimisation de la norme infinie sur $[a, b]$. Le problème (43) équivaut donc à trouver dans $\mathcal{P}$ l'elément de plus petite norme :

$$\min_{P \in \mathcal{P}} \|P\|_{\infty/[a,b]} \, . \tag{45}$$

Soit $T_k(x)$ le k-ème polynôme de Tchebychev, à savoir le polynôme défini pour $x \in [-1, 1]$ par

$$T_k(\cos\theta) = \cos(k\theta) \, . \tag{46}$$

Ce polynôme est à coefficients réels, de degré k exactement, et ses zéros sont les réels suivants :

$$\xi_j = \cos\left(\frac{(2j-1)\pi}{2k}\right) \quad (j = 1, 2, ..., k) \, , \tag{47}$$

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 :

$$\boxed{c = \frac{b+a}{b-a} = \cosh\sigma} \, , \tag{48}$$

et

$$A_k = T_k(c) \, , \tag{49}$$

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 :

$$\forall \, x \in I\!R \, , \; \forall \, k \in I\!N^* \; : \; T_{k+1}(x) + T_{k-1}(x) = 2x \, T_k(x) \, . \tag{50}$$

En conséquence :

$$\forall \, k \in I\!N^* \; : \; A_{k+1} + A_{k-1} = 2c \, A_k \, , \tag{51}$$

et comme de plus,

$$A_0 = 1\,,\ A_1 = c\,, \tag{52}$$

on obtient facilement que :

$$\boxed{A_k = \cosh(k\,\sigma) = \cosh(k\ \cosh^{-1} c)}\,. \tag{53}$$

Soit alors le polynôme :

$$\boxed{P^*(\lambda) = \frac{1}{A_k}\,T_k\left(\frac{b+a-2\lambda}{b-a}\right) = \frac{(-1)^k}{A_k}\,T_k\left(\frac{2\lambda-(b+a)}{b-a}\right)}\,, \tag{54}$$

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

$$\mu_j = \frac{b+a}{2} + \frac{b-a}{2}\,\xi_j \quad (j = 1, 2, ..., k)\,, \tag{55}$$

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 extrèmes 1 et -1 aux points :

$$\eta_j = \cos\left(\frac{j\pi}{k}\right) \quad (j = 0, 1, 2, ..., k)\,, \tag{56}$$

de sorte que :

$$-1 = \eta_k < \eta_{k-1} < ... < \eta_1 < \eta_0 = 1\,. \tag{57}$$

En conséquence, quand λ croit de a à b, la variable

$$\frac{b+a-2\lambda}{b-a}$$

décroit 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

$$\nu_j = \frac{b+a}{2} + \frac{b-a}{2}\,\eta_j \quad (j = 0, 1, 2, ..., k)\,, \tag{58}$$

avec (exactement) k alternances de signe. Soit alors le polynôme :

$$R(\lambda) = P^*(\lambda) - Q(\lambda)\,. \tag{59}$$

On a, quel que soit $j = 0, 1, 2, ..., k$:

$$|P^*(\nu_j)| = \frac{1}{A_k} = \|P^*\|_{\infty/[a,b]}\,, \tag{60}$$

et

$$|Q(\nu_j)| \leq \|Q\|_{\infty/[a,b]} < \|P^*\|_{\infty/[a,b]}\,, \tag{61}$$

la deuxième inégalité étant vraie par hypothèse. Il en résulte que

$$\text{signe}\,(R(\nu_j)) = \text{signe}\,(P^*(\nu_j)) \quad (j = 0, 1, 2, ..., k)\,. \tag{62}$$

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, (43), sont donc les inverses des zéros du polynôme optimal P^* :

$$\boxed{\tau_j = \frac{1}{\mu_j}\,.} \tag{63}$$

Le cycle itératif qui en résulte est connu sous le nom de "**Méthode itérative de Richardson**" ou "**accélération de Tchebychev**" [1].

Enfin, on peut mesurer la vitesse de convergence de l'itération par la quantité :

$$\boxed{v = -\frac{1}{k}\ln\left(\|P^*\|_{\infty/[a,b]}\right) = \frac{1}{k}\ln\left(A_k\right) = \frac{1}{k}\ln\left[\cosh(k\,\sigma)\right].} \tag{64}$$

La quantité $1/v$ représente le nombre moyen d'itérations (i.e. d'évaluations de la fonction $\mathbf{g}$) qu'il faut pour réduire le mode le plus lent à converger d'un facteur égal à ϵ. D'ailleurs, on a l'équivalence :

$$A_k \sim \frac{\exp\left(k\,\sigma\right)}{2} \quad (k \to \infty)\,, \tag{65}$$

de sorte que :

$$\begin{aligned} \lim_{k\to\infty}\, v &= \sigma \\ &= \cosh^{-1} c \\ &= \ln\left(c + \sqrt{c^2 - 1}\right)\,. \end{aligned} \tag{66}$$

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 :

$$\rho^* = \lim_{k \to \infty} (A_k)^{-\frac{1}{k}} \, . \tag{67}$$

On obtient donc l'expression suivante du **rayon spectral équivalent** de l'itération accélérée :

$$\boxed{\rho^* = \frac{1}{c + \sqrt{c^2 - 1}} \, .} \tag{68}$$

Cette méthode s'applique en particulier à la résolution itérative d'un système discrétisant un problème elliptique. A titre d'exemple, considérons le cas de la discrétisation classique sur maillage uniforme du problème unidimensionnel :

$$\begin{cases} -u_{xx} = f \, , & 0 \leq x \leq 1 \, , \\ u(0) = c_0 \, , & u(1) = c_1 \, , \end{cases} \tag{69}$$

à savoir :

$$\boxed{\begin{array}{c} A\,u = b \, , \text{ où :} \\ \\ A = \text{Trid } (-1, 2, -1) \, , \\ \\ u = (u_1, u_2, ..., u_M)^T \, , \\ \\ b = (f_1, f_2, \ldots, f_M)^T \, \Delta x^2 + (c_0, 0, \ldots, 0, c_1)^T \, . \end{array}} \tag{70}$$

Dans ce cas, où des conditions de Dirichlet sont appliquées aux deux bords, les valeurs propres de la matrice réelle, tri-diagonale, symétrique, définie-positive A sont bien connues :

$$\lambda_m = 2 - 2 \cos \theta_m \, , \quad (m = 1, 2, ..., M) \, , \tag{71}$$

où le paramètre de "fréquence" θ_m est donné par

$$\theta_m = \frac{m\pi}{M+1} \, , \quad (m = 1, 2, ..., M) \, . \tag{72}$$

Par conséquent,

$$\boxed{\forall m = 1, 2, ..., M \; : \; 0 \leq \lambda_m \leq 4 \, .} \tag{73}$$

On va maintenant examiner plusieurs choix possibles des paramètres a et b. Dans chaque cas, on calcule c par (48), A_k par (53) et v par (64).

1er essai : $a = 0$, $b = 4$.
Alors quel que soit k, $c = 1 = \eta_0$, $A_k = 1$ et

$$\boxed{\upsilon = 0\,.} \tag{74}$$

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.

2ème 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 :

$$c = \frac{\kappa+1}{\kappa-1} = 1 + \frac{2}{\kappa} + O\left(\frac{1}{\kappa^2}\right) \tag{75}$$

où

$$\kappa = \frac{\lambda_M}{\lambda_1} = \tan^{-2}\left(\frac{\pi}{2(M+1)}\right) = O(M^2) \gg 1\,, \tag{76}$$

est le nombre de conditionnement du système discret. Quelques calculs de développements limités permettent de tirer de (53) l'expression suivante :

$$A_k = 1 + \frac{2\,k^2}{\kappa} + O\left(\frac{1}{\kappa^2}\right)\,, \tag{77}$$

ce qui donne finalement d'après (64) :

$$\boxed{\upsilon = \frac{2\,k}{\kappa} + O\left(\frac{1}{\kappa^2}\right).} \tag{78}$$

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.

3ème essai :
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

$$a = 2\,,\ b = 4\,, \tag{79}$$

ce qui donne dans (48) :

$$c = 3\,. \tag{80}$$

D'où,

$$\sigma = \cosh^{-1}(3) = \ln\left(3 + \sqrt{8}\right), \tag{81}$$

$$A_k = \frac{(3+\sqrt{8})^k + (3-\sqrt{8})^k}{2}\,, \tag{82}$$

et enfin,

$$\boxed{\upsilon = \frac{1}{k}\,\ln\left(\frac{(3+\sqrt{8})^k + (3-\sqrt{8})^k}{2}\right),} \tag{83}$$

de sorte que

$$\lim_{k\to\infty}\ \upsilon = \ln\left(3+\sqrt{8}\right) \approx 1.256\,. \tag{84}$$

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 1, l'expression du polynôme $T_k(x)$, et les valeurs de A_k et υ pour $k = 1,\ 2,\ 3,\ 6$ et ∞.

k	$T_k(x)$	A_k	υ
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 1: Vitesse de convergence de la méthode de Richardson appliquée seulement à la partie "haute-fréquence" du spectre

On constate en particulier, qu'un cycle de trois 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 sensé être trivial. Les transferts inverses sont des prolongements. Moralement, ils introduisent des erreurs de type "haute-fréquence", car ce sont des composantes selon ceux 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.

Au passage, on peut évaluer la complexité de la méthode multi-grille. Pour cela, on note la grille la plus fine $\mathcal{M}_h$ (où $h = \Delta x$) et on suppose que le nombre de degrés de liberté qui lui sont associés est donné par

$$M = M_0 = 2^I \, \mu - 1 \,, \tag{85}$$

où I et μ sont des entiers strictement positifs. Les grilles suivantes dites "grossières", sont notées $\mathcal{M}_{2h}$, $\mathcal{M}_{4h}$, etc, et le nombre de degrés de liberté qui leur sont associés sont respectivement $M_1 = 2^{I-1} \, \mu - 1$, $M_2 = 2^{I-2} \, \mu - 1$, etc. Pour réduire les modes de haute-fréquence d'un facteur égal à 1/99, il suffit d'effectuer 6 itérations par niveau de grille (3 après une restriction, 3 après un prolongement). Soit C_0 le coût de ces itérations sur la grille fine, $\mathcal{M}_h$. Si on suppose que ce coût est proportionnel au nombre de degrés de liberté, ce qui est le cas lorsque la matrice G a une structure bande (sur tous les niveaux), le coût sur la grille $\mathcal{M}_{2h}$ est $C_0/2$, sur la grille $\mathcal{M}_{4h}$ est $C_0/4$, etc. Le coût total du V-cycle est donc :

$$C_V = C_0 \, \left(1 + \frac{1}{2} + \frac{1}{4} + \ldots \right) = 2 \, C_0 = O(M) \,. \tag{86}$$

D'autre part, il est naturel de réduire le résidu itératif à un niveau comparable à l'erreur d'approximation. Puisque le V-cycle a un rayon spectral théorique indépendant de M, à savoir $\rho = 1/99$, le nombre p d'applications du V-cycle nécessaires à réduire le résidu itératif à un niveau de l'ordre de $O(\frac{1}{M^2})$ est donné par :

$$\rho^p = O(\frac{1}{M^2}) \,, \tag{87}$$

de sorte que

$$p = O(\log M) \,, \tag{88}$$

et le coût total du calcul en résulte :

$$\boxed{C_{MG} = p\, C_V = O(M\ \log M)\,.} \tag{89}$$

Examinons maintenant la complexité de la stratégie de "multi-grille complète". Le résidu initial lors du dernier cycle multi-grille (celui qui utilise toutes les grilles $\mathcal{M}_h$, $\mathcal{M}_{2h}$, $\mathcal{M}_{4h}$, etc) est de l'ordre de l'erreur d'approximation sur la grille immédiatement plus grossière, $\mathcal{M}_{2h}$, donc 4 fois trop grand si l'approximation est d'ordre deux. Plus généralement pour une approximation d'ordre quelconque α, il suffit de réduire ce résidu d'un facteur constant égal à $(2h)^\alpha/h^\alpha = 2^\alpha$. Pour cela, un nombre constant de V-cycle est suffisant. Le coût total de l'application du dernier cycle, C_M, est donc proportionnel au nombre de degrés de liberté sur la grille fine. Pour le V-cycle immédiatement antérieur, le coût est moindre d'un facteur 2, et pour l'antépénultième d'un facteur 4, etc. Le coût total est donc :

$$\boxed{C_{MGC} = C_M\ \left(1 + \tfrac{1}{2} + \tfrac{1}{4} + \ldots\right) = 2\, C_M = O(M)\,.} \tag{90}$$

La complexité de la méthode multi-grille complète est donc proportionnelle au nombre de degrés de liberté (sur la grille fine).

4.2 Cas d'un spectre complexe

On ne connait pas la solution générale du problème du min-max, (43), lorsque le domaine Ω est quelconque.

Cependant, considérons toutes les ellipses dont l'équation est de la forme :

$$\frac{(d-x)^2}{a^2} + \frac{y^2}{a^2 - c^2} = 1\,. \tag{91}$$

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 (43) est connue grâce à un résultat du à Manteufel [4].

Malheureusement, on vera que, dans les applications qui nous concernent, cette dernière condition n'est pas réalisée, rendant le résultat de Manteufel inutilisable. Pour cette raison, on réfère à [4], pour une explicitation complète du théorème.

5 Applications

5.1 La technique de sur-relaxation appliquée au schéma prédicteur-correcteur de MacCormack

En 1969, MacCormack a proposé son célèbre schéma prédicteur-correcteur [5]. Dans le cas du problème mono-dimensionnel,

$$u_t + \partial/\partial x\, f(u) = 0\,, \tag{92}$$

ce schéma s'écrit :

Prédicteur :

$$u^{\overline{n+1}} = u^n - \frac{\Delta t}{\Delta x}\,\Delta f\,(u^n) \tag{93}$$

Correcteur :

$$u^{n+1} = \frac{1}{2}\left\{u^n + u^{\overline{n+1}} - \frac{\Delta t}{\Delta x}\,\nabla f\left(u^{\overline{n+1}}\right)\right\} \tag{94}$$

où les symboles Δ et ∇ représentent respectivement les opérateurs (du premier ordre) de différences avancée et rétrograde. Ce schéma a les propriétés suivantes :

- il est explicite et très simple à implémenter, et nécessite peu de place mémoire;
- on ne manipule que des flux, aucun jacobien (à l'inverse du schéma de Lax-Wendroff), et sans décomposition (concept apparu plus tard, du moins dans la littérature occidentale);
- dans le cas linéaire, le schéma s'identifie formellement au schéma de Lax-Wendroff, il est donc précis au second ordre en espace et en temps, et soumis aux mêmes limitations de stabilité.

L'ingénuosité de la construction réside dans le fait qu'en permutant la direction de décentrement, *de facto*, on réalise au correcteur des différences centrées, d'où la précision en espace. La précision en temps est le résultat d'une intégration de type Runge-Kutta d'ordre deux. Dans le cas d'une équation parabolique, par exemple lorsque

$$f(u) = c\,u - \sigma\,u_x\,, \tag{95}$$

on approche u_x par une différence rétrograde au prédicteur, et une différence avancée au correcteur, de manière à former dans les deux cas l'opérateur

$$\Delta\,\nabla = \nabla\,\Delta = \text{Trid}\,(1, -2, 1)\,, \tag{96}$$

c'est-à-dire l'opérateur (centré) de différence seconde, δ_{xx}. Par sa simplicité et sa précision, ce schéma s'applique donc aux problèmes hyperboliques et paraboliques, pour la recherche de solutions stationnaires ou instationnaires. Il a donc été très largement et peut-être abusivement utilisé dans les années 70 pour résoudre aussi bien les équations d'Euler que de Navier-Stokes.

Plaçons-nous maintenant dans le cas où l'on cherche la solution stationnaire de l'équation de la chaleur,

$$u_t = \sigma\, u_{xx}\,, \tag{97}$$

par application du schéma qui prend alors la forme :

Prédicteur :

$$\overline{u^{n+1}} = u^n + \mu\, \delta_{xx} u^n \tag{98}$$

Correcteur :

$$u^{n+1} = \frac{1}{2}\left\{u^n + \overline{u^{n+1}} + \mu\, \delta_{xx} \overline{u^{n+1}}\right\} \tag{99}$$

où l'on a posé

$$\mu = \frac{\sigma\, \Delta t}{\Delta x^2}\,. \tag{100}$$

On montre facilement qu'à un changement de notation près, ce schéma équivaut à l'algorithme prédicteur-correcteur (38)-(39) dans le cas particulier où :

$$A = -\,\mu\, \delta_{xx}\,,\ \omega_1 = \frac{1}{2}\,,\ \omega_2 = 1 \tag{101}$$

(i.e. u^{n+1} donné par (99) est égal à u^{n+2} donné par (39)). Il s'agit donc d'un cycle de deux pseudo-pas de temps $\tau_r \pm i\,\tau_i$, où d'après (41) :

$$\tau_r = \tau_i = \frac{1}{2}\,, \tag{102}$$

ce qui équivaut à annihiler les valeurs propres $1 \pm i$. Or, dans le mode normal de fonctionnement de l'algorithme $\mu = 1/2$, et les valeurs propres de l'opérateur A sont réelles et dans l'intervalle $[0, 2]$. Par conséquent d'après les résultats de la section précédente, le cycle optimal de deux pas de temps est unique et correspond à l'annihilation des valeurs réelles $1 \pm 1/\sqrt{2}$. Le cycle équivalent à l'algorithme de MacCormack n'est donc pas optimal pour l'équation de la chaleur. Il en est donc de même pour l'équation d'advection-diffusion,

$$u_t + c\, u_x = \sigma\, u_{xx}\,, \tag{103}$$

lorsque le paramètre μ est grand par rapport au nombre de Courant,

$$\nu = \frac{c\, \Delta t}{\Delta x}\,. \tag{104}$$

En d'autres termes, pour l'équation d'advection-diffusion, on peut améliorer les propriétés itératives du schéma de MacCormack par un cycle de double sur(sous)-relaxation et ceci d'autant mieux que le nombre de Reynolds de maille,

$$Re_{\Delta x} = \frac{\nu}{\mu} = \frac{c\,\Delta x}{\sigma}\,, \tag{105}$$

est petit.

En 1976-77, Désidéri et Tannehill [6] ont précisément introduit un tel algorithme. Bien que leur justification ait été différente, ils ont abouti à la même conclusion, et ont mis en évidence, notamment par un calcul d'écoulement hypersonique visqueux (équations complètes de Navier-Stokes), que des gains en temps substantiels pouvaient être réalisés de la sorte. Bien que le schéma explicite de MacCormack ait aujourd'hui perdu de son intérêt du fait de la plus grande efficacité des méthodes implicites, il est intéressant de noter que cette étude ancienne peut s'interpréter comme une technique d'annihilation.

5.2 Application à un schéma implicite en approximation centrée

On considère ici le cas de l'algorithme implicite d'Euler appliquée à l'équation d'advection,

$$u_t + c\,u_x = 0\,, \tag{106}$$

qui s'écrit :

$$\left(I + \nu\,\delta_x^C\right)\left(u^{n+1} - u^n\right) = -\,\nu\,\delta_x^C\,u^n\,, \tag{107}$$

où ν est à nouveau le nombre de Courant, et δ_x^C est l'opérateur de différence centrée :

$$\delta_x^C = \text{Trid}\,\left(-\frac{1}{2}, 0, \frac{1}{2}\right)\,, \tag{108}$$

dont les première et dernière lignes dépendent des procédures aux limites, dont on suppose ici qu'elles ne détruisent pas la structure tri-diagonale de la matrice. (Ici Δt est le vrai pas de temps et non le paramètre d'annihilation.)

Dans ce cas, la matrice G est pleine,

$$G = I - \nu\,\left(I + \nu\,\delta_x^C\right)^{-1}\,\delta_x^C\,, \tag{109}$$

mais son application effective se réalise par un algorithme d'inversion de système linéaire tri-diagonal de complexité $O(M)$. L'expression de la matrice A en résulte :

$$A = I - G = \nu\,\left(I + \nu\,\delta_x^C\right)^{-1}\,\delta_x^C\,, \tag{110}$$

et ses valeurs propres s'expriment simplement par :

$$\lambda_m = \frac{\nu\,\zeta_m}{1 + \nu\,\zeta_m}\,, \tag{111}$$

en fonction des valeurs propres $\{\zeta_m\}$ de l'opérateur δ_x^C, qui sont par exemple dans le cas périodique, données par

$$\zeta_m = i \, \sin\theta_m \,, \tag{112}$$

où le paramètre de fréquence a l'expression suivante :

$$\theta_m = \frac{2\,\pi\,m}{M} \,. \tag{113}$$

Lorsque le pas de temps de la méthode de base est grand, ν est grand, et quelles que soient les procédures aux bords, il résulte de (111) que le spectre de A forme un agrégat proche du point $z = 1$. Par conséquent, si on devait utiliser la technique d'annihilation, il serait naturel d'utiliser un pseudo-pas de temps τ égal à $1/z = 1$, ce qui équivaudrait à ne pas modifier l'algorithme de base.

Ceci explique pourquoi dans [7], les gains obtenus par cyclage du pas de temps ont été marginaux comparés à ceux réalisés par la simple utilisation d'un grand pas de temps.

5.3 Remarques sur un lisseur de type Runge-Kutta

On se place ici dans le cas où l'on résout l'équation d'advection, (106), par la méthode de Runge-Kutta 4 suivante :

$$\begin{array}{l} \text{Poser : } v^{(0)} = u^n \,, \text{ et faire :} \\ \left\{ \begin{array}{l} v^{(1)} = v^{(0)} - \alpha_1\,\nu\,\delta v^{(0)} \,, \\ v^{(2)} = v^{(0)} - \alpha_2\,\nu\,\delta v^{(1)} \,, \\ v^{(3)} = v^{(0)} - \alpha_3\,\nu\,\delta v^{(2)} \,, \\ v^{(4)} = v^{(0)} - \alpha_4\,\nu\,\delta v^{(3)} \,; \end{array} \right. \\ \text{poser : } u^{n+1} = v^{(4)} \,. \end{array} \tag{114}$$

Ici δ est l'opérateur de différence demi-décentrée,

$$\delta = \frac{1}{2}\left\{ \text{Trid}\left(-\frac{1}{2}, 0, \frac{1}{2}\right) + \text{Penta}\left(\frac{1}{2}, -2, \frac{3}{2}, 0, 0\right)\right\} = \text{Penta}\left(\frac{1}{4}, -\frac{5}{4}, \frac{3}{4}, \frac{1}{4}, 0\right) \tag{115}$$

(voir [9] pour une présentation détaillée des formes matricielles correspondantes, incluant des conditions aux bords raisonnables), et les coefficients $\{\alpha_i\}$ et le nombre de Courant ν sont ceux proposés par M. H. Lallemand dans sa thèse [8] pour optimiser les propriétés de lissage, c'est-à-dire d'atténuation des modes de haute-fréquence du schéma :

$$\alpha_1 = 0.10 \,,\ \alpha_2 = 0.26 \,,\ \alpha_3 = \frac{1}{2} \,,\ \alpha_4 = 1 \,,\ \nu = 1.7736 \,. \tag{116}$$

A la Figure 6, on a représenté les valeurs propres de l'opérateur δ dans le cas de conditions de Dirichlet-Neumann, et pour $M = 21$ ($\diamond$), et $M = 41$ ($+$). Ces valeurs propres sont distribuées sur un arc quasi-rectiligne, parallèle à l'axe des imaginaires, à l'abscisse $\Re(\lambda) \approx 3/4 = \lim_{M\to\infty} \text{Trace}\,(\delta)/M$. Les valeurs propres les plus proches de l'axe des réels

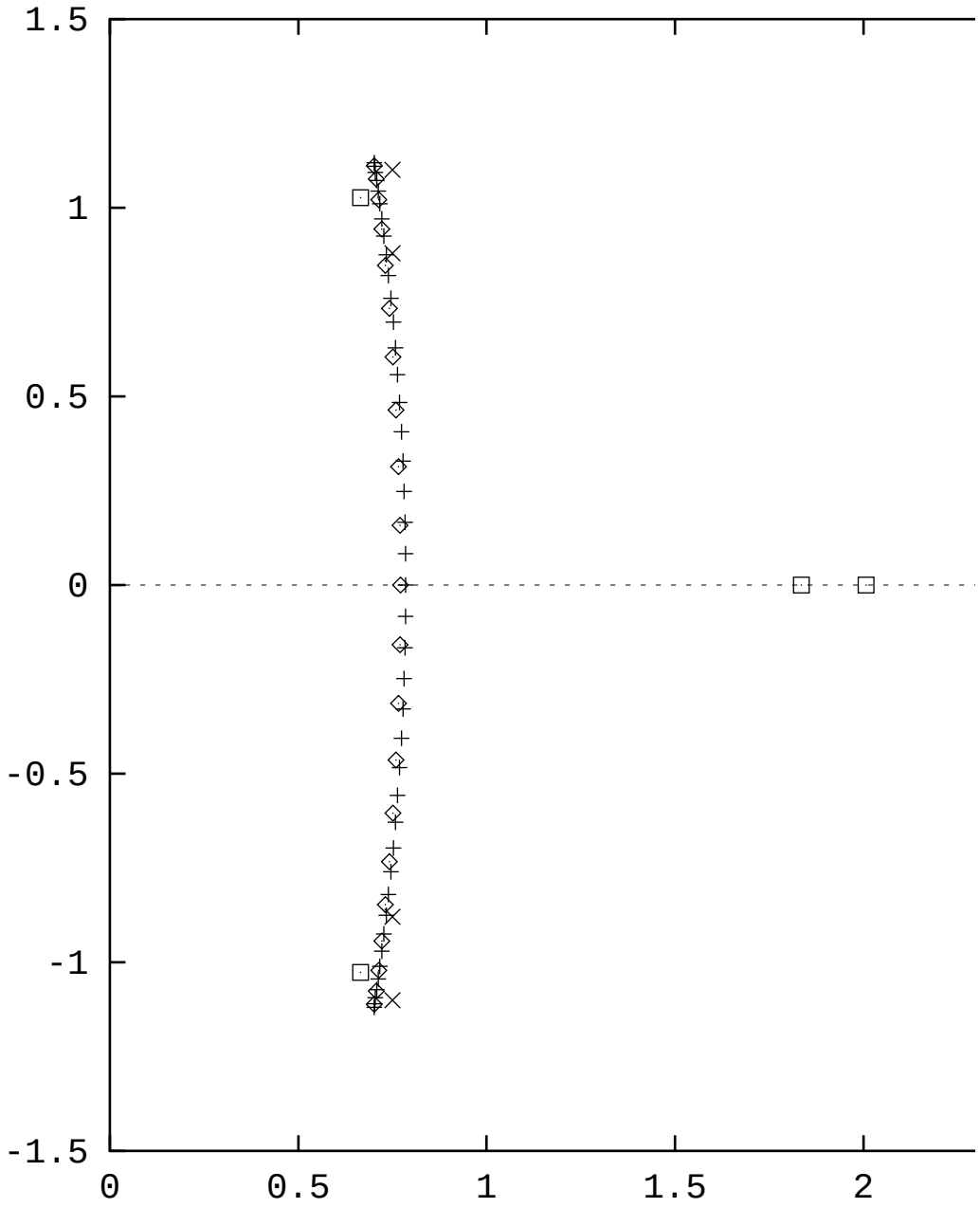

Figure 6: Valeurs propres de l'opérateur différence de Fromm, δ ($\diamond : M = 21$; $+ : M = 41$), et valeurs propres annihilées : *(i)* par le schéma de Runge-Kutta 4 à la limite de stabilité $\nu = \nu_{\max} \approx 1.9329$ ($\square$), et *(ii)* par l'algorithme alternatif de deux séquences prédicteur-correcteur ($\times$).

sont associées aux modes de basses fréquences; ce sont aussi les valeurs propres de plus faibles modules. Les valeurs propres associées aux modes de hautes fréquences sont celles, de plus grands modules, aux deux extrémités de l'arc. On constate qu'une valeur propre sur deux du spectre associé à la grille fine ($M = 41$) est très proche d'une valeur propre du spectre associé à la grille grossière ($M = 21$), et ceci est particulièrement visible pour les modes de basses fréquences.

Le facteur d'amplification de la méthode de Runge-Kutta proposée est donné par

$$g(z) = 1 - \alpha_4 z \, \{1 - \alpha_3 z \, [1 - \alpha_2 z \, (1 - \alpha_1 z)]\} \, , \tag{117}$$

où l'on a posé,

$$z = \lambda \, \nu \, , \tag{118}$$

et ses racines sont les complexes :

$$\left\{ \begin{array}{l} z_1 \approx 3.8797 \, , \\ z_2 \approx 3.5473 \, , \\ z_{3,4} \approx 1.2865 \pm 1.9835 \, i \, . \end{array} \right. \tag{119}$$

Lorsqu'on donne au nombre de Courant la valeur ν, on annihile de fait les valeurs propres

$$\lambda_m = \frac{z_m}{\nu} \, . \tag{120}$$

En particulier, à la limite de stabilité,

$$\nu = \nu_{\max} \approx 1.9329 \tag{121}$$

([8]), les valeurs propres annihilées sont :

$$\left\{ \begin{array}{l} \lambda_1 \approx 2.0072 \, , \\ \lambda_2 \approx 1.8352 \, , \\ \lambda_{3,4} \approx 0.6656 \pm 1.0262 \, i \, . \end{array} \right. \tag{122}$$

Ces valeurs sont représentées par le symbole $\Box$ à la Figure 6. Deux sont réelles, les deux autres tendent effectivement à "attaquer" les hautes fréquences. On voit aussi qu'une légère diminution du nombre de Courant en éloignant ces points de l'origine dans la direction des rayons polaires, a l'effet favorable de rapprocher les valeurs annihilées du spectre. Ceci confirme le bien fondé de la recommendation de M. H. Lallemand d'utiliser un pas de temps plus faible ($\nu \approx 1.7736$).

La Figure 6 révèle également que la forme même donnée au schéma de Runge-Kutta, (114), est réductrice. En effet, parmi les quatre valeurs propres annihilées, il en est deux, réelles, très éloignées du spectre. Il faut voir là une source d'inefficacité puisqu'on annihile seulement deux valeurs propres au prix de quatre calculs de flux. Comme alternative à coût

égal, on propose d'appliquer deux séquences "prédicteur-correcteur" telles que (38)-(39), la première pour annihiler les valeurs propres :

$$\lambda^*_{1,2} = 3/4 \pm 0.88\, i\,, \tag{123}$$

la seconde les valeurs :

$$\lambda^*_{3,4} = 3/4 \pm 1.10\, i\,. \tag{124}$$

Ces valeurs sont également représentées à la Figure 6 par le symbole $\times$. (Pour la réalisation de l'algorithme on renvoie à (40).) Ce choix fait, on connait le facteur d'amplification g_m d'un mode propre quelconque de l'opérateur δ associé à la valeur propre λ_m :

$$g_m = \prod_{k=1}^{4} \left(1 - \frac{\lambda_m}{\lambda^*_k}\right)\,. \tag{125}$$

Afin d'aprécier l'efficacité du l'algorithme, on a considéré le cas où $M = 41$ et ordonné la partie supérieure du spectre par valeurs croissantes de $\Im(\lambda_m) \geq 0$, $m = 0, 1, \ldots, 20$. A la Figure 7, la ligne continue passant par les symboles $\diamond$ donne la variation (croissante) de $|\lambda_m|$ en fonction de m. On identifie la plage $0 \leq m \leq 10$ aux basses fréquences, et la plage complémentaire $11 \leq m \leq 20$ aux hautes fréquences. Les lignes en pointillés indiquent la variation avec m du module du facteur d'amplification, $|g_m|$: *(i)* pour la méthode de Runge-Kutta 4, le nombre de Courant étant optimisé ($\nu = 1.7736$) (+), et *(ii)* pour l'algorithme de deux séquences prédicteur-correcteur ($\Box$). On constate que ce dernier est bien plus discriminatif des hautes fréquences, le facteur d'amplification étant uniformément inférieur à 2.4×10^{-2} sur la plage correspondante. Ce résultat fait de cet algorithme un meilleur "lisseur" potentiel.

Il faut néanmoins nuancer ce résultat du fait que la méthode de Runge Kutta précédente a été optimisée sur la base d'une analyse de Fourier, c'est à dire pour des conditions périodiques. On sait bien que le cas périodique est spectralement très différent du cas Dirichlet-Neumann, e.g. [9]. A la Figure 8, on a représenté les valeurs propres de l'opérateur différence de Fromm dans le cas périodique et pour $M = 21$ ($\diamond$) et $M = 41$ (+). On rappelle que les valeurs propres de l'opérateur de différence centrée sont distribuées sur l'intervalle $[-i, i]$ de l'axe des imaginaires, et celles de l'opérateur différence du premier ordre sur le cercle de centre $z = 1$ et de rayon 1. Si on moyennait ces deux schémas, on obtiendrait un spectre sur une courbe fermée symétrique par rapport à l'axe des réels et tangente à l'origine à l'axe des imaginaires. C'est aussi ce que l'on obtient à l'ordre deux, à la différence près que le contact de la courbe avec l'axe des imaginaires augmente avec l'ordre de précision; à l'ordre infini, c'est à dire dans le cas continu, le spectre est purement imaginaire. Les valeurs propres annihilées par la méthode de Runge Kutta 4 (ici dans le cas d'un nombre de Courant optimisé $\nu \approx 1.7736$), sont aussi indiquées sur la figure, par le symbole $\Box$. Il apparait clairement que les deux valeurs propres réelles ont été ajustées pour attaquer la plus haute fréquence associée à $\lambda = 2$. Malheureusement, les valeurs optimisées (123)-(124) ne paraissent plus du tout adaptées à la partie hautes-fréquences de ce spectre. A la Figure 9, on a représenté le module de la valeur propre générique, $|\lambda_m|$, de l'opérateur différence de Fromm, δ, dans le cas

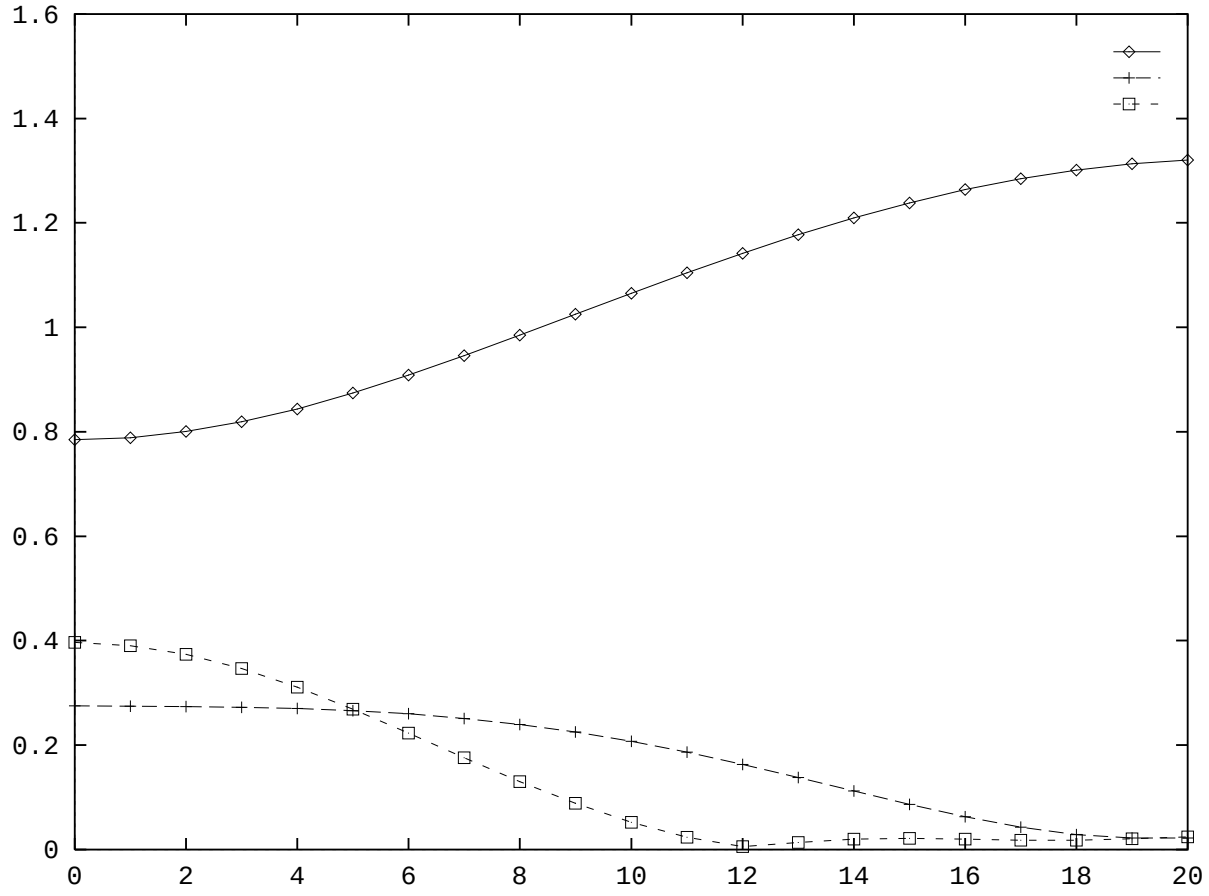

Figure 7: Module de la valeur propre générique, $|\lambda_m|$, de l'opérateur différence de Fromm, δ ($M = 41$) ($\diamond$), et module du facteur d'amplification, $|g_m|$, en fonction de l'indice m : *(i)* pour la méthode de Runge-Kutta 4 avec le nombre de Courant optimisé $\nu = 1.7736$ ($+$), et *(ii)* pour l'algorithme alternatif de deux séquences prédicteur-correcteur ($\Box$).

périodique ($M = 41$), ($\diamond$), et module du facteur d'amplification, $|g_m|$, en fonction de l'indice m : *(i)* de la méthode de Runge-Kutta 4 optimisée ($\nu = 1.7736$) ($+$), et *(ii)* de l'algorithme de deux séquences prédicteur-correcteur optimisé pour des conditions de Dirichlet-Neumann ($\Box$). A nouveau, quand m croit, la fréquence augmente ainsi que le module de λ_m. On constate que la méthode de Runge Kutta 4 n'est pas très discriminative des hautes fréquences.[1] L'algorithme de deux séquences prédicteur-correcteur optimisé pour les conditions de Dirichlet-Neumann est pire qu'inadapté, en fait fortement instable aux hautes-fréquences. Le meilleur lisseur précédent est donc ici le pire, ce qui démontre la difficulté de construire un algorithme ayant de bonnes propriétés de lissage quelles que soient les conditions aux limites.

[1] La courbe en pointillés associée au symbole + est conforme à la Figure 2.19 de [8].

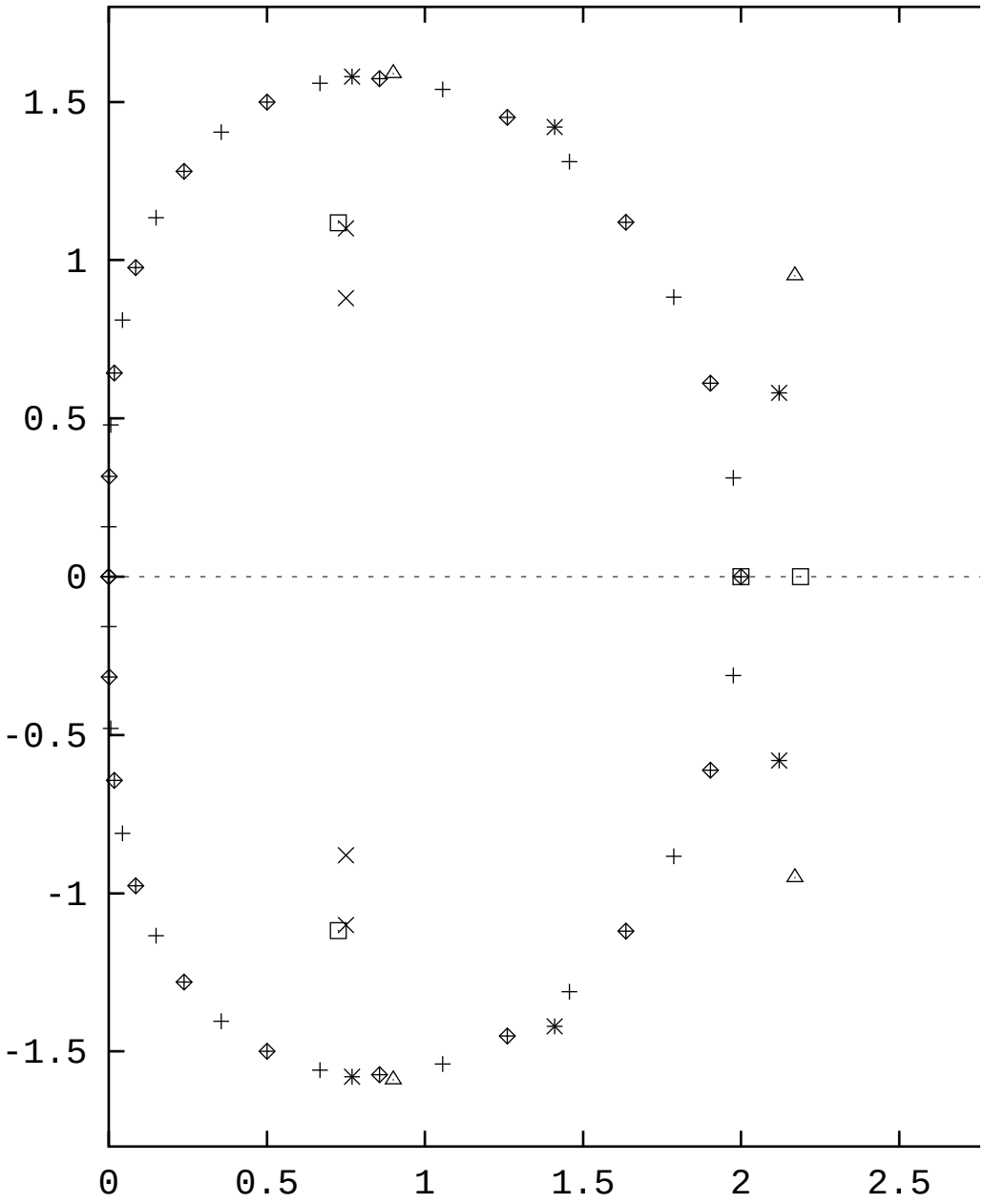

Figure 8: Valeurs propres de l'opérateur différence de Fromm, δ ($\diamond : M = 21$; $+ : M = 41$), dans le cas périodique et valeurs propres annihilées : *(i)* par le schéma de Runge-Kutta 4 optimisé ($\nu = 1.7736$) ($\Box$), *(ii)* par deux séquences prédicteur-correcteur optimisées pour des conditions de Dirichlet-Neumann, ($\times$), et *(iii)* optimisées pour des conditions périodiques, ($\triangle$), et *(iv)* par trois séquences prédicteur-correcteur optimisées ($*$).

Néanmoins, on peut adapter un algorithme de même structure au spectre périodique. Une optimisation sommaire a conduit au choix suivant de valeurs propres annihilées :

$$\begin{cases} \lambda_{1,2} = 0.90 \pm 1.59i\,, \\ \lambda_{3,4} = 2.17 \pm 0.95i\,. \end{cases} \tag{126}$$

Ces valeurs sont indiquées par le symbole $\triangle$ à la Figure 8. Le facteur d'amplification qui en résulte est représenté à la Figure 9 par la courbe en pointillés associée au symbole $\times$. La valeur maximale prise par le facteur d'amplification sur la plage des hautes fréquences est de 0.682 pour la méthode de Runge-Kutta 4 et de 0.186 pour l'algorithme d'annihilation des valeurs données ci-dessus en (126).

On peut améliorer encore le lisseur, en augmentant le coût. Par exemple avec trois séquences prédicteur-correcteur optimisées sommairement pour annihiler les valeurs suivantes :

$$\begin{cases} \lambda_{1,2} = 0.77 \pm 1.58i\,, \\ \lambda_{3,4} = 1.41 \pm 1.42i\,, \\ \lambda_{5,6} = 2.12 \pm 0.58i\,, \end{cases} \tag{127}$$

qui sont rerésentées à la Figure 8 par le symbole $*$, on obtient un facteur d'amplification représenté à la Figure 9 par la courbe en pointillés associée au symbole $\triangle$. La valeur maximale sur la plage des hautes fréquences est proche de 0.057. En termes de vitesse asymptotique, il est toujours meilleur d'augmenter le nombre de valeurs annihilées, puisqu'on optimise chaque nouvelle fois sur un plus grand domaine. Ce qui détermine précisément la séquence du schéma optimal c'est la tolérance que l'on fixe sur la valeur maximale du facteur d'amplification des hautes-fréquences. Cette tolérance doit se fixer en fonction de l'erreur de troncature, et le cas échéant du niveau estimé des non-linéarités.

Pour clore cette section, on résume les principales conclusions que l'on peut tirer :

- la structure du spectre de l'opérateur différence de Fromm (demi-décentrée) diffère fortement en périodique du cas où des conditions de Dirichlet-Neumann sont appliquées aux bords;
- dans les deux cas, le module de la valeur propre générique de cet opérateur augmente avec la fréquence;
- à coût égal, deux séquences prédicteur-correcteur optimisées réalisent un meilleur lisseur que la méthode optimisée de Runge-Kutta 4; malheureusement les paramètres définissant ces séquences dépendent fortement du cas étudié;
- la construction d'un bon lisseur s'avère plus difficile en périodique qu'avec des conditions de Dirichlet-Neumann; il semble qu'un bon lisseur est obtenu par deux séquences prédicteur-correcteur pour le cas Dirichlet-Neumann, et trois pour le cas périodique;

- en hyperbolique, un lisseur de qualité comparable au lisseur optimal du cas elliptique fait intervenir deux à trois fois plus de calculs de flux, et nécessite la mise en mémoire d'un vecteur supplémentaire.

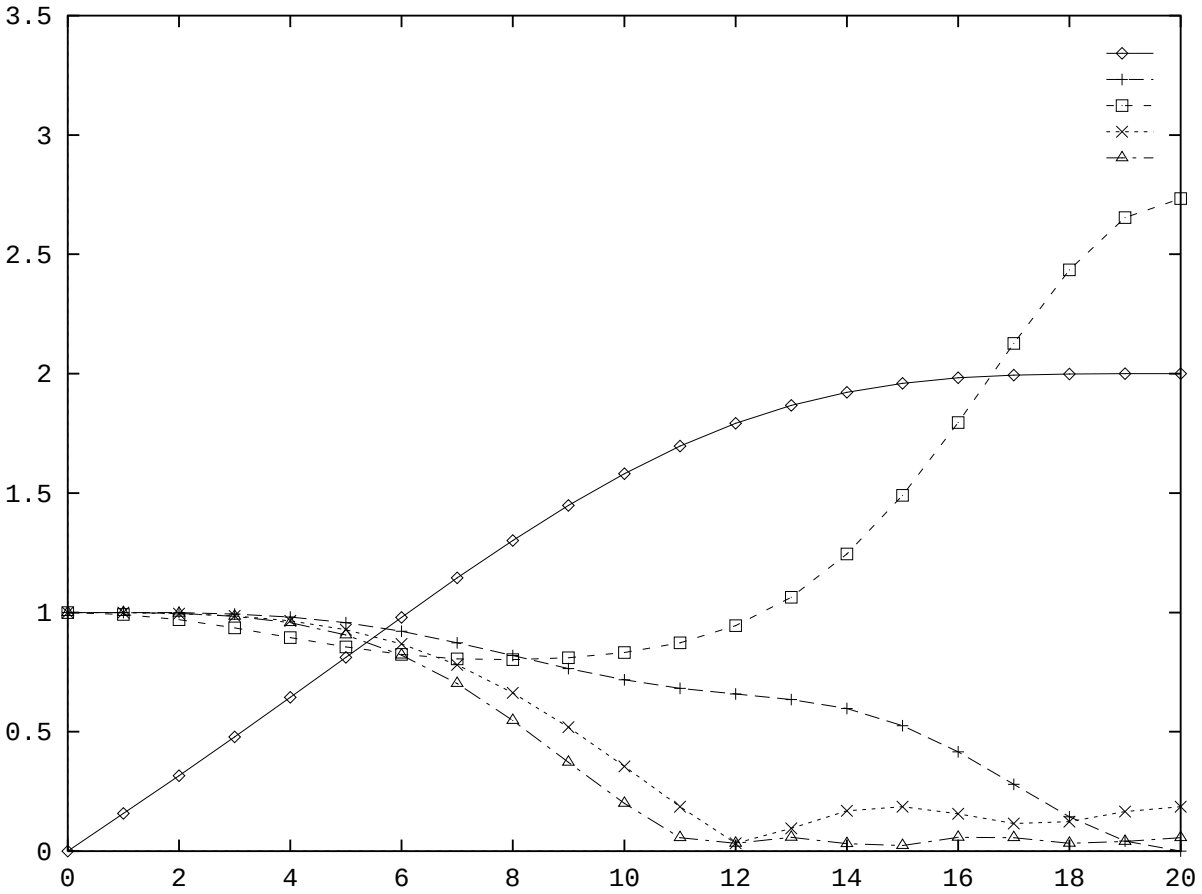

Figure 9: Module de la valeur propre générique, $|\lambda_m|$, de l'opérateur différence de Fromm, δ dans le cas périodique ($M = 41$), ($\diamond$), et module du facteur d'amplification, $|g_m|$, en fonction de l'indice m : *(i)* de la méthode de Runge-Kutta 4 optimisée ($\nu = 1.7736$) (+), *(ii)* de l'algorithme de deux séquences prédicteur-correcteur optimisé pour des conditions de Dirichlet-Neumann, ($\Box$), et *(iii)* optimisé pour des conditions périodiques, ($\times$), et *(iv)* de l'algorithme optimisé de trois séquences prédicteur-correcteur, ($\triangle$).

5.4 Application à la Méthode du Résidu-Corrigé

Dans cette section, on s'intéresse à l'application de la technique d'annihilation en boucle externe de la "Méthode Itérative du Résidu-Corrigé". Cette méthode a été examinée en profondeur dans [9] auquel on renvoie pour une description détaillée de modèles discrets

hyperboliques linéaires, périodiques ou non, en une et deux dimensions d'espace, et de leur étude par l'analyse de Fourier et l'analyse matricielle, dont on rappelle ici les principaux résultats.

5.4.1 Modèle Mono-Dimensionnel

On s'intéresse en particulier à un modèle mono-dimensionnel dans lequel on résout à nouveau l'équation d'advection, (106), par l'algorithme implicite d'Euler en approximation décentrée, de sorte que (107) est remplacé par :

$$\left(I + \nu\, \delta^U_{x,1}\right) \left(u^{n+1} - u^n\right) = -\,\nu\, \delta^\beta_{x,2}\, u^n\,, \tag{128}$$

où ν le nombre de Courant, $\delta^U_{x,1}$ est l'"opérateur de différence décentrée d'ordre un", qui pour $c > 0$, s'identifie à l'opérateur de différence rétrograde :

$$\delta^U_{x,1} = \nabla = \text{Trid}\ (-1, 1, 0)\ , \tag{129}$$

et $\delta^\beta_{x,2}$ est l'"opérateur de différence (partiellement) décentrée d'ordre deux" obtenu par la combinaison linéaire convexe suivante :

$$\delta^\beta_{x,2} = (1-\beta)\, \delta^C_x + \beta\, \delta^U_{x,2}\,, \tag{130}$$

où $\beta \in [0,1]$ est le "paramètre de décentrage", δ^C_x l'opérateur de différence centrée, (108), et $\delta^U_{x,2}$ l'opérateur de différence décentrée (ici rétrograde) du second-ordre :

$$\delta^U_{x,2} = \text{Penta}\ \left(\frac{1}{2}, -2, \frac{3}{2}, 0, 0\right)\ . \tag{131}$$

L'opérateur $\delta^\beta_{x,2}$ au membre de droite de l'algorithme (128) contrôle la précision de la solution stationnaire, alors que l'opérateur au membre de gauche, appelé "préconditionneur", garantit la stabilté inconditionnelle de l'intégration temporelle. Par le choix d'une approximation d'ordre un dans le préconditionneur, celui-ci est à diagonale dominante, et la résolution du système linéaire nécessaire à chaque pas de temps peut s'effectuer par relaxation. (Ceci est particulièrement important pour les extensions aux approximations sur des maillages non-structurés). Néanmoins, à convergence, la solution stationnaire est du second ordre et fonction de β. On s'interesse plus particulièrement aux valeurs suivantes de ce paramètre :

- $\beta = 0$: "Schéma Centré",
- $\beta = 1/3$: "Schéma du 3ème ordre de van Leer",
- $\beta = 1/2$: "Schéma demi-décentré de Fromm" (voir (115)),
- $\beta = 1$: "Schéma totalement décentré".

Lorsque $\nu \to \infty$ ("pas de temps infini"), (128) devient équivalent à l'algorithme suivant dit "**Méthode du Résidu-Corrigé**" :

$$\boxed{\delta^U_{x,1}\, u^{n+1} = \delta^U_{x,1}\, u^n - \delta^\beta_{x,2}\, u^n\,.} \tag{132}$$

Dans cet algorithme, non seulement le pas de temps a disparu, mais également la constante "physique" c. Ici, la matrice d'amplification G prend la forme suivante :

$$G = I - \left(\delta^U_{x,1}\right)^{-1}\, \delta^\beta_{x,2}\,, \tag{133}$$

et la matrice d'approximation (ici préconditionnée) :

$$\boxed{A = I - G = \left(\delta^U_{x,1}\right)^{-1}\, \delta^\beta_{x,2}\,.} \tag{134}$$

Les valeurs propres de la matrice G sont connues non seulement dans le cas périodique (analyse de Fourier) mais aussi dans le cas de conditions aux limites de Dirichlet-Neumann (analyse matricielle) où ce sont les nombres complexes suivants [9] :

$$\left\{ \begin{array}{l} g_0 = 0\,, \\ g_m = \frac{1}{2} - \beta + i\,\sqrt{\beta(1-\beta)}\,\cos\frac{m\pi}{M} \quad (m = 1, 2, \ldots, M-1)\,, \end{array} \right. \tag{135}$$

et celles de la matrice A s'en déduisent :

$$\boxed{\begin{array}{c} \lambda_0 = 1\,, \\ \lambda_m = \frac{1}{2} + \beta - i\,\sqrt{\beta(1-\beta)}\,\cos\frac{m\pi}{M} \quad (m = 1, 2, \ldots, M-1)\,. \end{array}} \tag{136}$$

La valeur propre $\lambda_0 = 1$ est associée au mode de plus basse fréquence qui est annihilé par une seule application de l'algorithme ($g_0 = 0$). Dans tout ce qui suit, il est sous-entendu que l'on se restreint à étudier l'élimination itérative des autres modes ($m = 1, 2, \ldots, M-1$). Les valeurs propres de la matrice G qui leur sont associées sont réparties sur un segment parallèle à l'axe des imaginaires à l'intérieur strictement du disque centré à l'origine de rayon $1/2$. Ce segment, à l'abscisse $\Re(g) = \frac{1}{2} - \beta$, est représenté à la Figure 10a.

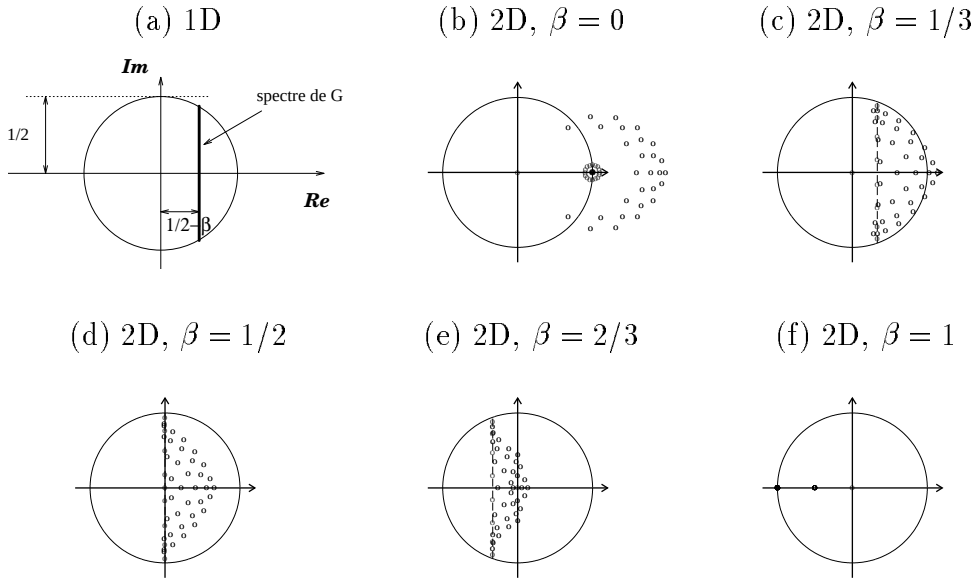

Figure 10: Spectre des valeurs propres de la matrice d'amplification G de la Méthode du Résidu-Corrigé ((a): 1D; (b-f): 2D).

Lorsque $\beta = 0$ ou 1, le schéma est centré ou totalement décentré, et $M - 1$ valeurs propres de la matrice d'amplification G sont confondues. Cette matrice est alors défective (non diagonalisable), et la convergence est pathologique. On se place désormais dans le cas inverse où

$$0 < \beta < 1\,, \tag{137}$$

pour lequel la convergence asymptotique est contrôlée par le rayon spectral :

$$\rho = |g_1| = \frac{1}{2}\sqrt{1 - 4\beta(1-\beta)\,\sin^2\frac{\pi}{M}} < \frac{1}{2}\,,\ \text{et}\ \approx \frac{1}{2}\,. \tag{138}$$

Asymptotiquement, les erreurs décroissent donc à peu près à la même vitesse que la suite 2^{-n}.

A la Figure 11, on a représenté le spectre $\sigma(A)$ (ou plutot le domaine Ω, ici un segment, qui le contient) dans le cas où $1/2 < \beta < 1$. Ce spectre est très localisé, et sa forme suggère que l'on cherche à accélérer le processus par annihilation des valeurs propres :

$$\lambda^*_{1,2} = \frac{1}{2} + \beta \pm r\,\sqrt{\beta(1-\beta)}\,i\,, \tag{139}$$

où $r \in [0,1]$, par l'utilisation séquentielle des pseudo-pas de temps conjugués suivants :

$$\tau_{1,2} = \frac{1}{\lambda^*_{1,2}}\,. \tag{140}$$

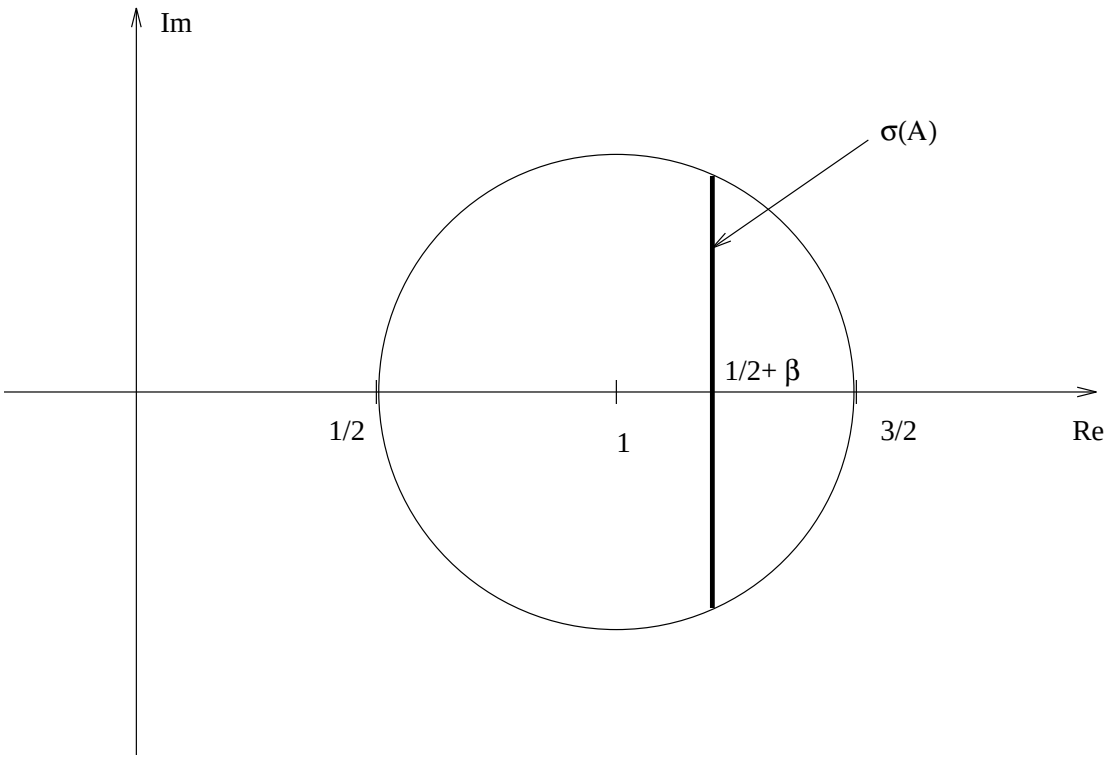

Figure 11: Spectre de la matrice A associée à la Méthode du Résidu Corrigé dans le cas où $1/2 < \beta < 1$.

Le mode associé à la valeur propre $\lambda_m \in \sigma(A)$ $(m = 1, 2, \ldots, M-1)$ est alors atténué par le facteur suivant :

$$g'_m = (1 - \lambda_m \tau_1)\,(1 - \lambda_m \tau_2)\ . \tag{141}$$

Or,

$$1 - \lambda_m \tau_{1,2} = \frac{(\gamma_m \pm r)\,\sqrt{\beta(1-\beta)}\,i}{\frac{1}{2} + \beta \pm r\,\sqrt{\beta(1-\beta)}\,i}\,, \tag{142}$$

où l'on a posé :

$$\gamma_m = \cos\frac{m\pi}{M}\,, \tag{143}$$

de sorte que

$$g'_m = \frac{(r^2 - \gamma_m^2)\,\beta(1-\beta)}{(\frac{1}{2} + \beta)^2 + r^2\,\beta(1-\beta)}\,. \tag{144}$$

Pour simplifier, remplaçons γ_m par un paramètre γ variant continûment de -1 à +1, et g'_m par :

$$g' = \frac{(r^2 - \gamma^2)\,\beta(1-\beta)}{(\frac{1}{2}+\beta)^2 + r^2\,\beta(1-\beta)}\,, \tag{145}$$

de sorte que la valeur optimale du paramètre r est la solution du problème suivant :

$$\min_{r\in[0,1]} \max_{\gamma\in[0,1]} |g'|\,. \tag{146}$$

Soit r^* cette valeur; il est évident que $r^* < 1$. Pour r fixé, le maximum de $|g'|$ est atteint à une limite : $\gamma = 0$ ou $\gamma = 1$; en conséquence, r^* est la valeur pour laquelle les valeurs correspondantes de g' sont opposées, ce qui donne :

$$\boxed{\begin{aligned} r^* &= \frac{1}{\sqrt{2}}\,, \\ g'_m &= \frac{1-2\gamma_m^2}{1+2\,\varpi^2} = -\frac{\cos\frac{2m\pi}{M}}{1+2\,\varpi^2} \quad (m = 1, 2, \ldots, M-1)\,, \end{aligned}} \tag{147}$$

où l'on a posé :

$$\boxed{\varpi = \frac{1+2\beta}{\sqrt{4\beta(1-\beta)}}\,.} \tag{148}$$

Le facteur d'atténuation du cycle a donc la valeur :

$$\max_{m=1,2,\ldots,M-1} |g'_m| = |g'_1| = \frac{\cos\frac{2\pi}{M}}{1+2\,\varpi^2}\,. \tag{149}$$

Ainsi, pour M infini, le rayon spectral équivalent (par calcul de flux) de l'itération est donc :

$$\boxed{\rho' = \frac{1}{\sqrt{1+2\,\varpi^2}}\,.} \tag{150}$$

A la Figure 12, on a représenté la variation de ρ' en fonction de β. On constate en particulier que :

$$\boxed{\lim_{\beta\to 0 \text{ ou } 1} \varpi = \infty\,,} \tag{151}$$

de sorte que :

$$\boxed{\lim_{\beta \to 0 \text{ ou } 1} \rho' = 0\,.} \tag{152}$$

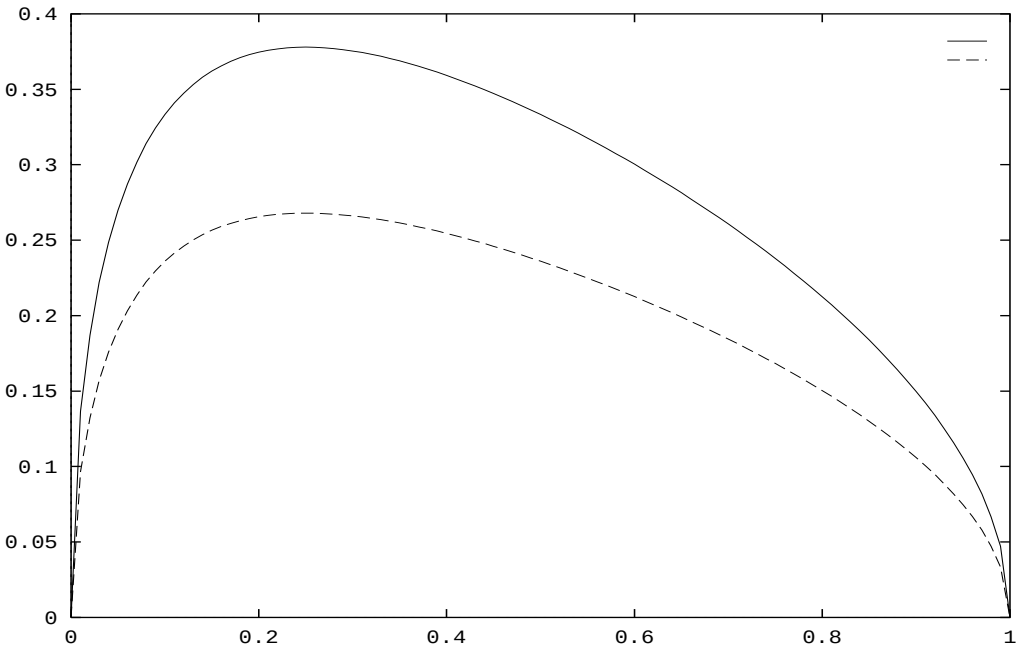

Figure 12: Variation des rayons spectraux ρ' et ρ^* en fonction de β (M infini).

Par conséquent, en introduisant une séquence prédicteur-correcteur de sur-relaxation optimisée en boucle externe de la Méthode Itérative du Résidu Corrigé, on obtient une itération dont le rayon spectral tend vers zéro lorsque le paramètre de décentrage β tend vers 0 ou 1. Ce résultat très favorable doit néanmoins être nuancé du fait que l'itération est défective lorsque β atteint l'une de ces deux limites. Il est donc recommandé d'utiliser une valeur intermédiaire du paramètre.

Enfin, à la vue de (147), on remarque que par l'application de cette technique d'annihilation, le spectre est transformé en une image réelle contenue dans l'intervalle $[-\rho'^2, \rho'^2]$. Par conséquent, la technique d'accélération de Tchebychev s'applique également. Or les valeurs propres de la "matrice A" équivalente associée à cette itération sont donc comprises entre $a = 1 - \rho'^2$ et $b = 1 + \rho'^2$. Il en résulte d'après (48) que :

$$c = \frac{1}{\rho'^2} \gg 1\,, \tag{153}$$

et d'après (68) le facteur d'atténuation de l'itération accélérée est donc :

$$\begin{aligned} g^* &= \frac{1}{1/\rho'^2 + \sqrt{1/\rho'^4 - 1}} \\ &= \frac{1}{1 + 2\varpi^2 + 2\varpi\sqrt{\varpi^2+1}} \\ &= \frac{1}{\left(\varpi + \sqrt{\varpi^2+1}\right)^2}, \end{aligned} \tag{154}$$

de sorte que le rayon spectral équivalent (par calcul de flux) de l'algorithme accéléré a pour expression :

$$\boxed{\rho^* = \sqrt{g^*} = \frac{1}{\varpi + \sqrt{\varpi^2+1}}.} \tag{155}$$

Ce paramètre est également représenté à la Figure 12.

A titre d'illustration considérons le cas où $\beta = 1/2$, pour lequel le conditionnement des vecteurs propres est optimal ([9]). On a alors, $\varpi = 2$, $\rho' = 1/3$ et $\rho^* = 1/(2+\sqrt{5}) \approx 0.236 < 1/2^2$. On a donc plus que doublé l'efficacité de l'algorithme. D'une manière générale, le gain en efficacité par rapport à l'algorithme de base se mesure par le facteur :

$$\boxed{f' = -\frac{\log\ \rho'}{\log\ 2}} \tag{156}$$

dans le cas où l'on n'applique pas l'accélération de Tchebychev, et autrement par le facteur

$$\boxed{f^* = -\frac{\log\ \rho^*}{\log\ 2}.} \tag{157}$$

Ces grandeurs sont représentées à la Figure 13. Cette figure suggère que les gains susceptibles d'être réalisés en augmentant le degré de décentrage β dans le but d'optimiser l'efficacité de la procédure d'accélération, ne semblent pas suffisants pour justifier de prendre le risque d'une dégradation importante du conditionnement du système (vis-à-vis de la diagonalisation).

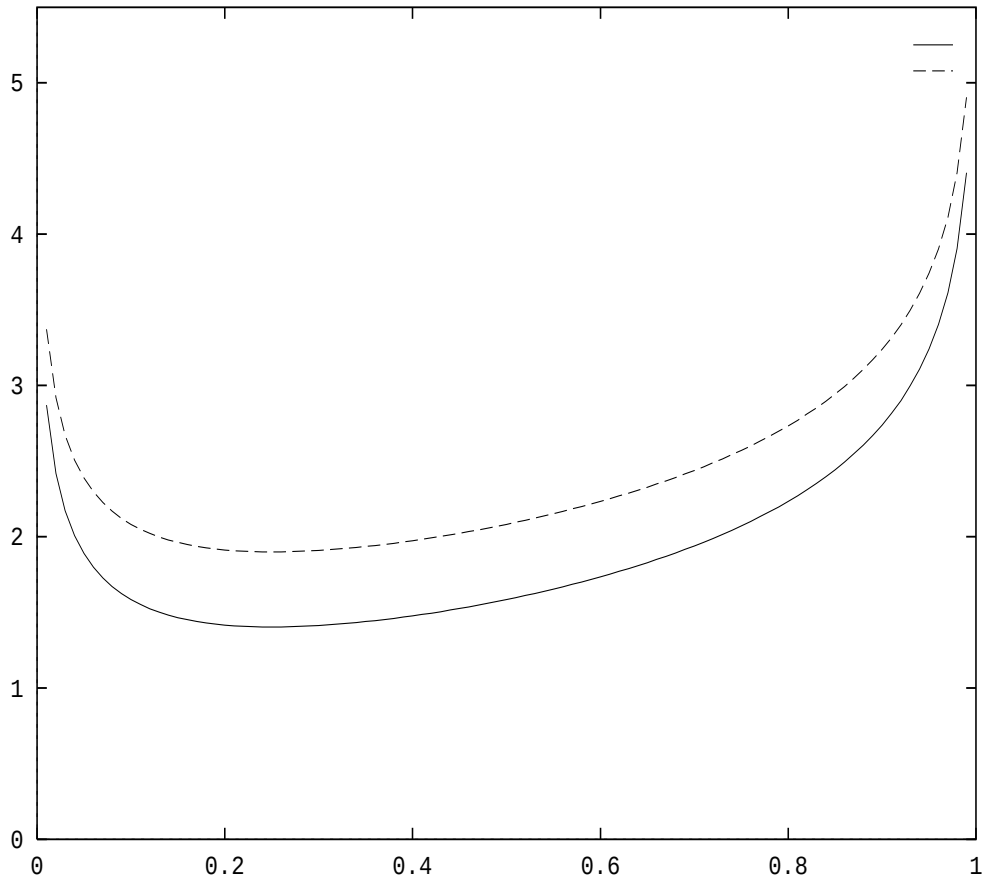

Figure 13: Gains en efficacité f' et f^* en fonction de β (M infini).

Comme alternative à l'application en boucle externe de l'accélération de Tchebychev, on propose l'algorithme d'annihilation séquentielle des valeurs propres suivantes :

$$\lambda_j^{\pm} = \frac{1}{2} + \beta \pm r_j \sqrt{\beta(1-\beta)}\, i \quad (j = 1, 2, \ldots, k)\,, \tag{158}$$

où les paramètres $r_j \in [0,1]$. On voit d'après (145) que l'optimisation de ces paramètres revient à résoudre le problème suivant :

$$\min_{r_1,\, r_2,\, \ldots\, r_k} \|P'(\gamma)\|_{\infty/[0,1]}\,, \tag{159}$$

où le polynôme $P'(\gamma)$ a la forme suivante :

$$P'(\gamma) = \prod_{j=1}^{k} \frac{r_j^2 - \gamma^2}{r_j^2 + \varpi^2} = (-1)^k \prod_{j=1}^{k} \frac{(\gamma + r_j)(\gamma - r_j)}{r_j^2 + \varpi^2}\,. \tag{160}$$

Ce problème est de même nature que (45). Ici, on optimise sur l'ensemble $\mathcal{P}'$ des polynômes pairs de degré $2k$ ayant des racines réelles et satisfaisant la condition de normalisation suivante :

$$P'(\pm\, i\,\varpi) = 1\,. \tag{161}$$

A cette normalisation près, le polynôme optimal est identique au polynôme de Tchebychev de degré $2k$:

$$P'(\gamma) = \frac{1}{A'_{2k}}\, T_{2k}(\gamma)\,, \tag{162}$$

où :

$$A'_{2k} = T_{2k}(i\,\varpi)\,. \tag{163}$$

La démonstration est semblable à celle de la Section 4.1 à quelques notations près. Les valeurs optimales des paramètres $\{r_j\}$ sont donc les suivantes :

$$\boxed{r_j^* = \cos\left(\frac{(2j-1)\pi}{4k}\right) \quad (j = 1, 2, \ldots, k)\,,} \tag{164}$$

et le rayon spectral du cycle complet est donné par :

$$\rho'_k = \frac{1}{|A'_{2k}|}\,. \tag{165}$$

En vertu de (50), la suite $\{A'_j\}$ $(j \in I\!N)$ définie par :

$$A'_j = T_k(i\,\varpi)\,, \tag{166}$$

est la solution du système suivant d'équations :

$$\left\{\begin{array}{l} A'_{j+1} + A'_{j-1} = 2\,i\,\varpi\, A'_j \quad (j \geq 1)\,, \\ \\ A'_0 = 1\,,\ A'_1 = i\,\varpi\,. \end{array}\right. \tag{167}$$

On en déduit facilement que :

$$A'_j = \frac{\exp(j\,\theta) + (-1)^j \exp(-j\,\theta)}{2}\, i^j\,, \tag{168}$$

où $\theta = \sinh^{-1}\,\varpi$, de sorte que

$$A'_{2k} = (-1)^k\, \cosh\,(2k\theta)\,, \tag{169}$$

et finalement :

$$\boxed{\rho'_k = \frac{1}{\cosh\,\left(2k\,\sinh^{-1}\,\varpi\right)}\,.} \tag{170}$$

Puisque la séquence fait intervenir $2k$ calculs de flux, le rayon spectral équivalent (par calcul de flux) est pour k grand, donné par :

$$\boxed{\rho_{\text{equiv}} = \lim_{k\to\infty} \left(\rho'_k\right)^{\frac{1}{2k}} = \frac{1}{\exp\left(\sinh^{-1}\varpi\right)} = \frac{1}{\varpi + \sqrt{\varpi^2+1}} = \rho^* .} \tag{171}$$

Ce résultat n'a rien d'étonnant ! Puisqu'on a trouvé que les valeurs à annihiler optimales sont à une transformation près les zéros du polynôme de Tchebychev du degré approprié, l'algoritme d'annihilation optimal est équivalent à l'accélération de Tchebychev appliquée à l'algorithme d'annihilation d'un seul couple (d'ailleurs arbitraire) de valeurs propres complexes conjuguées de la forme considérée.

5.4.2 Réalisation de l'Algorithme Mono-Dimensionnel Optimal

En résumé, l'algorithme d'annihilation optimal suppose le choix des paramètres suivants :

- β, paramètre de décentrage; on recommande $\beta = 1/2$ (valeur pour laquelle on observe la meilleure séparation possible des valeurs propres de l'algorithme de base, et en conséquence, un conditionnement optimal vis-à-vis de la diagonalisation), ou une valeur supérieure, néanmoins éloignée de 1, e.g. $\beta = 3/4$;

- ν, le nombre de Courant dans le cas où l'on donne à l'algorithme de base une formulation instationnaire; la condition $\nu \gg 1$ doit alors être vérifiée afin que cet algorithme ait effectivement les propriétés théoriques de la Méthode du Résidu-Corrigé;

- k, le nombre d'étapes d'annihilation; en principe, l'efficacité augmente avec ce paramètre; cependant, pour une plus grande simplicité algorithmique, et pour éviter des étapes de sur-relaxation excessive qui peuvent être instables en non-linéaire, on choisira sans doute une valeur "modéremment grande", e.g. $k = 2$, 3...

L'expression des valeurs propres annihilées, (158), et des valeurs optimales $\{r_j^*\}$, (164), permettent de calculer les paramètres de relaxation ω_1 et ω_2 à partir de (40). L'algorithme complet est constitué d'une phase d'annihilation du mode de plus basse fréquence (Phase 1) réalisée par un seul pas de l'algorithme de base, (128), et d'une phase d'annihilation des autres modes (Phase 2) réalisée par k séquences "prédicteur-correcteur" de la forme (38)-(39).

En définitive, une application de l'algorithme complet à partir de la solution initiale u^0 est donc réalisée par les deux phases suivantes :

- **Phase 1 :** Une application de l'algorithme de base (128) :

$$\boxed{\begin{array}{c}\left(I+\nu\,\delta^U_{x,1}\right)\,\delta u^1 = -\,\nu\,\delta^\beta_{x,2}\,u^0\,,\\ u^1 = u^0 + \delta u^1\,.\end{array}} \tag{172}$$

- **Phase 2 :** Pour $j = 1, 2, \ldots, k$, exécuter la séquence prédicteur-correcteur suivante :

Prédicteur :

$$\boxed{\begin{array}{c}\left(I+\nu\,\delta^U_{x,1}\right)\,\delta u^{2j} = -\,\nu\,\delta^\beta_{x,2}\,u^{2j-1}\,,\\ \omega_{1,j} = 1/(1+2\beta)\,,\\ u^{2j} = u^{2j-1} + \omega_{1,j}\,\delta u^{2j}\,;\end{array}} \tag{173}$$

Correcteur :

$$\boxed{\begin{array}{c}\left(I+\nu\,\delta^U_{x,1}\right)\,\delta u^{2j+1} = -\,\nu\,\delta^\beta_{x,2}\,u^{2j}\,,\\ \omega_{2,j} = 1/\left[\dfrac{1+2\beta}{4} + \dfrac{\beta(1-\beta)}{1+2\beta}\,\cos^2\,\dfrac{(2j-1)\pi}{4k}\right],\\ u^{2j+1} = u^{2j-1} + \omega_{2,j}\,\delta u^{2j+1}\,.\end{array}} \tag{174}$$

On remarque que le prédicteur correspond toujours à une phase stabilisatrice de sous-relaxation, alors que le positionnement de $\omega_{2,j}$ par rapport à 1 dépend de j et de β. On constate à nouveau, que cet algorithme nécessite la mise en mémoire d'un vecteur de plus que l'algorithme de base. Notons enfin que si dans le cas linéaire, la phase 1 s'effectue une fois pour toutes. en non-linéaire, elle doit être répétée à intervalles.

5.4.3 Extension au Modèle Bi-Dimensionnel

On s'intéresse ici au modèle hyperbolique fourni par l'équation d'advection linéaire, à coefficients constants, en deux dimensions d'espace :

$$u_t + a\,u_x + b\,u_y = 0\,, \tag{175}$$

des conditions de Dirichlet-Neumann appropriées étant appliquées aux bords.

L'application de la Méthode du Résidu-Corrigé à ce modèle a été étudiée dans [9] (auquel on renvoie pour les détails), et on sait en particulier, que les effets liés au caractère bidimensionnel du modèle sont les plus importants lorsque l'on a :

$$\frac{a}{\Delta x} = \frac{b}{\Delta y}\,, \tag{176}$$

ce que l'on suppose désormais. Les valeurs propres de la matrice d'amplification correspondante G sont représentées à la Figure 10b-f pour différentes valeurs du paramètre β. Ces valeurs propres forment un spectre constitué des valeurs propres du modèle monodimensionnel qui sont situées sur un segment parallèle à l'axe des imaginaires à l'abscisse $\Re(g) = \frac{1}{2} - \beta$, complémenté d'un nuage généralement à la droite de ce segment qui se déplace aussi vers la gauche lorsque β augmente. Par conséquent, dans le cas bi-dimensionnel, la Figure 11 ne représente qu'une partie du spectre de la matrice d'approximation A; les valeurs propres non représentées étant visiblement situées pour $\beta \geq \frac{1}{2}$, à l'intérieur du disque de centre $z = 1$ et de rayon $\frac{1}{2}$, et à gauche de la corde d'abscisse $\Re(\lambda) = \frac{1}{2} + \beta$. On conçoit que dans ce cas l'algorithme (173)-(174) n'est plus optimal.

On peut cependant chercher un algorithme efficace en enrichissant le jeu des valeurs propres annihilées du modèle mono-dimensionnel (158). La forme quasi-triangulaire de l'enveloppe convexe du spectre ainsi que sa structure pour le cas limite $\beta = 1$ suggèrent de rajouter par exemple, les ℓ triplets suivants :

$$\begin{cases} \mu_{j,1} = p_j + i\, q_j\,, \\ \mu_{j,2} = p_j - i\, q_j\,, \\ \mu_{j,3} = r_j\,, \end{cases} \tag{177}$$

$(j = 1, \ldots, \ell)$ où :

$$\begin{cases} \frac{1}{2} \leq p_j \leq \frac{1}{2} + \beta\,, \\ (p_j - 1)^2 + q_j^2 \leq \frac{1}{4}\,. \end{cases} \tag{178}$$

L'algorithme correspondant s'effectue ici en trois phases, sa réalisation est explicitée dans la section suivante dans un cas particulier. Les phases 1 et 2 sont identiques respectivement à (172) et à (173)-(174) à condition de remplacer les opérateurs différences mono-dimensionnels par leur homologues bidimensionnels (des sommes directes, [9]). La phase 3 comprend une suite de ℓ séquences prédicteur-correcteur dont les paramètres sont en vertu de (40) donnés

par :

$$\boxed{\begin{array}{c} \omega_{1,j} = \dfrac{1}{2\,p_j}\,, \\ \\ \omega_{2,j} = \dfrac{2\,p_j}{p_j^2 + q_j^2}\,; \end{array}} \tag{179}$$

mais chaque séquence est complétée d'une étape de sur(sous)-relaxation simple correspondant au paramètre suivant :

$$\boxed{\omega_{3,j} = \frac{1}{r_j}\,.} \tag{180}$$

L'algorithme optimal de ce type est la solution du problème suivant :

$$\min_{\ell,\{p_j\},\{q_j\}\,\{r_j\}\,(1\le j\le \ell)} \ \rho\,, \tag{181}$$

où

$$\rho = \max_{\lambda\,\in\,\sigma(A)-\{1\}} \ |H_{k,\ell}\,(\lambda)|^{1/(2k+3\ell)}\,, \tag{182}$$

où :

$$H_{k,\ell}\,(\lambda) = h_k\,(\lambda) \times \prod_{j=1}^{\ell} \prod_{\imath=1}^{3} \left(1 - \frac{\lambda}{\mu_{\imath,j}}\right)\,, \tag{183}$$

et

$$h_k\,(\lambda) = \prod_{j=1}^{k} \left(1 - \frac{\lambda}{\lambda_j^+}\right) \left(1 - \frac{\lambda}{\lambda_j^-}\right)\,, \tag{184}$$

est une fonction connue fixée par le choix des valeurs propres annihilées du cas monodimensionnel. On ne connait pas la solution de ce problème, d'ailleurs on ignore s'il est bien posé (même pour k fixé). Afin de proposer une recommandation pratique, on particularise désormais l'algorithme au cas où $k = 2$ et $\ell = 1$ (7 valeurs propres annihilées au total), pour lequel une optimisation sommaire a donné, pour $\beta = \frac{1}{2}$:

$$\left\{\begin{array}{l} p^* = 0.92\,, \\ q^* = 0.23\,, \\ r^* = 0.74\,. \end{array}\right. \tag{185}$$

Ces valeurs correspondent aux valeurs suivantes des paramètres de relaxation :

$$\left\{\begin{array}{l} \omega_1^* \approx 0.54\,, \\ \omega_2^* \approx 2.04\,, \\ \omega_3^* \approx 1.35\,. \end{array}\right. \tag{186}$$

omettant désormais l'indice j. Le rayon spectral équivalent (par calcul de flux) de l'algorithme qui en résulte est donné par :

$$\rho^*(1/2) \approx 0.332\,, \tag{187}$$

et le gain en efficacité par rapport à l'algorithme de base f^* est de l'ordre de 1.59 dans ce cas.

On a effectué une optimisation analogue pour les valeurs de β comprises entre $\frac{1}{2}$ et 1. On a représenté la variation avec β du triplet "optimal" (p^*, q^*, r^*) à la Figure 14, et des valeurs correspondantes du triplet $(\omega_1^*, \omega_2^*, \omega_3^*)$, du rayon spectral équivalent ρ^*, et enfin du gain en efficacité (par rapport à l'algorithme de base) $f^* = -\log \rho^* / \log 2$ respectivement aux Figures 15, 16 et 17.

En conclusion, on constate en particulier qu'un choix judicieux des paramètres de relaxation permet de réaliser un gain en efficacité qui tend vers l'infini quand β tend vers 1. Bien évidemment, la matrice d'amplification étant défective dans ce cas, on ne recommande pas d'opérer proche de cette limite. Néanmoins, il semble avantageux d'augmenter sensiblement le décentrage par rapport au schéma de Fromm ($\beta = \frac{1}{2}$). En particulier, la valeur théorique du gain en efficacité dépasse 2 lorsque β est supérieur à (une valeur proche de) $\frac{2}{3}$.

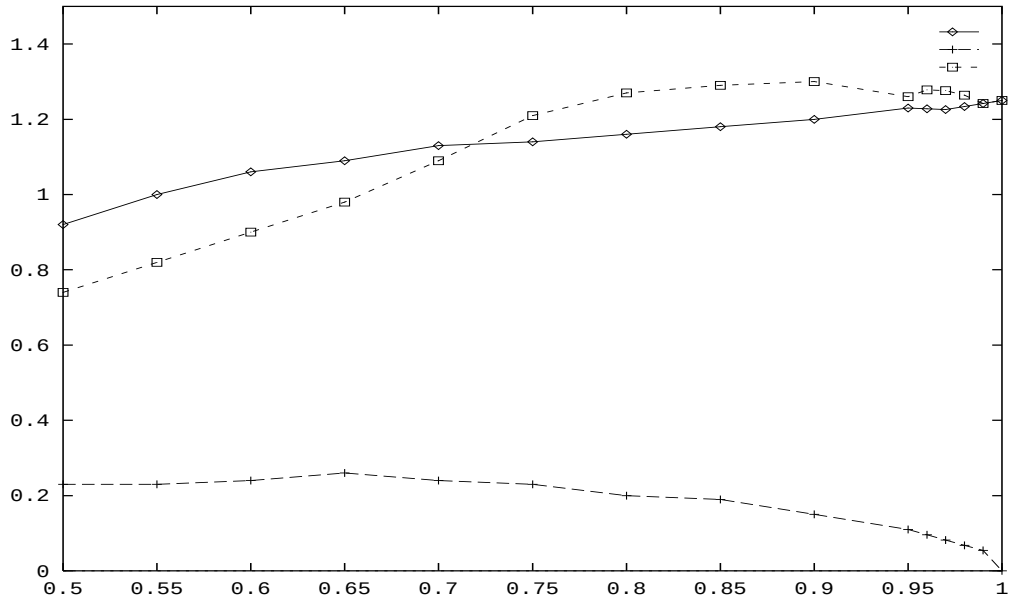

Figure 14: Variation des paramètres p^*, q^* et r^* en fonction de β.

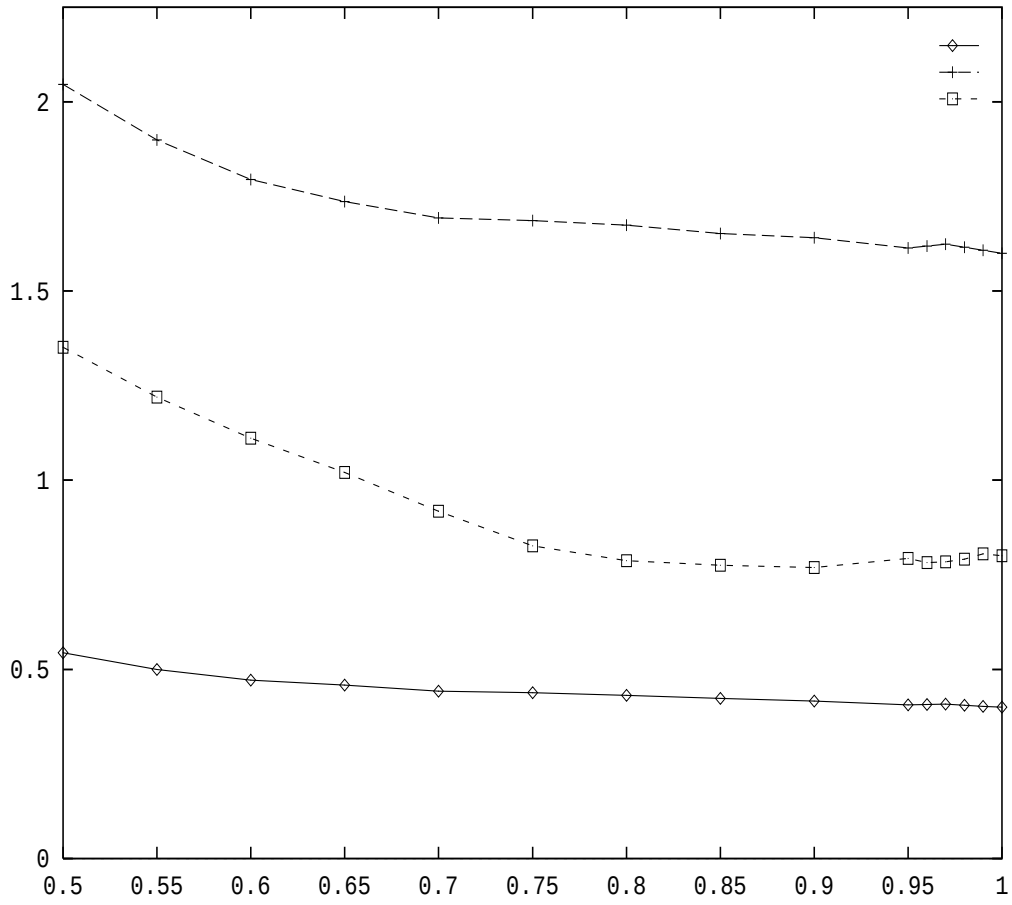

Figure 15: Variation des paramètres ω_1^*, ω_2^* et ω_3^* en fonction de β.

5.4.4 Algorithme Optimal du Cas Bi-Dimensionnel pour $\beta = \frac{2}{3}$

Dans le cas particulier où $\beta = \frac{2}{3}$ et $k = 2$, les valeurs optimales des paramètres de relaxation de la phase 2 sont les suivantes pour le modèle mono-dimensionnel :

$$\begin{aligned}(\omega_1, \omega_2) &= \left(3/7, \left[7/12 + 2/21 \cos^2 \tfrac{\pi}{8}\right]^{-1}\right) \text{ et } \left(3/7, \left[7/12 + 2/21 \cos^2 \tfrac{3\pi}{8}\right]^{-1}\right), \\ &\approx (0.4286, 1.5046) \text{ et } (0.4286, 1.6743) .\end{aligned} \tag{188}$$

Pour le modèle bi-dimensionnel, l'optimisation sommaire des paramètres de relaxation de la phase 3 a fourni les valeurs suivantes :

$$\begin{cases} \omega_1 \approx 0.45 , \\ \omega_2 \approx 1.725 , \\ \omega_3 \approx 1 . \end{cases} \tag{189}$$

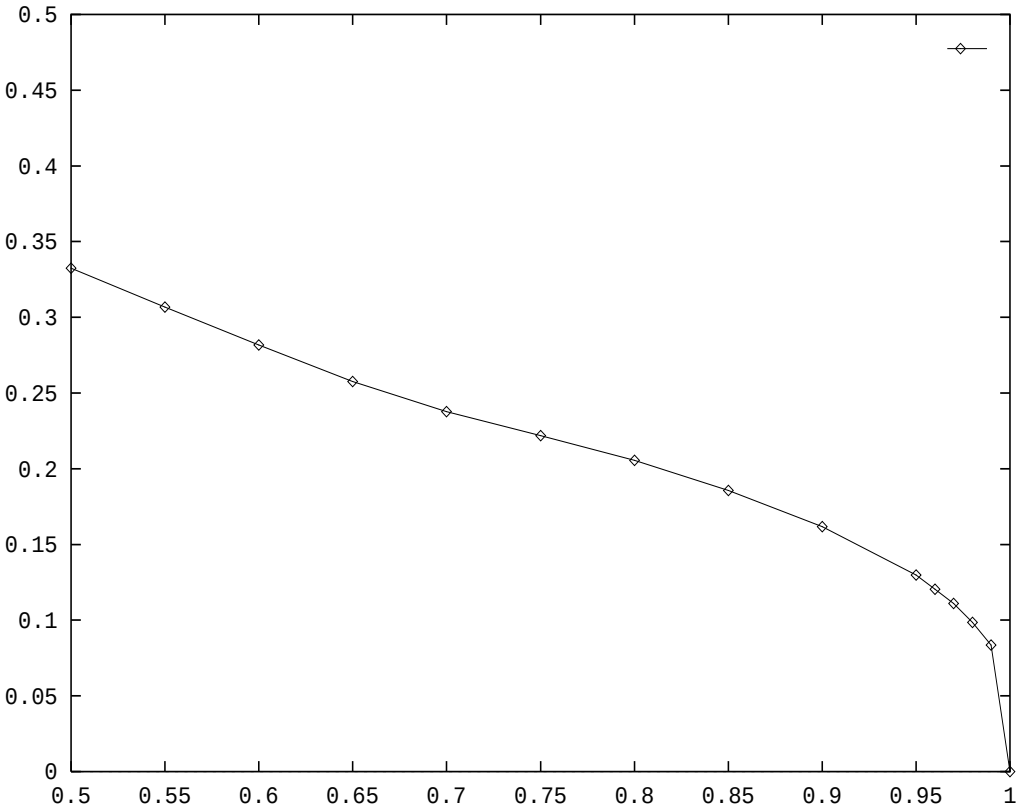

Figure 16: Variation du rayon spectral équivalent ρ^* en fonction de β.

Il en résulte, que l'étape de relaxation de la phase 3 est identique à l'application de l'algorithme de base, c'est-à-dire à la phase 1.

En résumé, on effectue la phase 1 ("algorithme de base") à chaque cycle; dans la phase 2, (173)-(174), on affecte successivement au couple (ω_1, ω_2) les trois valeurs suivantes :

$$(0.4286, 1.5046) \ , \ (0.4286, 1.6743) \ \text{ et } \ (0.45, 1.725) \ . \tag{190}$$

Le gain théorique en efficacité réalisé par cet algorithme comparé à l'algorithme de base est proche de 2. Ajoutons enfin que des gains supérieurs pourraient être atteints en enrichissant encore le jeu des valeurs propres annihiliées; on a choisi ici dans ce document préliminaire une stratégie de simplicité algorithmique.

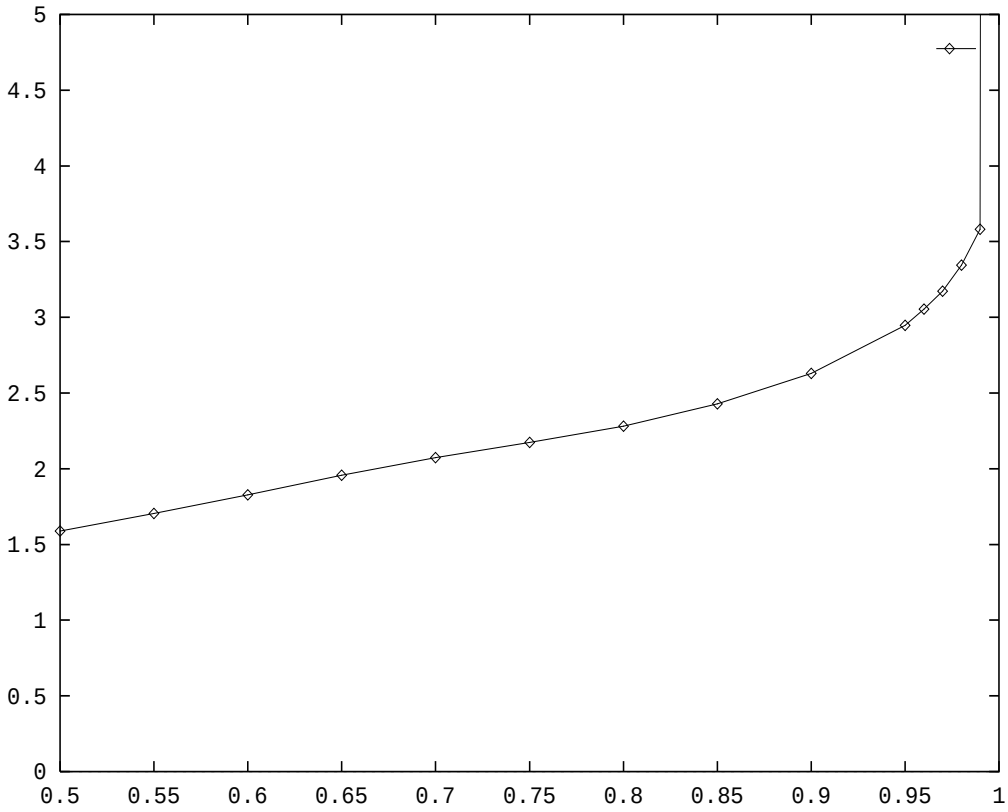

Figure 17: Variation du gain en efficacité f^* en fonction de β.

6 Conclusions

On a considéré un algorithme d'annihilation de modes propres qui dans sa forme la plus générale équivaut à brancher sur l'algorithme de base des boucles de sur(sous)-relaxation. Grâce à cette équivalence, lorsque les valeurs optimales des paramètres de relaxation ne sont pas connues, un choix raisonnable peut être fait après examen du spectre de la matrice d'approximation. De plus, par un choix adéquat des modes annihilés, on peut aussi adapter la construction du schéma itératif à certains besoins spécifiques. En particulier, un bon "lisseur" se construit par annihilation préférentielle des modes de hautes fréquences. Enfin, on a montré comment cette technique permettait d'obtenir certains gains théoriques en efficacité pour la méthode implicite du Résidu Corrigé.

Remerciements

L'auteur tient à remercier vivement Marie-Claude Ciccoli et Bruno Koobus du Projet SINUS pour leur lecture critique du manuscrit et leur aide efficace à la finalisation du document.

Bibliographie

[1] R. S. VARGA, "Matrix Iterative Analysis", *Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1962.*

[2] H. LOMAX, communication privée.

[3] D. JESPERSEN, communication privée.

[4] T. A. MANTEUFEL, "The Tchebychev Iteration for Nonsymmetric Linear Systems", *Numerical Mathematics, 28, 307-327, 1977.*

[5] R. W. MACCORMACK, "The Effect of Viscosity in Hypervelocity Impact Cratering", AIAA Paper No. 69-354, 1969.

[6] J. A. DESIDERI, "Over-Relaxation of a Finite-Difference Technique to Solve Fluid Flow Problems", *Master Thesis, Iowa State University, 1976*, et

J. A. DESIDERI, J. C. TANNEHILL, "Over-Relaxation Applied to the MacCormack Finite-Difference Scheme", *Journal of Computational Physics, Vol. 23, No. 3, March 1977.*

[7] J. A. DESIDERI, "On Improving the Iterative Convergence Properties of an Implicit Approximate-Factorization Finite-Difference Algorithms", *Ph. D. Thesis, Iowa State University, 1978*, et

J. A. DESIDERI, J. L. STEGER, J. C. TANNEHILL, "On Improving the Iterative Convergence Properties of an Implicit Approximate-Factorization Finite-Difference Algorithms", *NASA TM 78495, June 1978.*

[8] M. H. LALLEMAND, "Schémas Décentrés Multigrilles pour la Résolution des Equations d'Euler en Eléments Finis", Thèse Doctorale, Université de Provence, 1988.

[9] J. A. DESIDERI, P. W. HEMKER, "Analysis of the Convergence of Iterative Implicit and Defect-Correction Algorithms for Hyperbolic Problems", *Rapport INRIA N° 1200, Mars 1990.*