Raffinement d'éléments propres approchés d'un opérateur compact

Mario Paul Ahues Blanchait

THESE

présentée à

l' Université Scientifique et Médicale de Grenoble
et à
l' Institut National Polytechnique de Grenoble

POUR OBTENIR LE DIPLOME DE
DOCTEUR - INGENIEUR
en Mathématiques Appliquées

par

Mario Paul AHUES BLANCHAIT

RAFFINEMENT D'ELEMENTS PROPRES APPROCHES D'UN OPERATEUR COMPACT

Thèse soutenue le 6 juin 1983 devant la Commission d'Examen :

Monsieur	François ROBERT	:	Président
Madame	Françoise CHATELIN		Examinateurs
Messieurs	Pierrre Jean LAURENT		
	Fulbert MIGNOT		
	Jacques RAPPAZ		

à mes parents

à ma femme

à mes filles, un jour vous comprendrez
le sens profond de cet
effort : celui de retrouver
tant de valeurs perdues et
tant de choses promises
pour un lendemain
toujours lointain

et l'on verra flamber à nouveau
et très haut
mon coeur brûlant et étoilé.

PABLO NERUDA.

Je tiens à exprimer ma profonde gratitude à Madame Françoise CHATELIN qui a bien voulu m'accueillir dans l'Equipe d'Analyse Numérique et diriger ce travail.

Que Monsieur François ROBERT reçoive ma plus vive reconnaissance pour avoir accepté la présidence du Jury.

Je suis très sensible à l'honneur que me fait Monsieur Jacques RAPPAZ en acceptant de siéger à ce Jury.

A Monsieur Fulbert MIGNOT j'adresse mes remerciements les plus sincères pour sa présence parmi ce Jury.

Que Monsieur Pierre Jean LAURENT trouve ici l'expression de ma reconnaissance pour avoir accepté de juger mon travail.

Je remercie amicalement Mauricio TELIAS pour son inestimable collaboration dans ces recherches.

Je remercie Mory pour l'aide qu'elle m'a donnée lors de la correction des manuscrits.

Mes remerciements vont aussi à Madame Geneviève BICAÏS qui a fourni une nouvelle preuve de sa grande compétence dans la dactylographie des textes mathématiques et à Messieurs Daniel IGLESIAS et Claude ANGUILLE, du Service de Reprographie, pour l'excellente qualité de leur travail ainsi que pour la sympathie qu'ils m'ont témoignée.

L'Equipe d'Analyse Numérique m'a permis de vivre une expérience très riche. Pendant mon séjour auprès d'elle j'ai aussi pu me faire de précieuses amitiés dont je garderai un très beau souvenir : Filomena d'ALMEIDA, du Portugal et Rekha KULKARNI de l'Inde.

AVANT - PROPOS

Cette thèse a été réalisée au sein de l'Equipe d'Analyse Numérique du Laboratoire IMAG. Elle s'inscrit dans le cadre des activités poursuivies durant 1981 par ce qui se voulut un groupe de recherche axé sur les méthodes de raffinement itératif, groupe constitué de Filomena d'Almeida, Mauricio Telias et moi-même. C'est, peut-être, de ce fait que j'ai choisi tout naturellement la première personne du pluriel lors de la rédaction de ce document. Je dois, pourtant, signaler que je suis le seul responsable d'éventuelles erreurs commises.

Parmi les divers résultats présentés dans cette thèse, une partie de la section 2.3 et la plupart des expériences numériques du chapitre 5 sont le fruit de cette collaboration.

Je ne peux pas m'empêcher de dire que je considère cet esprit de travail en communauté comme indispensable dans le domaine scientifique, excluant par là-même tout esprit de concurrence malsaine. Le poids d'un travail scientifique n'a pas à être divisé par le nombre de ses auteurs pour en déduire le salaire de chacun d'eux... Par contre, chaque chercheur doit participer avec sérieux et responsabilité au travail du groupe, aussi bien dans une activité de recherche que dans celle qui lui est impérativement associée et qui s'appelle l'enseignement. Car si ces deux activités connaissent des moments de création et de réflexion, c'est réellement lors de l'enseignement qu'elles sont vraiment centrées sur l'homme et sur la société.

Cela veut dire que le travail scientifique ne devrait se concevoir que dans un contexte profondément humain. Et je tiens à remarquer que, dans ce cadre et pour que ce travail ait un sens créateur dont pourra profiter la société, il faut accepter que ce soit une autre logique -différente de celle des mathématiques- qui rende compte des phénomènes humains, infiniment plus riches et par tant plus complexes que ceux de notre domaine mathématique restreint.

Nul ne peut s'estimer exclu de cette responsabilité qui est à la base de tout espoir...

Ce travail a été fait dans le cadre de l'accord de Coopération Scientifique et Technique entre le Gouvernement Français et le Laboratoire Departamento de Matemáticas y Ciencias de la Computación de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, à Santiago du Chili, signé au début des années 1970.

NOTATIONS

On utilise les notations suivantes, la plupart étant usuelles.

$(A)_{ij}$	Coefficient de la matrice A en ligne i, colonne j
(a_{ij})	Matrice dont le coefficient en ligne i, colonne j est a_{ij}
A^H	Matrice transposée-conjuguée de A. Si $A = (a_{ij})$ alors $A^H = (\overline{a_{ji}})$
A^{-1}	Opérateur (ou matrice) inverse de A ou preimage par A.
$\bar{\alpha}$	Nombre complexe conjugué de α.
B(x,r)	Boule ouverte de centre x et de rayon $r > 0$. C'est le sous-ensemble des y qui vérifient $\|x-y\| < r$.
C	L'ensemble des nombres complexes.
:=	Symbole qui sert à définir ce que l'on place à sa gauche par ce qui est écrit à sa droite.
dim X	Dimension, au sens algébrique, de l'espace vectoriel X.
GF	Composition d'opérateurs. $(GF)(x) := G(F(x))$.
i	Nombre complexe dont la partie réelle est 0 et la partie imaginaire est 1.
I_N	Matrice unité (ou identité) de taille N.
KerT	Noyau (ou kernel) de l'opérateur linéaire T. C'est le sous-ensemble preimage de 0 : $T^{-1}\{0\}$.
$\mathbb{N}$	L'ensemble des entiers naturels ou non négatifs.
$\mathbb{R}$	L'ensemble des nombres réels.
[T]	Matrice qui sert à représenter (une partie de) l'opérateur linéaire T.
$V\|_B$	Restriction de l'opérateur V au sous-ensemble B de son domaine de définition.

TABLE DE MATIERES

INTRODUCTION GENERALE

I.1 - QU'EST CE QU'UNE METHODE DE RAFFINEMENT ITERATIF ?

Supposons que l'on ait à résoudre l'équation $F(x) = 0$ pour laquelle existe une solution ξ que l'on ne sait pas calculer de façon exacte, ou -comme on dit très souvent- par une méthode analytique.

Une méthode numérique itérative consiste à définir une suite $\{\xi^k\}$ -où, en général, ξ^0 est un élément arbitraire- qui converge (en un sens à préciser) vers ξ.

Or, *une méthode de raffinement itératif* a quelque chose de plus : le caractère arbitraire de ξ^0 peut être notablement diminué de la manière suivante. A l'opérateur F on substitue ce que l'on appelle couramment une approximation notée F_0. On suppose que l'on sait résoudre l'équation approchée $F_0(x) = 0$ avec une précision aussi grande que l'on veut. Soit x_0 la solution de l'équation approchée. En général, F_0 est un opérateur tel que x_0 est proche de ξ. On dit que x_0 est une solution approchée de ξ, expression qui est aussi fréquente que contradictoire : ce que l'on veut dire réellement c'est que x_0 est une approximation à la solution ξ. Si cette approximation n'est pas satisfaisante alors on peut soit améliorer l'approximation F_0 et résoudre de nouveau l'équation approchée correspondante, soit appliquer à x_0 une méthode de raffinement itératif. On dit que l'on *raffine itérativement* l'approximation x_0 lorsque l'on calcule une suite $\{\xi^k\}$, qui converge vers ξ , telle que $\xi^0 = x_0$. Très souvent, en pratique, pour calculer la suite $\{\xi^k\}$ on doit :

i) Résoudre des équations du type $F_0(x) = d$ (où F_0 est, bien entendu, l'approximation utilisée pour calculer x_0) et

ii) Calculer la valeur de F en des points donnés.

On préfère l'option de raffiner x_0 lorsque l'option d'améliorer x_0 via une amélioration de F_0 est numériquement déconseillée du fait que l'équation approchée associée devient de plus en plus difficile à résoudre lorsque F_0 se rapproche de F. Il est important de remarquer que les méthodes de raffinement nécessitent seulement la résolution d'équations concernant l'approximation F_0.

En général, ξ^k est la valeur d'un certain opérateur V_k aux points $(\xi^0,\xi^1,\ldots,\xi^{k-1})$:

$$\xi^k = V_k(\xi^0,\ldots,\xi^{k-1}), \qquad k \geq 1.$$

Très souvent, on a simplement :

$$\xi^k = V_k(\xi^{k-1}), \qquad k \geq 1,$$

ou encore :

$$\xi^k = V(\xi^{k-1}), \qquad k \geq 1,$$

où V est un opérateur indépendant de k. Dans ce dernier cas on reconnait une itération de point fixe sur V.

I.2-L'INTERET DU RAFFINEMENT ITERATIF POUR DES EQUATIONS INTEGRALES

Les méthodes que l'on utilise normalement pour discrétiser un opérateur intégral T de type Fredholm :

$$(Tx)(t) = \int_0^1 \kappa(t,s)x(s)ds$$

fournissent une représentation matricielle A_N, de taille N, qui est pleine. N représente la qualité de la discrétisation (par exemple, le nombre de points de la formule de quadrature ou la dimension de l'espace sur lequel on fait une projection d'interpolation ou une projection orthogonale).

L'équation de Fredholm de deuxième espèce

$$(T-z)x = f, \qquad (T-z \text{ bijectif})$$

et le problème de valeurs propres associé

$$T\phi = \lambda\phi , \qquad \phi \neq 0$$

nous amènent à considérer la résolution numérique des problèmes matriciels suivants :

$$(A_N - zI_N)x_N = f_N , \qquad (A_N - zI_N \text{ régulière})$$

$$A_N\phi_N = \lambda_N\phi_N, \qquad \phi_N \neq 0$$

où I_N est la matrice identité de taille N.

Si N est grand (disons N > 100) la résolution numérique de ces équations matricielles constitue un problème coûteux, voire impossible, lorsque nous ne disposons que des méthodes usuelles de l'algèbre linéaire numérique. Or, en pratique, il est parfois nécessaire que A_N soit une matrice de grande taille pour que les solutions x_N et ϕ_N approchent x et ϕ, respectivement, de façon satisfaisante. L'intérêt d'une méthode de raffinement est donc de ne nécessiter que la résolution de problèmes du type

$$(A_n - z\, I_n)x_n = f_n , \qquad (A_n - zI_n \text{ régulière})$$

$$A_n\phi_n = \lambda_n\phi_n , \qquad \phi_n \neq 0$$

où n est beaucoup plus petit que N , pour obtenir les points de départ des méthodes itératives qui convergent vers x_N et ϕ_N respectivement. Nous présentons dans cette thèse diverses méthodes itératives pour le raffinement d'un vecteur propre approché de l'opérateur T. Toutes ces méthodes ont en commun le fait qu'elles n'utilisent la discrétisation de grande taille A_N que pour le calcul de produits $A_N y_N$ où y_N est un vecteur donné.

Il faut bien remarquer que ces méthodes exploitent le fait que A_n et A_N sont deux discrétisations d'un même opérateur intégral T. Ceci les rend plus avantageuses que les méthodes générales pour le problème de valeurs propres des grandes matrices, telles que, par exemple, les méthodes de Lanczos, Lanczos par blocks, Arnoldi ou des itérations simultanées. On reviendra sur ce point dans les commentaires à la fin du chapitre 4.

I.3 - COMMENT GARANTIR LA CONVERGENCE DU RAFFINEMENT LORS DES EQUATIONS INTEGRALES ?

Signalons tout d'abord que, d'une manière générale, la convergence sera assurée si le point de départ des itérations est suffisamment proche de la solution.

Prenons premièrement le cas de l'équation intégrale dite de Fredholm de deuxième espèce :

$$(T-z)x = f$$

où T est l'opérateur intégral du paragraphe précédent et $z \neq 0$ est un nombre (en général complexe) tel que T-z est bijectif. Si l'on substitue à T une approximation T_n on aura à résoudre l'équation approchée

$$(T_n-z)x = f$$

Ceci est possible lorsque T_n-z est bijectif. La solution et son approximation sont alors données respectivement par

$$u = (T-z)^{-1}f \quad \text{et} \quad u_n = (T_n-z)^{-1}f$$

On conclut ainsi que u_n sera proche de u dans la mesure où pour n assez grand T_n satisfait aux conditions suivantes

. T_n- z est bijectif

. $(T_n-z)^{-1}f$ est proche de $(T-z)^{-1}f$.

Ces exigences nous conduiront à la notion de *convergence stable* de T_n-z vers T-z.

Considérons deuxièmement le problème de valeurs propres. Dans ce cas on aura à résoudre le problème approché

$$T_n\phi_n = \lambda_n\phi_n \qquad \phi_n \neq 0$$

Supposons que λ soit une valeur propre simple et non nulle de T. Alors, étant donné λ, la solution de $T\phi = \lambda\phi$ est un sous-espace uni-dimensionnel que l'on démontre être égal à l'espace image de l'opérateur

$$P = -\frac{1}{2\pi i}\int_{\Gamma}(T-z)^{-1}dz$$

où Γ est une courbe fermée, assez régulière, tracée dans l'ensemble des z tels que T-z est bijectif, entourant λ et aucune autre singularité de $(T-z)^{-1}$.

Or, étant donné λ_n, valeur propre de T_n on est amené à considérer l'opérateur

$$P_n = -\frac{1}{2\pi i}\int_{\Gamma_n}(T_n-z)^{-1}dz$$

où Γ_n est une courbe isolant λ_n. Donc, si l'on veut que la solution du problème approché soit proche de la solution du problème en T on doit au moins imposer les conditions suivantes :

. Que Γ_n puisse être prise égale à Γ et celle-ci arbitrairement réduite : ainsi λ_n sera proche de λ.

. Que T_n-z soit bijectif pour tout z sur Γ.

. Que l'espace image de P_n soit proche -en un sens aussi géométrique que possible!- de celui de P.

. Que la dite notion de proximité de deux sous-espaces implique, au moins, l'égalité des dimensions : $\dim P_n X = \dim PX$, pour n assez grand : ainsi λ_n sera, elle aussi, une valeur propre simple de T_n.

On verra que tout ceci nous conduit à la notion de *convergence fortement stable* de T_n-z vers T-z autour de λ.

I.4 - DESCRIPTION SOMMAIRE DES CHAPITRES DE CETTE THESE

Le premier chapitre contient les éléments de base de l'analyse fonctionnelle et de l'approximation spectrale d'opérateurs linéaires continus dans un espace de Banach complexe qui s'avèrent nécessaires pour la compréhension des contenus des chapitres suivants. La plupart des démonstrations sont omises mais on donne des références.

Dans le deuxième chapitre on présente la méthode de Newton, le principe de correction du résidu et l'on propose une méthode de raffinement itératif dite mixte, tout en donnant des exemples qui seront repris au chapitre suivant dans un contexte et dans un but plus précis.

Le troisième chapitre constitue le noyau de ce travail. Il contient les apports théoriques originaux de cette thèse. Nous y présentons quatre familles de méthodes itératives pour le raffinement des éléments propres simples d'un opérateur compact dans un espace de Banach complexe. Selon le choix d'un paramètre qui caractérise ces familles on peut atteindre une convergence quadratique, dans le pire des cas la convergence étant linéaire. Certains choix du paramètre nous ramènent à des méthodes qui ont été déjà proposées dans la littérature sur ce sujet, bien que parfois dans un contexte plus restreint.

Le quatrième chapitre est destiné à l'étude de la représentation matricielle des différentes approximations d'un opérateur intégral de type Fredholm à noyau continu défini sur l'espace de Banach des fonctions numériques complexes définies et continues sur un intervalle de longueur finie de la droite réelle. On en déduit le coût des itérations proposées au chapitre précédent.

Finalement, le cinquième chapitre contient les résultats des expériences numériques dûment analysées et commentées.

PREMIER CHAPITRE

SUR L'ANALYSE FONCTIONNELLE ET L'APPROXIMATION SPECTRALE

Notions fondamentales et propriétés

INTRODUCTION

Afin que cette thèse puisse être utilisée pour la formation mathématique des étudiants qui s'intéressent à l'approximation numérique des problèmes spectraux (ou pour lesquels on voudrait qu'il en soit ainsi !) nous avons conçu ce premier chapitre comme étant un recueil des définitions et des théorèmes principaux dans les domaines de l'analyse fonctionnelle, de la théorie spectrale d'opérateurs linéaires continus et de l'approximation de leurs éléments propres.

On ne donne pas, pourtant, la démonstration de chacun de ces résultats mais on indique de manière précise les références appropriées.

Font exception à cette règle les propriétés relatives aux bornes d'erreur des éléments propres approchés dont on fournit de façon détaillée les démonstrations correspondantes en essayant ainsi de familiariser le lecteur avec les techniques que l'on emploie couramment dans ce domaine.

Remarquons finalement que nous avons restreint ce recueil mathématique au minimum nécessaire, ce qui veut dire que l'on ne donne ici que les définitions et les résultats dont on a besoin au cours des chapitres suivants.

1.1 NOTIONS FONDAMENTALES

Soit X un espace de Banach sur le champ C des nombres complexes. Notons $\|\ \|$ la norme de X. On définit *l'espace adjoint* de X comme l'ensemble X^* des fonctionnelles $\psi : X \to C$ qui satisfont aux conditions suivantes :

(i) $\quad \forall x,y \in X \quad \forall \alpha \in C \quad \psi(\alpha x+y) = \bar{\alpha}\, \psi(x)+\psi(y)$

(ii) $\quad |\psi| := \text{Sup}\ \{|\psi(x)| : x \in X,\ \|x\| = 1\} < +\infty.$

L'ensemble X^* est un espace vectoriel sur C et l'application $\psi \in X \to |\psi| \in \mathbb{R}$ est une norme sur X^*. On vérifie aisément que X^* muni de cette norme est un espace de Banach complexe. Il est utile d'introduire la notation

$$\langle \psi,x\rangle := \psi(x) \qquad \langle x,\psi\rangle := \overline{\psi(x)} \qquad \forall x \in X \quad \forall \psi \in X^*.$$

En fait, par l'intermédiaire d'un théorème d'analyse très important, on peut identifier X à un sous-espace de X^{**}, donc il n'y a pas d'ambiguïté dans ces notations.

Exemple 1.1.1 : Dans les applications numériques présentées dans cette thèse l'espace C[0,1] des fonctions $x : [0,1] \to C$ continues sur [0,1] jouera un rôle important. Il s'agit d'un espace de Banach lorsque l'on définit comme norme $\|x\| := \text{Max}\ \{|x(t)| : t \in [0,1]\}$. L'espace adjoint est constitué des fonctions à variation bornée (dûment normalisées) et $\langle \psi,x\rangle := \int_0^1 \overline{x(t)}\, d\psi(t)$ (cf. Yosida 1971, p. 119). □

Exemple 1.1.2 : Soit N un entier ≥ 2. On définit X_N comme le sous-espace de C[0,1] des fonctions x telles que $x\big|_{\Delta_i^{(N)}}$ est un polynôme de degré ≤ 1 où $\Delta_i^{(N)} := [t_i^{(N)}, t_{i+1}^{(N)}]$ $i = 1,\ldots,N-1$ est la partition de [0,1] induite par les points $t_i^{(N)}$ qui vérifient $0 = t_1^{(N)} < t_2^{(N)} < \ldots < t_{N-1}^{(N)} < t_N^{(N)} = 1$. On dit que X_N est l'espace des *fonctions continues linéaires par morceaux* associé à la maille

$\{t_1^{(N)},\ldots,t_N^{(N)}\}$. X_N est de dimension N.

L'espace adjoint X_N^* est engendré par exemple, par la base $\{e_i^{(N)*}: i = 1,\ldots,N\}$ où $\forall x \in X_N \quad \langle e_i^{(N)*},x\rangle := x(t_i^{(N)}) \quad i = 1,\ldots,N.$

□

Etant donnés deux espaces de Banach X,Y on note $L(X,Y)$ l'ensemble des *applications linéaires continues* de X dans Y. $L(X,Y)$ est un espace de Banach si l'on prend comme norme l'application

$$T \in L(X,Y) \to \text{Sup}\{\|Tx\| : x \in X, \|x\| = 1\} \in \mathbb{R}\ .$$

Dans un souci d'alléger l'écriture on va noter $\|\ \|$ la norme de X, de Y et celle de $L(X,Y)$. D'autre part on écrit $L(X)$ pour $L(X,X)$, 1 pour l'opérateur *identité* sur X et z (avec $z \in C$) pour l'opérateur $z1 : x \in X \to zx \in X$.
Il est évident que $z \in L(X)$ $\forall z \in C$.

Etant donné $T \in L(X,Y)$ et $\psi \in Y^*$ il existe un élément unique de X^*, noté $T^*\psi$ tel que $(T^*\psi)x = \psi Tx$ $\forall x \in X$, ou, en utilisant la notation introduite précédemment :

$$\langle T^*\psi,x\rangle = \langle\psi,Tx\rangle \qquad \forall x \in X\ .$$

Cela définit un opérateur $T^* : Y^* \to X^*$ dit *adjoint* de T.

Propriété 1.1.1

Si $T \in L(X,Y)$ *alors* $T^* \in L(Y^*,X^*)$ *et* $\|T^*\| = \|T\|$.

Preuve : Cf. Kato 1966, p. 154. □

Un opérateur $T \in L(X,Y)$ est dit *compact* lorsque T applique les sous-ensembles bornés de X sur des sous-ensembles relativement compacts de Y.

Exemple 1.1.3 : Le théorème d'Arzela-Ascoli caractérise les sous-ensembles relativement compacts de l'espace de Banach $X = C[0,1]$ de la façon suivante :

Pour que $W \subseteq X$ soit relativement compact il faut et il suffit que les deux conditions qui suivent soient satisfaites :

i) W est borné, c'est-à-dire $\text{Sup}\{\|x\| : x \in W\}$ est fini.

ii) W est équicontinu, c'est-à-dire : $\forall \varepsilon > 0\ \exists \delta > 0$ tel que $\forall t_1, t_2 \in [0,1]$, la condition $|t_1 - t_2| < \delta$ implique $\text{Sup}\{|x(t_1) - x(t_2)| : x \in W\} < \varepsilon$.

(cf. Lang 1977, p. 220).

Ce théorème permet de démontrer la compacité de certains opérateurs sur X, dits *opérateurs intégraux :*
Si $\kappa : [0,1] \times [0,1] \to C$ est une fonction continue alors, étant donné $x \in X$, la fonction $Tx : [0,1] \to C$ définie par

$$(Tx)(t) = \int_0^1 \kappa(t,s)x(s)ds \quad \forall t \in [0,1]$$

appartient elle aussi à l'espace X. L'opérateur $T : X \to X$ ainsi construit est linéaire et continu :

$$\|T\| \leq \text{Max}\left\{ \int_0^1 |\kappa(t,s)|ds : t \in [0,1]\right\}.$$

T est un *opérateur intégral de type Fredholm* à noyau κ continu. D'après le théorème d'Arzela-Ascoli on montre sans effort que T est un opérateur compact.
(cf. Chatelin 1983, p. 83). □

Propriété 1.1.2

Si $A \in L(X,Y)$ *et si* $B \in L(Y,X)$ *est compact alors* $AB \in L(Y)$ *et* $BA \in L(X)$ *sont compacts.*

Preuve : Cf. Kato 1966, p. 158 . □

Propriété 1.1.3

Si $T \in L(X,Y)$ *est compact alors* $T^* \in L(Y^*,X^*)$ *est compact.*

Preuve : Cf. Kato 1966, p. 159. □

Propriété 1.1.4

$1 \in L(X)$ *est compact si et seulement si* X *est un espace de dimension finie.*

Preuve : Cf. Kato 1966, pp. 7, 131 et 157. □

Propriété 1.1.5

Si $T \in L(X,Y)$ *est tel que* $\dim TX < +\infty$ *alors* T *est compact.*

Preuve : Cette propriété est une conséquence des propriétés 1.1.2 et 1.1.4. □

Dans la suite nous considérons des fonctions définies dans un ouvert Ω du plan complexe et à valeurs dans $L(X)$.

D'une manière générale, les définitions et les propriétés de la théorie des fonctions de C dans C peuvent être généralisées au cas des fonctions de C dans un espace de Banach complexe quelconque.

Soit X un espace de Banach sur C et soit Ω un ouvert dans C. Une fonction $f : \Omega \to X$ est dite *analytique* sur Ω si $\frac{df}{dz}$ existe en tout point de Ω.

La généralisation de la plupart des théorèmes classiques se fait à l'aide du résultat suivant :

Si $f : \Omega \to \mathbf{X}$ est analytique sur Ω alors pour tout $\psi \in \mathbf{X}^*$, $\langle \psi, f \rangle$ est analytique sur Ω.

En utilisant cette propriété on démontre par exemple, le Théorème de Liouville pour le cas général :

Si $f : C \to \mathbf{X}$ *est analytique et bornée sur* C *alors* f *est une fonction constante.*

Parmi les ouvrages qui traitent de ce sujet on cite Dunford et Schwartz 1958 (pp. 224-232).

Rappelons maintenant quelques notions de la théorie spectrale.

Soit $T \in L(X)$. On définit

L'ensemble résolvant de T : $\rho(T) := \{z \in C : (T-z)^{-1} \in L(X)\}$

Le spectre de T : $\sigma(T) := C \setminus \rho(T)$

Le rayon spectral de T : $r_\sigma(T) := \mathrm{Inf}\{\|T^n\|^{1/n} : n \in \mathbb{N}, n > 0\}$.

Propriété 1.1.6

Si $A \in L(X,Y)$ *et* $B \in L(Y,X)$ *alors* $r_\sigma(AB) = r_\sigma(BA)$.

Preuve : $\|(AB)^n\|^{1/n} \leq (\|A\|.\|B\|)^{1/n} \|(BA)^n\|^{1/n}$ montre l'inégalité $r_\sigma(AB) \leq r_\sigma(BA)$ ce qui suffit. □

Propriété 1.1.7

Si $T \in L(X)$ *vérifie* $r_\sigma(T) < 1$ *alors*

i) $1 \in \rho(T)$

ii) $\|(1-T)^{-1} - \sum_{i=0}^{n} T^i\| \to 0$ *si* $n \to \infty$.

Preuve : Cf. Chatelin 1983 p. 101. □

Propriété 1.1.8

Si $T \in L(X)$ *alors*

i) $\sigma(T)$ *est non vide et compact dans* $\mathbb{C}$

ii) $r_\sigma(T) = \text{Max}\{|\lambda| : \lambda \in \sigma(T)\}$

Preuve : Cf. Chatelin 1983 p. 101.
Remarquons que d'après cette propriété $r_\sigma(T)$ est invariant sous un changement de norme équivalente dans $L(X)$. □

Propriété 1.1.9

Etant donnés $T \in L(X)$ *et un réel* $\varepsilon > 0$ *il existe une norme équivalente* $\|\ \|_*$ *dans* $L(X)$ *qui vérifie* $r_\sigma(T) \leq \|T\|_* \leq r_\sigma(T)+\varepsilon$.

Preuve : Cf Chatelin 1983 p. 77. □

Propriété 1.1.10

Si $T \in L(X)$ *alors*

$$\lambda \in \sigma(T^*) \iff \bar{\lambda} \in \sigma(T)$$

Preuve : On montre, de façon équivalente, que $z \in \rho(T^*) \iff \bar{z} \in \rho(T)$.
Cf. Kato 1966, p. 184. □

L'ensemble des *isomorphismes* sur X est défini ainsi
$\text{Iso}(X) := \{T \in L(X) : 0 \in \rho(T)\}$

Propriété 1.1.11

$\text{Iso}(X)$ *est un sous-ensemble ouvert de* $L(X)$.
Plus précisément, si $A \in \text{Iso}(X)$ *alors* $\forall B \in L(X) : \|A-B\| < 1/\|A^{-1}\|$ *implique* $B \in \text{Iso}(X)$.

Preuve : $r_\sigma(1-A^{-1}B) \leq \|A^{-1}(A-B)\| \leq \|A-B\|\ \|A^{-1}\|$
et il suffit d'appliquer la propriété 1.1.7. □

Propriété 1.1.12

Si $T \in \mathrm{Iso}(X)$ *alors* $T^{-1} \in \mathrm{Iso}(X)$ *et l'inversion* $T \to T^{-1}$ *est une application continue de* $\mathrm{Iso}(X)$ *sur lui-même.*

Preuve : Cf. Cartan 1972 p. 23. □

Propriété 1.1.13

Si $T \in \mathrm{Iso}(X)$ *alors* $T^* \in \mathrm{Iso}(X^*)$ *et* $(T^*)^{-1} = (T^{-1})^*$.

Preuve : Cf. Kato 1966, p. 169. □

Soit $T \in L(X)$. La condition $z \in \rho(T)$ entraîne que la solution $x(f)$ de l'équation $(T-z)x = f$ est continue comme fonction de f. En effet $x(f) = (T-z)^{-1}f$ et $(T-z)^{-1} \in L(X)$ lorsque $z \in \rho(T)$. Cela motive la définition qui suit :

Etant donnés $T \in L(X)$ et $z \in \rho(T)$ *l'opérateur résolvant* (ou simplement *la résolvante*) de T au point z est

$$R(T,z) := (T-z)^{-1}.$$

Propriété 1.1.14

Si $T \in L(X)$, $z \in \rho(T)$ *et* $z \neq 0$ *alors* $R(T,z) = \frac{1}{z}(R(T,z)T-1)$.

Preuve : Triviale. □

Propriété 1.1.15

Si $A,B \in L(X)$ *et* $z \in \rho(A) \cap \rho(B)$ *alors*
$R(A,z) - R(B,z) = R(A,z)\,(B-A)\,R(B,z)$.

Preuve : Il suffit d'écrire $B-A = B-z - (A-z)$. □

Propriété 1.1.16

Si $T \in L(X)$ *et* $z \in \rho(T)$ *alors* $R(T,z)^* = R(T^*, \bar{z})$.

Preuve : Cf. Kato 1966, p. 184. □

Propriété 1.1.17

Si $T \in L(X)$ *alors l'application* $z \in \rho(T) \to R(T,z) \in L(X)$
est analytique sur $\rho(T)$ *et l'application*
$z \in \rho(T) \to r_\sigma(R(T,z)) \in \mathbb{R}$
est semi-continue supérieurement.

Preuve : Cf. Chatelin 1983, p. 96. □

On dit que $\lambda \in C$ est une *valeur propre* de $T \in L(X)$ si et seulement si $T-\lambda$ n'est pas injectif. Dans ce cas-là le sous-espace $\mathrm{Ker}(T-\lambda)$ ne se réduit pas à $\{0\}$ et tout vecteur $\phi \neq 0$ de cet espace est appelé *vecteur propre* de T associé à λ. Une valeur propre λ de T est dite *isolée* lorsqu'il existe une courbe de Jordan fermée Γ_λ renfermant une partie compacte Δ_λ de C telle que $\sigma(T) \cap \Delta_\lambda = \{\lambda\}$. Pour un tel λ on définit :

$$P(T,\lambda) := -\frac{1}{2\pi i}\int_{\Gamma_\lambda} R(T,z)dz$$

$$M(T,\lambda) := P(T,\lambda)X$$

$$m(T,\lambda) := \dim M(T,\lambda)$$

$$S(T,z) := \begin{cases} R(T,z)(1-P(T,\lambda)) & \text{si } z \neq \lambda \\ \frac{1}{2\pi i}\int_{\Gamma_\lambda} R(T,w)\frac{dw}{w-\lambda} & \text{si } z = \lambda \end{cases}$$

Si aucune confusion n'est possible on utilisera la notation simplifiée :

$$P := P(T,\lambda)$$

$$M := M(T,\lambda)$$

$$m := m(T,\lambda)$$

$S(z) := S(T,z)$

$S := S(T,\lambda)$

Propriété 1.1.18

Si $T \in L(X)$ *et* λ *est une valeur propre isolée de* T *alors*

i) P *est une projection de* X *sur* M *:*

$P \in L(X)$

$P^2 = P$

ii) $\mathrm{Ker}(T-\lambda) \subseteq M$

iii) M *est invariant par* T *:*

$TM \subseteq M$

iv) $\forall z \in \rho(T) \cup \{\lambda\}$, $S(z) \in L(X)$.

v) $\rho(T) \cup \{\lambda\}$ *est un ouvert dans* C *et l'application*

$z \in \rho(T) \cup \{\lambda\} \to S(z) \in L(X)$

est analytique sur $\rho(T) \cup \{\lambda\}$

vi) L'application

$z \in \rho(T) \cup \{\lambda\} \to r_\sigma(S(z))$

est semi-continue supérieurement sur $\rho(T) \cup \{\lambda\}$

vii) $S = [(T-\lambda)_{|(1-P)X}]^{-1}(1-P)$

Preuve : Cf. Chatelin 1983, pp. 104-107, **Vesentini** 1968. □

$P(T,\lambda)$ est la *projection spectrale* ; $M(T,\lambda)$ le *sous-espace invariant* ; $m(T,\lambda)$ la *multiplicité algébrique* ; $S(T,\lambda)$, la *résolvante réduite* associés tous à λ. λ est dit *simple* si $m(T,\lambda) = 1$. On note $Q\sigma(T)$ l'ensemble de valeurs propres isolées dont $m(T,\lambda)$ est fini. Il est clair que $Q\sigma(T) \subseteq \sigma(T)$.

Propriété 1.1.19

Si $T \in L(X)$ *alors*

i) $\lambda \in Q\sigma(T) \iff \bar{\lambda} \in Q\sigma(T^*)$

ii) $P(T,\lambda)^* = P(T^*,\bar{\lambda})$ $\quad \forall \lambda \in Q\sigma(T)$

iii) $S(T,\lambda)^* = S(T^*,\bar{\lambda})$ $\quad \forall \lambda \in Q\sigma(T)$

iv) $m(T,\lambda) = m(T^*,\bar{\lambda})$ $\quad \forall \lambda \in Q\sigma(T)$

Preuve : Cf. Chatelin 1983. p. 113. □

Propriété 1.1.20

Si $T \in L(X)$, $z \in \rho(T)$ *et* $\lambda \in Q\sigma(T)$ *alors*

i) $T, R(T,z), P(T,\lambda)$ *et* $S(T,\lambda)$ *commutent*

ii) *si* $\lambda \neq 0$,

$$S(T,\lambda) = \frac{1}{\lambda}\,(S(T,\lambda)T + P(T,\lambda) - 1)$$

Preuve : Triviale. □

Propriété 1.1.21

Etant donnés $T \in L(X)$, $\lambda \in Q\sigma(T)$,

$\phi \in M(T,\lambda)$, $\phi^* \in M(T^*,\bar{\lambda})$ *tels que*
$m(T,\lambda) = 1$ *et* $\langle \phi^*, \phi \rangle = 1$

alors $P(T,\lambda)x = \langle x, \phi^* \rangle \phi$ $\quad \forall x \in X$.

Preuve : Cf. Chatelin 1983, p. 113. □

Propriété 1.1.22

Si $T \in L(X)$, $\lambda \in Q\sigma(T)$ *et* $m(T,\lambda) = 1$
alors $(T-\lambda)x \in M(T,\lambda)$ *implique* $x \in M(T,\lambda)$.

Preuve : $(T-\lambda)x = \alpha\phi$ ($\alpha \in C$, $\phi \in M(T,\lambda)-\{0\}$) implique
$0 = P(T,\lambda)(T-\lambda)x = \alpha\phi$, donc $\alpha = 0$. □

En ce qui concerne les opérateurs compacts on a le résultat suivant tout à fait remarquable :

Propriété 1.1.23

Si $T \in L(X)$ *est compact alors*
$\sigma(T) \subseteq Q\sigma(T) \cup \{0\}$
l'égalité n'ayant lieu que dans chacun des cas suivants :

i) X *est de dimension infinie .*

ii) X *est de dimension finie et* T *n'est pas bijectif.*

Preuve : Ce théorème est démontré dans Kato 1966, p. 185, Yosida 1971, p. 283. Une preuve très simple qui utilise la notion algébrique *d'indice* d'un opérateur : indT := dim KerT-dim(X/TX), se trouve dans Lang 1977 p. 214. Une conséquence immédiate de cette propriété est la suivante : si $z \neq 0$ et T-z est injectif alors $z \in \rho(T)$ et donc $T-z \in Iso(X)$. □

1.2 CONVERGENCE D'OPERATEURS

On considère toujours un espace de Banach X. Soit
$\bar{B} := \{x \in X : \|x\| \leq 1\}$, la *boule unité fermée* de X.
Etant donnés une suite $\{T_n\}$ dans $\mathcal{L}(X)$ et un opérateur $T \in L(X)$ nous serons concernés avec certaines notions de *convergence* $T_n \to T$ dont nous aurons besoin pour assurer la convergence des procédés numériques que l'on va proposer.

Du point de vue topologique la notion la plus naturelle de convergence de la suite $\{T_n\}$ vers T et celle qui s'appuie sur la structure normée de $L(X)$: la *convergence en norme* (appelée aussi *convergence uniforme* ou tout simplement convergence dans $L(X)$). On la définit et la note ainsi :

$$T_n \xrightarrow{\|\ \|} T \quad \text{si et seulement si } \|T_n - T\| \to 0 \text{ lorsque } n \to \infty.$$

Pourtant, cette notion ignore le fait que T_n et T sont des opérateurs; elle les considère tout simplement comme éléments d'un espace vectoriel normé : $L(X)$. Par contre la *convergence ponctuelle* tient compte essentiellement de la nature des objets T_n et T et la topologie de $L(X)$ n'est pas concernée :

$$T_n \xrightarrow{p} T \text{ si et seulement si } \forall x \in X \quad \|T_n x - Tx\| \to 0 \quad \text{si } n \to \infty.$$

Propriété 1.2.1

Si $T_n \xrightarrow{\|\ \|} T$ *alors* $T_n \xrightarrow{p} T$.

Preuve : Triviale. La réciproque est évidemment fausse. □

Propriété 1.2.2

Si $T_n \xrightarrow{p} T$ *alors* $\text{Sup}\{\|T_n\| : n \in \mathbb{N}\}$ *est fini.*

Preuve : Il s'agit d'un corollaire du théorème de Banach-Steinhaus dont on peut trouver une démonstration dans Lang 1977 p. 199.

□

Si $\{T_n - T\}$ est une suite d'opérateurs compacts et si $\{T_n^*\}$ ne converge pas ponctuellement vers T^* alors la notion de convergence que l'on va présenter maintenant est, pour ainsi dire, intermédiaire entre la convergence ponctuelle et la convergence en norme : la *convergence collectivement compacte*. On la définit et la note comme suit :

$T_n \xrightarrow{cc} T$ si et seulement si $T_n \xrightarrow{p} T$ et $\bigcup_{n\in\mathbb{N}} (T_n-T)\bar{B}$ est relativement compact dans X.

Propriété 1.2.3

Si T_n-T *est compact* $\forall n \in \mathbb{N}$ *alors*
$\| \ \|$
$T_n \longrightarrow T$ *implique* $T_n \xrightarrow{cc} T$.

Preuve : Cf. Chatelin 1983, p. 126. □

Propriété 1.2.4

Si $T_n \xrightarrow{cc} T$ *alors* T_n-T *est compact* $\forall n \in \mathbb{N}$.

Preuve : Triviale. □

Propriété 1.2.5

Si $W \subseteq X$ *est relativement compact et* $T_n \xrightarrow{p} T$ *alors*

$\sup \{\| (T_n-T)x\| : x \in W\} \to 0$ *si* $n \to \infty$.

Preuve : Cf. Chatelin 1983, p. 124. Remarquons que cette propriété montre que lorsque X est de dimension finie alors les trois notions de convergence données jusqu'ici s'identifient. □

Propriété 1.2.6

Si $T_n \xrightarrow{p} T$, $P_n \xrightarrow{cc} P$ *et* P *est compact alors* $\|(T_n-T)P_n\| \to 0$ *si* $n \to \infty$.

Preuve : On déduit ce résultat du précédent.
On en tire aussi que si $T_n \xrightarrow{p} T$ et P est compact alors $\|(T_n-T)P\| \to 0$ lorsque $n \to \infty$. □

Propriété 1.2.7

Si $T_n \xrightarrow{p} T$, $P_n \xrightarrow{cc} P$ *et* P *est compact alors* $T_n P_n \xrightarrow{cc} TP$ *et* $P_n T_n \xrightarrow{cc} PT$.

Preuve : Cf. Chatelin 1983. p. 125. □

Propriété 1.2.8

Si $\{P_n\}$ *est une suite de projections et* P *est une projection de rang fini alors les deux conditions suivantes sont équivalentes :*

i) $P_n \xrightarrow{cc} P$.

ii) $P_n \xrightarrow{p} P$ *et* dim $P_n X$ = dim PX *pour* n *suffisamment grand.*

Preuve : L'implication i) => ii) est une conséquence des propriétés suivantes :

. Si $P_n \xrightarrow{p} P$ et dim PX < ∞ alors pour n assez grand dim $P_n X \geq$ dim PX.

. Si pour toute suite bornée $\{x_n\}$ telle que $x_n \in P_n X$ il existe une sous-suite qui converge vers un élément de PX et si dim PX < ∞ alors dim $P_n X \leq$ dim PX pour n assez grand.

L'implication ii) => i) découle du résultat suivant :

. Si $\{x_i : i = 1,\ldots,m\}$ est une base de PX alors pour n assez grand $\{P_n x_i : i = 1,\ldots,m\}$ est une base de $P_n X$ et sa base adjointe $\{x^*_{in} : i = 1,\ldots,m\}$ vérifie

$$\text{Sup}\{|x^*_{in}| : n \text{ assez grand}\} < \infty \text{ et}$$

$$|x^*_{in} - P^*_n x^*_i| \to 0 \quad \text{si } n \to \infty$$

pour i = 1,...,m (Cf. Chatelin 1983 pp. 127-129).

□

Exemple 1.2.1 : Soit $\{\pi_n\}$ une suite de projections de rang fini telle que $\pi_n \xrightarrow{p} 1$ et soit $T \in L(X)$ un opérateur compact.

L'opérateur $T_n^P := \pi_n T$ sera appelé *approximation de projection* de T. D'après la propriété 1.2.6 on a $T_n^P \overset{\|\|}{\longrightarrow} T$.

L'opérateur $T_n^S := T\pi_n$ s'appelle *approximation de Sloan* de T et, d'après la propriété 1.2.7, $T_n^S \xrightarrow{cc} T$.

L'opérateur $T_n^G := \pi_n T_n^P$ s'appelle *approximation de Galerkin* de T, et, d'après la propriété 1.2.7., $T_n^G \xrightarrow{cc} T$.

Ces trois approximations seront reprises dans le cas des opérateurs intégraux lors des applications numériques. □

Exemple 1.2.2 : Soit $X = C[0,1]$ et X_n le sous-espace de X des fonctions linéaires par morceaux (cf. exemple 1.1.2) associé à la partition uniforme $\Delta_i^{(n)} = [\frac{i-1}{n-1}, \frac{i}{n-1}]$ $i = 1,\ldots,n-1$. On note $h_n = \frac{1}{n-1}$ le pas de la discrétisation. Les fonctions suivantes, appelées *fonctions chapeaux*, constituent une base de X_n : $\forall i = 1,\ldots,n$.

$$e_i^{(n)}(t) = \begin{cases} 1 - \frac{1}{h_n}|t - t_i^{(n)}| & \text{si } t \in [t_{i-1}^{(n)}, t_{i+1}^{(n)}] \\ 0 & \text{si non} \end{cases}$$

où $t_i^{(n)} = \frac{i-1}{n-1}$, $i = 1,\ldots,n$, $t_0^{(n)} = 0$, $t_{n+1}^{(n)} = 1$.

Etant donné un opérateur intégral de Fredholm T à noyau κ continu on définit *l'approximation de Nyström* ainsi

$$(T_n^N x)(t) = \sum_{j=1}^{n} \omega_j^{(n)} \kappa(t, t_j^{(n)}) x(t_j^{(n)}) \qquad \forall x \in X.$$

où $$\omega_j^{(n)} = \begin{cases} h_n/2 & \text{si } j = 1 \text{ ou } j = n \\ h_n & \text{si } 1 < j < n. \end{cases}$$

On montre que $T_n^N \xrightarrow{cc} T$ (cf. Chatelin 1983, p. 196).

Soit π_n *l'opérateur d'interpolation* dans X_n aux points $t_i^{(n)}$:

$$(\pi_n x)(t) = \sum_{i=1}^{n} x(t_i^{(n)}) e_i^{(n)}(t).$$

On vérifie que π_n est une projection et $\pi_n \xrightarrow{p} 1$. (Cf. Ahués et al 1982 p. 53). *L'approximation de Fredholm* est donnée par

$$(T_n^F x)(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_j^{(n)} \kappa(t_i^{(n)}, t_j^{(n)}) x(t_j^{(n)}) e_i^{(n)}(t)$$

donc

$T_n^F = \pi_n T_n^N$ et, d'après la propriété 1.2.7, $T_n^F \xrightarrow{cc} T$. □

La notion de convergence que nous allons introduire maintenant est motivée par des faits qui concernent le calcul et l'analyse numérique .

Soit $T \in L(X)$ et $z \in \rho(T)$. Lorsque l'on veut "résoudre numériquement" l'équation $(T-z)x = f$, dont la solution exacte (parfois dite "analytique") est $x = R(T,z)f$, on substitue à T une "approximation" T_n de T, c'est-à-dire $\{T_n\}$ est une suite d'opérateurs dans $L(X)$ telle que $T_n \xrightarrow{p} T$.

Pourtant, ceci ne garantit pas que l'on pourra résoudre l'équation approchée $(T_n-z)x_n = f$. Il faut que $z \in \rho(T_n)$ et alors la "solution numérique" sera $x_n = R(T_n,z)f$. Mais est-elle proche de la solution exacte ? D'après la propriété 1.1.15

$$x_n - x = (R(T_n,z)-R(T,z))f = R(T_n,z)(T-T_n)R(T,z)f$$

$$= R(T_n,z)(T-T_n)x .$$

Donc, $\|x_n-x\| \leq \|R(T_n,z)\| . \|(T-T_n)x\|$.

Comme $T_n \xrightarrow{p} T$ on a $\|(T-T_n)x\| \to 0$ si $n \to \infty$.

Pour que $\|x_n-x\| \to 0$ il faut et il suffit que $\sup\{\|R(T_n,z)\| : n \in \mathbb{N}\}$ soit fini.

De cette façon on est arrivé à la notion de *convergence stable* de $\{T_n\}$ vers T au point $z \in \rho(T)$:

$T_n - z \xrightarrow{s} T-z$, $z \in \rho(T)$, si et seulement si les conditions suivantes sont satisfaites :

i) $T_n \xrightarrow{p} T$.

ii) Pour n assez grand $z \in \rho(T_n)$ et $R(T_n,z)$ est borné par rapport à n.

Propriété 1.2.9

Les deux conditions suivantes sont équivalentes

i) $T_n - z \xrightarrow{s} T-z \qquad z \in \rho(T)$.

ii) $T_n \xrightarrow{p} T$, $z \in \rho(T_n)$ *pour* n *assez grand et*

$R(T_n,z) \xrightarrow{p} R(T,z)$.

Preuve : Triviale. □

Propriété 1.2.10

Si Δ *est un compact dans* $\rho(T)$ *et* $T_n - z \xrightarrow{s} T-z \quad \forall z \in \Delta$,
alors $\exists\, n_0 \in \mathbb{N}$ *tel que* $\sup \{\|R(T_n,z)\| : z \in \Delta, n \geq n_0\}$ *est fini.*

Preuve : Cf. Chatelin 1983, p. 232. □

On considère maintenant le problème spectral.

Soient $T \in L(X)$, $z \in \rho(T)$, $\lambda \in \mathrm{Q}\sigma(T)$; Γ_λ est une courbe de Jordan isolant λ tracée dans $\rho(T)$ et $\{T_n\}$ une suite dans $L(X)$ telle que

$$T_n - z \xrightarrow{s} T-z \qquad \forall z \in \Gamma_\lambda .$$

Posons, pour simplifier :

$$R_n(z) = R(T_n,z).$$

Pour n assez grand on définit les opérateurs suivants qui appartiennent à $L(X)$:

$$P_n := \frac{-1}{2\pi i} \int_{\Gamma_\lambda} R_n(z)dz$$

$$S_n(z) := R_n(z)\ (1-P_n).$$

Propriété 1.2.11

Si $T_n - z \xrightarrow{s} T-z$ $\forall z \in \Gamma_\lambda$ *alors* $P_n \xrightarrow{p} P$ $\quad S_n(z) \xrightarrow{p} S(z)$.

Preuve : $\forall x \in X$ on a

$$\|(P_n-P)x\| \leq \frac{1}{2\pi} \int_{\Gamma_\lambda} \|R_n(z)(T-T_n)R(z)x\ dz\| .$$

Soient $\ell = \frac{1}{2\pi} \int_{\Gamma_\lambda} |dz|$, $\omega = \text{Sup}\ \{\|R_n(z)\| : z \in \Gamma_\lambda,\ n \geq n_0\}$
et $z_0 \in \Gamma_\lambda$ tel que $\|R(z_0)x\| = \text{Sup}\ \{\|R(z)x\| : z \in \Gamma_\lambda\}$.
On pose $y_0 = R(z_0)x$ et l'on a

$$\|(P_n-P)x\| \leq \ell\omega\|(T-T_n)y_0\| \to 0 \text{ si } n \to \infty.$$

Finalement $S_n(z) \xrightarrow{p} S(z)$ car $R_n(z) \xrightarrow{p} R(z)$ et $P_n \xrightarrow{p} P$.

□

L'étude de la convergence des sous-espaces invariants se fera à l'aide de la notion d'*ouverture*. Si l'on pose $M_n = P_nX$ alors l'*ouverture* entre M et M_n est définie par :

$$\theta(M,M_n) := \text{Max}\ \{\delta(M,M_n),\ \delta(M_n,M)\}$$

où :

$$\delta(M,M_n) = \text{Sup}\ \{\text{Inf}\{\|x-x_n\| : x_n \in M_n\} \quad : x \in M,\ \|x\| = 1\}$$

$$\delta(M_n,M) = \text{Sup}\ \{\text{Inf}\{\|x-x_n\| : x \in M\} \quad : x_n \in M_n,\ \|x_n\| = 1\}.$$

Propriété 1.2.12

$$\theta(M,M_n) \leq \text{Max } \{\|(P-P_n)P\|,\ \|(P-P_n)P_n\|\} .$$

Preuve : Il suffit de voir que

$$\delta(M,M_n) \leq \|(P-P_n)P\|$$

$$\delta(M_n,M) \leq \|(P-P_n)P_n\| . \qquad \square$$

Pour l'étude des approximations numériques du problème de valeurs propres sera d'une importance capitale la notion de convergence introduite par Chatelin (Cf. Chatelin 1983, pp. 228-253) : la *convergence fortement stable* sur Γ_λ :

$T_n - z \xrightarrow{fs} T-z$ sur Γ_λ si et seulement si $\{T_n\}$ satisfait aux conditions qui suivent :

i) $T_n - z \xrightarrow{s} T-z \qquad \forall z \in \Gamma_\lambda$.

ii) Pour n assez grand dim M_n = m.

Propriété 1.2.13

Si $T_n - z \xrightarrow{fs} T-z$ sur Γ_λ *alors* $P_n \xrightarrow{cc} P$.

Preuve : Evidente, d'après les propriétés 1.2.8 et 1.2.11. $\square$

Propriété 1.2.14

Si $T_n - z \xrightarrow{fs} T-z$ *sur* Γ_λ *alors*

i) $\theta(M,M_n) \to 0$ *lorsque* $n \to \infty$.

ii) Pour n *assez grand,* Δ_λ *-la region bornée dont la frontière est* Γ_λ*- contient exactement* m *valeurs propres de* T_n *comptées leurs multiplicités algébriques.*

Preuve : i) P étant de rang fini est compact. D'après les propriétés 1.2.6 et 1.2.12 on conclut $\theta(M,M_n) \to 0$ lorsque $n \to \infty$.

ii) Cf. Chatelin 1983, p.234.

□

La propriété précédente montre la convergence des sous-espaces invariants M_n vers M au sens de l'ouverture et la convergence des valeurs propres de T_n vers λ si l'on considère que Γ_λ peut être prise comme un cercle de centre λ et de rayon arbitrairement petit.

Dans le cas des opérateurs compacts on peut caractériser la convergence fortement stable autour de tous les points de $Q\sigma(T)$ par la convergence au sens suivant, dite *convergence uniformément radiale* dans $\rho(T)$:

$T_n - z \xrightarrow{u\sigma} T-z$ dans $\rho(T)$ si et seulement si les conditions suivantes sont vérifiées :

i) $T_n - z \xrightarrow{s} T-z \qquad \forall z \in \rho(T)$.

ii) $\forall \Delta$ compact dans $\rho(T)$: $\mathrm{Sup}\{r_\sigma((T-T_n)R(z)) : z \in \Delta\} \to 0$ lorsque $n \to \infty$.

Propriété 1.2.15

Si $T_n - z \xrightarrow{u\sigma} T-z$ *dans* $\rho(T)$ *alors* $\forall \Delta$ *compact dans* $\rho(T)$ $\mathrm{Sup}\{r_\sigma((T_n-T)R_n(z)) : z \in \Delta\} \to 0$ si $n \to \infty$.

Preuve : Cf. Chatelin 1983 p. 254. □

Le théorème de caractérisation qui suit est dû à Lemordant 1980.

Propriété 1.2.16

Si $T \in L(X)$ *est compact alors les conditions suivantes sont équivalentes*

i) $T_n - z \xrightarrow{u\sigma} T-z$ *dans* $\rho(T)$.

ii) $\forall \lambda \in Q\sigma(T) \quad T_n - z \xrightarrow{fs} T-z$ *sur* Γ_λ.

Preuve : Cf. Chatelin 1983 p. 248. □

Exemple 1.2.3 : Si $T_n \xrightarrow{\|\|\|} T$ alors $T_n - z \xrightarrow{u\sigma} T-z$ dans $\rho(T)$. En effet, on a pour $z \in \rho(T)$

$$T_n - z = [1-(T-T_n)R(z)](T-z) .$$

Soit $\varepsilon > 0$ tel que $\varepsilon < 1/\text{Sup}\{\|R(z)\| : z \in \Delta\}$ où Δ est un compact de $\rho(T)$. Alors pour $z \in \Delta$

$$r_\sigma((T-T_n)R(z)) \leq \|(T-T_n)R(z)\| < 1$$

si n est assez grand pour que $\|T-T_n\| < \varepsilon$.
D'après la propriété 1.1.7 on a $z \in \rho(T_n)$ et

$$R_n(z) = R(z) \sum_{i=0}^{\infty} ((T-T_n)R(z))^i .$$

En outre

$$\|R_n(z)\| \leq \frac{\|R(z)\|}{1-\varepsilon\|R(z)\|} \quad \text{donc} \quad T_n - z \xrightarrow{s} T-z \qquad \forall z \in \rho(T).$$

L'inégalité $r_\sigma((T-T_n)R(z)) \leq \|R(z)\|\|T-T_n\|$ sert finalement à montrer que $\text{Sup}\{r_\sigma((T-T_n)R(z)) : z \in \Delta\} \to 0$ lorsque $n \to \infty$.

□

Exemple 1.2.4 : Si $T_n \xrightarrow{cc} T$ alors $T_n - z \xrightarrow{u\sigma} T-z$ dans $\rho(T)$: d'après les propriétés 1.2.5 et 1.2.7 on a $\|((T-T_n)R(z))^2\| \to 0$ si $n \to \infty$ uniformément dans tout compact de $\rho(T)$. Ceci montre que $r_\sigma((T-T_n)R(z)) \to 0$ uniformément dans tout compact de $\rho(T)$.
Donc

$$R_n(z) = R(z)[1-(T-T_n)R(z)]^{-1} =$$

$$= R(z)[1+(T-T_n)R(z)] \sum_{i=0}^{\infty} [(T-T_n)R(z)]^{2i}$$

ce qui montre que $T_n - z \xrightarrow{s} T-z \quad \forall z \in \rho(T)$. □

On déduit de ces deux exemples que pour un opérateur $T \in L(X)$ compact les approximations T_n^P, T_n^S et T_n^G définies dans l'exemple 1.2.1. sont fortement stables autour de toute valeur propre $\lambda \in Q\sigma(T)$ aussi bien que les approximations T_n^N et T_n^F de l'exemple 1.2.2. dans le cas d'un opérateur intégral de Fredholm à noyau continu.

1.3 BORNES D'ERREUR DES ELEMENTS PROPRES

Dans cette section on va démontrer quelques bornes d'erreur théoriques dont nous aurons besoin pour montrer la convergence des méthodes numériques proposées dans le chapitre 3. On remarque pourtant que ces bornes sont des *bornes a priori* et ne sont pas calculables en pratique.

Soient $T \in L(X)$, $\lambda \in Q\sigma(T)$ isolée par Γ_λ et $\{T_n\}$ telle que $T_n - z \xrightarrow{fs} T - z$ sur Γ_λ. Il existe donc un entier n_0 tel que toutes les constantes suivantes sont finies :

$$\nu := \text{Sup}\{\|T_n\| : n \geq n_0\}$$

$$\eta := \text{Sup}\{\|P_n\| : n \geq n_0\}$$

$$\ell := \frac{1}{2\pi} \int_{\Gamma_\lambda} |dz|$$

$$c_\lambda := 2\ell \, \text{Sup}\{\|R_n(z)\| : z \in \Gamma_\lambda, n \geq n_0\} \text{Sup}\{\|R(z)\| : z \in \Gamma_\lambda\}.$$

Propriété 1.3.1

Pour n *assez grand,* $\forall \phi_n \in M_n$ *tel que* $\|\phi_n\| = 1$ $\exists$ $\phi^{(n)} \in M$ *tel que*

i) $P_n \phi^{(n)} = \phi_n$.

ii) $\{\phi^{(n)}\}$ *est une suite bornée.*

Preuve : Soit $x \in M$ tel que $\|x\| = 1$. Alors

$$|1 - \|P_n|_M x\|\ | = |\ \|x\| - \|P_n|_M x\|\ | \leq \|x - P_n|_M x\| =$$

$$= \|(P - P_n)Px\| \leq \|(P - P_n)P\| \to 0 \text{ si } n \to \infty. \text{ Il existe donc un entier}$$

n_* tel que $\forall n \geq n_*$

$$-\frac{1}{2} < 1 - \|P_n|_M x\| < \frac{1}{2}$$

donc $\frac{1}{2}\|x\| < \|P_n|_M x\|$, ce qui montre que $P_n|_M : M \to M_n$ est injective pour $n \geq n_*$. Mais $\dim M_n = m$ donc $P_n|_M$ est une bijection entre M et M_n. Ceci démontre i). Or l'inégalité précédente implique $\|(P_n|_M)^{-1}\| \leq 2$ donc pour $n \geq n_*$:

$$\|\phi^{(n)}\| = \|(P_n|_M)^{-1}\phi_n\| \leq 2 \text{ car } \|\phi_n\| = 1.$$

□

Propriété 1.3.2

La suite $\{\phi^{(n)}\}$ *définie par* $\phi^{(n)} = (P_n|_M)^{-1}\phi_n$ *où* $\phi_n \in M_n$ *et* $\|\phi_n\| = 1$ *vérifie pour* n *assez grand :*

$$\|\phi^{(n)} - \phi_n\| \leq c_\lambda \|(T-T_n)P\| .$$

Preuve : Il suffit de voir que

$$\phi^{(n)} - \phi_n = (P-P_n)P\phi^{(n)} = -\frac{1}{2\pi i}\int_{\Gamma_\lambda} (R(z)-R_n(z))P\phi^{(n)}dz$$

$$= -\frac{1}{2\pi i}\int_{\Gamma_\lambda} R_n(z)(T-T_n)PR(z)\phi^{(n)}dz \quad \text{car}$$

P et R(z) commutent. □

Propriété 1.3.3

La suite $\{\phi^{(n)}\}$ *vérifie* $\|\phi^{(n)}\| \to 1$ *si* $n \to \infty$.

Preuve : $|\ 1-\|\phi^{(n)}\|\ | = |\ \|\phi_n\|-\|\phi^{(n)}\|\ | \leq \|\phi_n - \phi^{(n)}\|$

$$\leq c_\lambda \|(T-T_n)P\| \to 0 \text{ si } n \to \infty.$$ □

Pour n assez grand $\phi^{(n)} \neq 0$ donc on peut définir la *suite normalisée* $\hat{\phi}^{(n)} := \frac{1}{\|\phi^{(n)}\|}\phi^{(n)}$.

Propriété 1.3.4

Pour n *assez grand* $\|\hat{\phi}^{(n)} - \phi_n\| \leq 2\, c_\lambda \|(T-T_n)P\|$.

Preuve : $\|\hat{\phi}^{(n)} - \phi^{(n)}\| = |1-\|\phi^{(n)}\|\ | \leq c_\lambda \|(T-T_n)P\|$ d'après la preuve de la propriété 1.3.3. Mais $\|\hat{\phi}^{(n)} - \phi_n\| \leq \|\hat{\phi}^{(n)} - \phi^{(n)}\| + \|\phi^{(n)} - \phi_n\|$.

□

Propriété 1.3.5

Si λ *est simple alors pour* n *assez grand il existe* λ_n *, valeur propre simple de* T_n *telle que* $|\lambda - \lambda_n| \leq 2n\|(T-T_n)P\|$.

Preuve : Pour n assez grand dim $M_n = 1$ et il existe λ_n valeur propre simple de T_n isolée par Γ_λ . Soit $\phi_n \in M_n$ tel que $\|\phi_n\| = 1$. $\bar{\lambda}_n$ est une valeur propre simple et isolée de T_n^*, donc il existe ϕ_n^* tel que $T_n^*\phi_n^* = \bar{\lambda}_n\phi_n^*$ et $\langle\phi_n^*,\phi_n\rangle = 1$. Or $\phi^{(n)} = (P_n|M)^{-1}\phi_n$ vérifie $\langle\phi^{(n)},\phi_n^*\rangle = 1$ car la projection P_n s'écrit $P_n x = \langle x,\phi_n^*\rangle\phi_n \quad \forall x \in X$.
(Cf. Propriété 1.1.20 appliquée à T_n). Alors nous avons $\lambda - \lambda_n = \langle(T-T_n)P\phi^{(n)},\phi_n^*\rangle$ et par conséquent

$$|\lambda-\lambda_n| = \|P_n(T-T_n)P\phi^{(n)}\| \le \|P_n\|.\|\phi^{(n)}\| \cdot \|(T-T_n)P\|$$

$$\le 2\eta\|(T-T_n)P\|, \text{ pour n assez grand.} \quad \square$$

Si λ est simple alors la résolvante réduite de T_n au point λ_n est, pour n suffisamment grand :

$$S_n := \lim_{z\to\lambda_n} S_n(z) = \frac{1}{2\pi i}\int_{\Gamma_\lambda} \frac{R_n(z)}{z-\lambda_n}\,dz = [(T_n-\lambda_n)|(1-P_n)X]^{-1}(1-P_n).$$

Propriété 1.3.6

Si λ est simple alors $S_n \xrightarrow{p} S$. En outre pour tout compact Δ de $\rho(T) \cup \{\lambda\}$ il existe un entier n_0 tel que

$\sup\{\|S_n(z)\| : z \in \Delta, n \ge n_0\}$ *est fini.*

Preuve : Cf. Chatelin 1983 , p. 244. $\square$

De ce résultat on déduit (en prenant par exemple $\Delta = \Delta_\lambda$) l'existence d'un entier n_* tel que la constante suivante est finie :

$$\gamma := \sup\{\|S_n\| : n \ge n_*\}.$$

Nous reprendrons les constantes ν, η , γ et c_λ au chapitre 3.

1.4 A PROPOS DES OPERATEURS NON LINEAIRES

Soit D un sous-ensemble non vide de l'espace de Banach X. Une application $F : D \to X$ sera dite un *opérateur* dans X de domaine $Dom(F) = D$. Bien entendu un tel sous-ensemble D n'est pas nécessairement un sous-espace vectoriel de X et, a fortiori, un tel opérateur F n'est pas nécessairement linéaire.

Rappelons que F est *continu* au point $x_0 \in D$ si et seulement si $\forall \varepsilon > 0 \quad \exists \delta > 0$ tel que $\forall x \in D$ la condition $\|x-x_0\| < \delta$ implique $\|F(x)-F(x_0)\| < \varepsilon$. Or, F est dit *uniformément continu* sur D si et seulement si $\forall \varepsilon > 0 \quad \exists \delta > 0$ tel que $\forall x_1, x_2 \in D$ l'inégalité $\|x_1-x_2\| < \delta$ entraîne l'inégalité $\|F(x_1)-F(x_2)\| < \varepsilon$.

Exemple 1.4.1 : L'opérateur $F : D \to X$ est dit *lipschitzien* sur D si et seulement si il existe une constante positive ℓ_F, appelée *constante de Lipschitz* pour F, telle que $\forall x_1, x_2 \in D \quad \|F(x_1)-F(x_2)\| \leq \ell_F \|x_1-x_2\|$. Un tel opérateur est certainement uniformément continu et, a fortiori, continu en tout point de D. Si l'on peut choisir une constante ℓ_F dans l'intervalle $]0,1[$ alors on dit que F est *contractant* sur D.

□

On considère maintenant un opérateur $F : D \to X$ dont le domaine D est un ouvert de X. F est un opérateur *différentiable* au point $x \in D$ si et seulement si il existe un opérateur (que l'on montre unique lorsqu'il existe), $D_x F \in L(X)$, appelé *dérivée de Fréchet* de F en x, qui vérifie :

$$\frac{\|F(y) - F(x) - (D_x F)(y-x)\|}{\|y-x\|} \to 0 \text{ si } y \to x.$$

Soit W un ouvert inclus dans D. On dit que F est de *classe* C^1 sur W si et seulement si l'application $x \in W \to D_x F \in L(X)$ est continue sur W.

Propriété 1.4.1

Soit $F : D \to X$ *de classe* C^1 *sur un ouvert* W *inclus dans* D . *Soit* $x \in W$ *tel que* $D_x F \in Iso(X)$. *Alors* F *est bijectif en tant qu'opérateur défini sur un certain voisinage de* x *et à valeurs sur un certain voisinage de* $F(x)$. *Ainsi restrient l'opérateur inverse est de classe* C^1 *sur son domaine.*

Preuve : Cf. Lang 1977 p. 120. □

Propriété 1.4.2

Soit D *un ouvert convexe de* X *et* $F : D \to X$ *un opérateur de classe* C^1 *sur* . *Notons* $J(x) := D_x F \quad \forall x \in D$. *Alors* $\forall x,y \in D$

$$F(y)-F(x) = \int_0^1 J(x+t(y-x))\,(y-x)dt .$$

Preuve : Soit $g : [0,1] \to X$ définie par $g(t) = F(x+t(y-x))$.
La propriété 1.4.2. ne fait que traduire la relation élémentaire suivante :

$$\int_0^1 g'(t)dt = g(1) - g(0).$$ □

1.5 REMARQUES ET COMMENTAIRES BIBLIOGRAPHIQUES

La référence que nous avons consultée le plus fréquemment, pour ce qui est de l'approximation spectrale d'opérateurs linéaires, fut Chatelin 1983. Elle contient un recueil très complet des résultats de base d'analyse fonctionnelle et de théorie spectrale aussi bien que l'état actuel de la recherche en analyse numérique concernant l'approximation des éléments propres des opérateurs intégraux et différentiels. Comme référence plus classique on doit citer Kato 1966. Les ouvrages de Lang 1977 et de Cartan 1972 sont, bien que d'un niveau plus élémentaire, des références très utiles en matière de calcul différentiel et d'analyse réelle. On y trouve aisément les théorèmes principaux de ces domaines.

Lors de la définition de l'opérateur projection spectrale nous avons introduit l'intégrale, sur une courbe de Jordan, d'une fonction de variable complexe et dont les valeurs appartiennent à l'espace de Banach $L(X)$. Il s'agit, en fait, des fonctions analytiques sur des domaines ouverts du plan complexe. Une référence importante à ce sujet est Dunford et Schwartz 1958 (Vol. 1). Signalons en outre que la semi-continuité supérieure des applications $z \to r_\sigma(R(T,z))$ sur $\rho(T)$ et $z \to r_\sigma(S(T,z))$ sur $\rho(T) \cup \{\lambda\}$ découle du fait qu'elles sont des fonctions sous-harmoniques et ceci est une conséquence du résultat plus général suivant (cf. Vesentini 1968) :

> Si D est un ouvert dans C et si $f : D \to L(X)$ est une fonction analytique sur D alors l'application $z \to r_\sigma(f(z))$ est sous-harmonique sur D.

Nous dirons maintenant quelques mots à propos des notions de convergence d'opérateurs. On a présenté la convergence ponctuelle, en norme, collectivement compacte, stable en $z \in \rho(T)$, fortement stable autour de $\lambda \in Q\sigma(T)$ et uniformément radiale dans $\rho(T)$. Or, dans la littérature on retrouve aussi d'autres définitions telles que la *convergence compacte* $T_n \xrightarrow{c} T$ et la *convergence régulière* $T_n \xrightarrow{r} T$. Nous ne les avons pas prises en considération parce que $T_n \xrightarrow{r} T$ est équivalente à $T_n - z \xrightarrow{s} T - z$ $z \in \rho(T)$ et que, quand $T_n - T$ est un opérateur compact -ce qui est le cas auquel nous nous intéressons lors des applications- alors $T_n \xrightarrow{c} T$ équivaut à $T_n \xrightarrow{cc} T$. Citons dans ce domaine les références Anselone 1971 et Anselone et Ansorge 1979. La notion de *convergence fortement stable* autour de $\lambda \in Q\sigma(T)$ est due à Chatelin 1983.

La notion de convergence fortement stable, dans le sens de Chatelin, sera d'une importance capitale lors de l'étude de la convergence des méthodes numériques de raffinement itératif que nous proposerons au Chapitre 3.

Finalement, nous ferons quelques remarques sur la notion d'*ouverture* (en anglais "gap") entre deux sous-espaces vectoriels fermés. Pour en donner une interprétation géométrique supposons que M et N soient deux sous-espaces de dimension finie r de l'espace euclidien complexe C^n ($n \geq 2r \geq 2$) muni du produit scalaire habituel

$$\langle x,y\rangle_2 = \sum_{i=1}^{n} x_i \bar{y}_i, \quad x = (x_1,\ldots,x_n),\ y = (y_1,\ldots,y_n)$$

et de la norme induite

$$\|x\|_2 = \langle x,x\rangle_2^{1/2}$$

Il existe une base $\{x^{(1)},\ldots,x^{(r)}\}$ de M et une base $\{y^{(1)},\ldots,y^{(r)}\}$ de N telles que

$$\text{Max}\{|\langle x,y\rangle_2| : x \in M,\ y \in N,\ \|x\|_2 = \|y\|_2 = 1\} = |\langle x^{(1)},y^{(1)}\rangle_2|$$

$$\text{Max}\{|\langle x,y\rangle_2| : x \in M,\ y \in N,\ \|x\|_2 = \|y\|_2 = 1,\ \langle x,x^{(j)}\rangle_2 = \langle y,y^{(j)}\rangle_2 = 0,\ j = 1,\ldots,i-1\} = |\langle x^{(i)},y^{(i)}\rangle_2|$$

pour $i = 2,\ldots,r$. (C. Björck et Golub 1973, Davis et Kahan 1970, Stewart 1973). On définit les angles canoniques $\alpha_1,\ldots,\alpha_r$ entre M et N par les conditions suivantes

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_r\ ,\quad \alpha_i \in [0,\pi/2],\quad \cos\alpha_i = |\langle x^{(i)},y^{(i)}\rangle_2|$$

Or, l'ouverture entre M et N (par rapport à la norme $\|\ \|_2$) est $\theta(M,N) = \sin\alpha_r$. On montre aussi que $\theta(M,N) = \|Q_M - Q_N\|_2$ où Q_M (resp. Q_N) est la *projection orthogonale* sur M (resp. sur N).

En général, dans un espace de Banach quelconque, si l'un des deux sous-espaces vectoriels M ou N est de dimension finie alors la condition $\theta(M,N) < 1$ entraîne dim M = dim N. Une référence où l'on peut approfondir l'étude de la notion d'ouverture est Kato 1966.

CHAPITRE 2

SUR LA METHODE DE NEWTON ET LE PRINCIPE DE CORRECTION DU RESIDU

Description et présentation d'exemples

INTRODUCTION

La méthode de Newton est peut être l'une des méthodes les plus célèbres en ce qui concerne le calcul numérique itératif des racines isolées d'un opérateur non linéaire suffisamment régulier. On fait dans ce chapitre une présentation de cette méthode dans un espace de Banach complexe et sous des hypothèses qui assurent soit une *convergence superlinéaire*, soit la *convergence quadratique* que l'on associe fréquemment à cette méthode.

Il est bien connu que les aspects les plus délicats de la méthode de Newton sont, d'une part, sa *convergence locale*, ce qui veut dire que le point de départ des itérations doit être choisi suffisamment proche de la solution exacte pour que la méthode converge vers elle, et, d'autre part, qu'elle nécessite l'*inversion de la dérivée* de l'opérateur dont on recherche une racine, la dérivée étant évaluée à l'itéré en cours, cette inversion doit être refaite à chaque itération.

Dans le but de surmonter numériquement la dernière difficulté citée ci-dessus on présente une classe de méthodes itératives que l'on appelle *méthodes de correction du résidu*, pour terminer en proposant une *méthode mixte* qui, d'une part, tient compte des caractéristiques de la méthode de Newton et, d'autre part, emploie une technique de correction du résidu lors de l'inversion de la dérivée au cours de l'itération de Newton.

On a illustré toutes ces présentations par des exemples qui en outre servent à préparer le chemin qui nous conduira aux méthodes que l'on va proposer au chapitre suivant pour le raffinement itératif d'un vecteur propre approché d'un opérateur compact.

2.1 GENERALITES

On considère un sous-ensemble non vide D dans un espace de Banach X.

Etant donné un opérateur $F : D \to X$ supposons qu'il existe un élément $\xi \in D$ solution de l'équation

$$F(x) = 0 \qquad (2.1.1.)$$

Supposons en outre que l'on ait une suite $\{\xi^k\}$ qui converge vers ξ. On s'intéresse à l'étude de la vitesse de convergence et pour cela on introduit les notions suivantes.

La convergence $\xi^k \to \xi$ est dite *linéaire* si et seulement si il existe une constante α dans l'intervalle $]0,1[$ telle que

$$\forall k \geq 0 \qquad \|\xi^{k+1}-\xi\| \leq \alpha\|\xi^k-\xi\|$$

La convergence $\xi^k \to \xi$ est dite *superlinéaire* si et seulement si il existe une suite $\{\alpha_k\}$ de nombres réels positifs tel que

i) $\forall k \geq 0 \qquad \|\xi^{k+1}-\xi\| \leq \alpha_k \|\xi^k-\xi\|$

ii) $\alpha_k \to 0 \qquad$ si $k \to \infty$.

La convergence $\xi^k \to \xi$ est dite *quadratique* si et seulement si il existe une constante $\beta > 0$ telle que

$$\forall k \geq 0 \qquad \|\xi^{k+1}-\xi\| \leq \beta\|\xi^k-\xi\|^2$$

Les implications qui ont lieu entre ces notions sont tout à fait évidentes, la convergence linéaire étant la plus faible et la convergence quadratique la plus forte.

Propriété 2.1.1

Si la convergence $\xi^k \to \xi$ *est superlinéaire, alors*

$$\frac{\|\xi^{k+1}-\xi^k\|}{\|\xi^k-\xi\|} \to 1 \quad \textit{lorsque } k \to \infty$$

Preuve : $\|\xi^k-\xi\| \le \|\xi^{k+1}-\xi^k\| + \|\xi^{k+1}-\xi\|$

$$\le \|\xi^{k+1}-\xi^k\| + \alpha_k\|\xi^k-\xi\|$$

Donc,

$$(1-\alpha_k)\|\xi^k-\xi\| \le \|\xi^{k+1}-\xi^k\| \le \|\xi^k-\xi\|+\|\xi^{k+1}-\xi\|$$

$$\le (1+\alpha_k)\|\xi^k-\xi\| \qquad \square$$

Si $\xi \neq 0$ alors la propriété précédente nous dit que quand la convergence $\xi^k \to \xi$ est superlinéaire alors, pour k suffisamment grand, on peut en pratique estimer *l'erreur relative* $\|\xi^k-\xi\|/\|\xi\|$ en substituant ξ^{k+1} à ξ car $\|\xi^{k+1}\| \to \|\xi\|$ et donc

$$\frac{\|\xi^k-\xi^{k+1}\|/\|\xi^{k+1}\|}{\|\xi^k-\xi\|/\|\xi\|} \to 1 \quad \text{si } k \to \infty.$$

2.2 LA METHODE DE NEWTON

Formellement la *méthode de Newton* pour la résolution de l'équation (2.1.1.) s'écrit :

$$\left.\begin{array}{l} \xi^0 \text{ donné, assez proche de la solution } \xi \\ \xi^{k+1} := \xi^k - J(\xi^k)^{-1}F(\xi^k) \quad k \ge 0 \end{array}\right\} \quad (2.2.1.)$$

où $J(x) := D_xF$ est supposé dans $Iso(X)$ pour $x \in \{\xi^k : k \ge 0\}$.

Les propositions suivantes donnent des conditions suffisantes pour que la suite de Newton $\{\xi^k\}$ définie par (2.2.1) soit convergente vers ξ .

Propriété 2.2.1

Supposons que

i) $J(\xi) \in \mathrm{Iso}(X)$

ii) $\exists r > 0$ *tel que* $J : B(\xi,r) \to L(X)$ *est uniformément continu sur* $B(\xi,r)$

Alors il existe un réel $\rho > 0$ *tel que* $\forall \xi^0 \in B(\xi,\rho)$ *la suite (2.2.1) converge vers* ξ *, la convergence étant superlinéaire.*

Preuve : D'après l'hypothèse ii) il existe $r_1 \in]0,r[$ tel que $\|x-\xi\| < r_1 \Rightarrow \|J(x)-J(\xi)\| < 1/\|J(\xi)^{-1}\|$ d'où $J(\xi) \in \mathrm{Iso}(X)$ pour $x \in B(\xi,r_1)$ (cf. Propriété 1.1.11).
L'application $x \in B(\xi,r_1) \to J(x)^{-1} \in L(X)$ étant continue, ils existent $r_2 \in]0,r_1[$ et $\mu > 0$ tels que $\mathrm{Sup}\{\|J(x)^{-1}\| : x \in B(\xi,r_2)\} \leq \mu$
Mais, de nouveau à partir de ii), on a l'existence de $\rho \in]0,r_2[$ tel que $\forall x,y \in B(\xi,\rho)$ $\|J(x)-J(y)\| < \frac{1}{2\mu}$.

Soit $\xi^k(t) := \xi+t(\xi^k-\xi)$ $t \in [0,1]$. Alors, d'après la propriété 1.4.2., on peut écrire la relation suivante :

$$\xi^{k+1}-\xi = J(\xi^k)^{-1} \int_0^1 [J(\xi^k(t))-J(\xi^k)](\xi^k-\xi)dt$$

si $\xi^k \in B(\xi,\rho)$. Ceci montre que $\|\xi^{k+1}-\xi\| \leq \frac{1}{2}\|\xi^k-\xi\|$ donc $\xi^k \to \xi$ (de façon linéaire) si $\xi^0 \in B(\xi,\rho)$.
Or la convergence est superlinéaire car

$$\|\xi^{k+1}-\xi\| \leq \alpha_k \|\xi^k-\xi\|$$

où

$$\alpha_k = \mu\, \mathrm{Sup}\{\|J(\xi^k(t))-J(\xi^k)\| : 0 \leq t \leq 1\} \to 0 \quad \text{si}$$

$k \to \infty$.

□

Propriété 2.2.2

Supposons que

i) $J(\xi) \in \mathrm{Iso}(X)$

ii) $\exists\, r > 0$ *tel que* $J : B(\xi,r) \to L(X)$ *est lipschitzien sur* $B(\xi,r)$

Alors il existe un réel $\rho > 0$ *tel que* $\forall \xi^0 \in B(\xi,\rho)$ *la suite (2.2.1) converge vers* ξ *de façon quadratique.*

Preuve : On suit la démonstration précédente mais cette fois-ci on a de plus

$$\alpha_k = \mu \ell_J \, \mathrm{Sup}\{\|\xi^k(t)-\xi^k\| : 0 \le t \le 1\} = \mu \, \ell_J \|\xi^k-\xi\|,$$

ℓ_J étant une constante de Lipschitz de J sur $B(\xi,r)$.

□

Exemple 2.2.1 : Soient $T \in L(X)$ un opérateur compact et $\lambda \in Q\sigma(T)$ telle que λ est simple et $\lambda \neq 0$. On reprend les notations simplifiées $P = P(T,\lambda)$, $S = S(T,\lambda)$, $M = M(T,\lambda)$.
Supposons donné $\psi \in X^*$ tel que $M \cap \mathrm{Ker}\ \psi = \{0\}$. On définit sur X l'opérateur F suivant

$$F(x) = Tx - \langle Tx,\psi\rangle x \qquad \forall x \in X$$

qui est évidemment non linéaire et de classe C^1 sur X, sa dérivée notée J(x), étant donnée par

$$J(x)u = Tu - \langle Tx,\psi\rangle u - \langle Tu,\psi\rangle x \quad \forall x,u \in X.$$

Soit $\phi \in M$ normalisé par $\langle\phi,\psi\rangle = 1$. L'existence d'un tel vecteur propre de T est assurée par un théorème de Riesz (Cf. Chatelin 1983, p. 70). ϕ est une solution de l'équation $F(x) = 0$ et l'on a les résultats suivants :

i) $J(\phi) \in \mathrm{Iso}(X)$

ii) J est lipschitzien sur X

Effectivement, on voit que $\langle T\phi,\psi\rangle = \lambda$, donc

$$J(\phi)u = (T-\lambda)u - \langle Tu,\psi\rangle\phi \qquad \forall u \in X$$

T étant compact, il suffit d'appliquer les propriétés 1.1.5, 1.1.22 et 1.1.23 pour conclure : $J(\phi)^{-1} \in L(X)$. D'autre part

$\forall x_1,x_2 \in X \quad (J(x_1)-J(x_2))u = \langle T(x_2-x_1),\psi\rangle u + \langle Tu,\psi\rangle(x_2-x_1)$

d'où $\|J(x_1)-J(x_2)\| \le 2\|T\|.|\psi|.\|x_1-x_2\|$. Alors $\ell_J = 2\|T\|\,|\psi|$ est une constante de Lipschitz pour J sur X.

D'après la propriété 2.2.2 il existe $\rho > 0$ tel que la suite

$$\phi^0 \in B(\phi,\rho)$$

$$\phi^{k+1} := \phi^k - J(\phi^k)^{-1}(T\phi^k - \langle T\phi^k,\psi\rangle\phi^k) \qquad k \geq 0$$

converge vers ϕ de façon quadratique. Remarquons que la suite $\{\langle T\phi^k,\psi\rangle\}$ converge, dans ces conditions, vers λ.

□

Exemple 2.2.2 : Sous les mêmes hypothèses que précédemment, il existe $r > 0$ tel que

$$\mathrm{Inf}\{|\langle Tx,\psi\rangle| : x \in B(\phi,r)\} \geq \frac{|\lambda|}{2} > 0,$$

car $\lambda = \langle T\phi,\psi\rangle \neq 0$. On peut donc définir un opérateur $\tilde{F}$ de domaine $D = B(\phi,r)$ de la manière suivante

$$\tilde{F}(x) = x - \frac{1}{\langle Tx,\psi\rangle}\, Tx \qquad \forall x \in D.$$

$\tilde{F}$ est non linéaire et de classe C^1 sur D. Sa dérivée, $\tilde{J}(x)$, est définie par

$$\tilde{J}(x)u = u - \frac{1}{\langle Tx,\psi\rangle}\, Tu + \frac{\langle Tu,\psi\rangle}{\langle Tx,\psi\rangle^2}\, Tx \qquad \forall u \in X, \quad \forall x \in D.$$

ϕ est une solution de l'équation $\tilde{F}(x) = 0$ et l'on a aussi :

i) $\tilde{J}(\phi) \in \mathrm{Iso}(X)$

ii) $\tilde{J}$ est lipschitzien sur D.

Le fait que $\tilde{J}(\phi)^{-1} \in L(X)$ est une conséquence des propriétés 1.1.5, 1.1.22 et 1.1.23 comme dans l'exemple 2.2.1. On peut vérifier, d'autre part que

$$\ell_{\tilde{J}} = \frac{8|\psi|\,\|T\|^2}{|\lambda|^2} + \frac{48\|T\|^3|\psi|^2(r+\|\phi\|)}{|\lambda|^3}$$

est une constante de Lipschitz pour $\tilde{J}$ sur D. On conclut qu'il existe $\rho > 0$ tel que

$$\phi^0 \in B(\phi,\rho)$$

$$\phi^{k+1} := \phi^k - J(\phi^k)^{-1}\left(\phi^k - \frac{1}{\langle T\phi^k,\psi\rangle} T\phi^k\right) \qquad k \geq 0$$

converge vers ϕ de façon quadratique et l'on en déduit la convergence de $\{\langle T\phi^k,\psi\rangle\}$ vers λ. □

On remarque que la condition ii) de la propriété 2.2.2., J lipschitzien sur un voisinage de la solution ξ, est une condition sans laquelle on risque de perdre la convergence quadratique. Cela se voit, par exemple, dans le cas suivant (Cf. Moré et Sorensen 1982, p. 9) :
$\xi = 0$ est une solution de l'équation 2.1.1., l'opérateur $F : \mathbb{R} \to \mathbb{R}$ étant donné par

$$F(x) = \begin{cases} x(1+\mathrm{Log}|x|) & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases}$$

La convergence de la suite (2.2.1) est superlinéaire, d'après la propriété 2.2.1 mais elle n'est pas quadratique car la suite $\left\{\frac{|\xi^{k+1}|}{|\xi^k|^{1+q}}\right\}$ est *non bornée* $\forall q > 0$.

2.3 LE PRINCIPE DE CORRECTION DU RESIDU

On garde les notations générales de la section 2.1.

Un opérateur G sera dit un *inverse approché local* de F autour de la solution ξ du problème (2.1.1.) si et seulement si la suite

$$\left.\begin{aligned} \xi^0 &:= G(0) \\ \xi^{k+1} &:= G(0) + (1-GF)(\xi^k) \qquad k \geq 0 \end{aligned}\right\} \quad (2.3.1.)$$

est bien définie $\forall k \geq 0$ et converge vers ξ lorsque $k \to \infty$.

Les propriétés 2.3.1. et 2.3.4. donnent des conditions suffisantes pour que G soit un inverse approché local.

Propriété 2.3.1

S'il existe un réel $\rho > 0$ *tel que*

i) $B(\xi,\rho) \subseteq D$

ii) $F(B(\xi,\rho)) \subseteq Dom(G)$

iii) $G(0) \in B(\xi,\rho)$

iv) $1-GF$ *est contractant sur* $B(\xi,\rho)$

alors G *est un inverse approché de* F *autour de* ξ .

Preuve : Les hypothèses i), ii) et iii) permettent de définir l'opérateur $V : B(\xi,\rho) \to X$ par

$$V(x) = G(0) + x - G(F(x))$$

On vérifie aisément que $V(\xi) = \xi$ et que $V(B(\xi,\rho)) \subseteq B(\xi,\rho)$.
L'hypothèse iv) assure que V est contractant sur $B(\xi,\rho)$.
L'itération (2.3.1.) correspond à l'itération de point fixe de V : $\xi^{k+1} = V(\xi^k)$ $\xi^0 \in B(\xi,\rho)$. Si $\ell_V \in]0,1[$ est une constante de Lipschitz de V sur $B(\xi,\rho)$ alors on en déduit $\|\xi^k-\xi\| \leq (\ell_V)^k\|\xi^0-\xi\|$ d'où le résultat.

□

Propriété 2.3.2

Sous les hypothèses de la propriété 2.3.1,

i) F *est injectif sur* $B(\xi,\rho)$

ii) G *est injectif sur* $F(B(\xi,\rho))$.

Preuve : Si ce n'est pas le cas alors il existe $x_1, x_2 \in B(\xi,\rho)$ tels que

$$(1-GF)(x_1) - (1-GF)(x_2) = x_1 - x_2$$

et l'on contredit l'hypothèse iv).

□

Propriété 2.3.3

Notons $F_B = F_{|B(\xi,\rho)}$. *Sous les hypothèses de la propriété 2.3.1 on a :*
Si $F_B^{-1} : F(B(\xi,\rho)) \to B(\xi,\rho)$ *est un opérateur continu alors* G *est continu sur* $F(B(\xi,\rho))$.

Preuve : $\forall x \in F(B(\xi,\rho))$ on a

$$G(x) = G(0)+F_B^{-1}(x)-V(F_B^{-1}(x))$$

où V est défini dans la preuve de la propriété 2.3.1.

□

Propriété 2.3.4

Si F *et* G *satisfont aux conditions suivantes*

i) $GF \in L(X)$

ii) $r_o(1-GF) < 1$

alors G *est un inverse approché de* F *autour de* ξ .

Preuve : L'hypothèse i) permet d'écrire

$$\xi^k = \left(\sum_{i=0}^{k} U^i\right)\xi^0$$

avec $U := 1-GF \in L(X)$. D'après la propriété 1.1.7 on est assuré, par ii), de la convergence de $\{\xi^k\}$ vers $(1-U)^{-1}\xi^0$.
Mais $(1-U)^{-1}\xi^0 = (GF)^{-1}\, G(0) = (GF)^{-1}\, GF(\xi) = \xi$ car l'existence de $(GF)^{-1}$ entraîne l'unicité de ξ .

□

Remarquons que l'hypothèse i) de la propriété 2.3.4, $GF \in L(X)$, n'implique pas forcément que $F \in L(X)$ ou que $G \in L(X)$.

Propriété 2.3.5

Sous les hypothèses de la propriété 2.3.1 aussi bien que sous celles de la propriété 2.3.4 la convergence de (2.3.1) est linéaire.

Preuve : Si les hypothèses de la propriété 2.3.1 sont satisfaites on a :

$$\|\xi^{k+1}-\xi\| = \|V(\xi^k)-V(\xi)\| \leq \ell_V\|\xi^k-\xi\| \qquad \forall k \geq 0$$

où V et $\ell_V \in]0,1[$ ont été définis au cours de la preuve de cette propriété là. Or, si les hypothèses de la propriété 2.3.4 sont vérifiées, soit $\varepsilon > 0$ tel que $r_\sigma(1-GF)+\varepsilon < 1$ et soit $\|\ \|_*$ une norme telle que $\|1-GF\|_* < r_\sigma(1-GF)+\varepsilon$ (Cf. propriété 1.1.9) Alors on a l'inégalité

$$\|\xi^{k+1}-\xi\|_* < (r_\sigma(1-GF)+\varepsilon)\|\xi^k-\xi\|_*$$

□

On peut interpréter (2.3.1) comme une *technique de raffinement* de la solution numérique $\xi^0 = G(0)$ du problème (2.1.1). Si l'on suppose F continu alors la suite des résidus associée à la suite d'approximations $\{\xi^k\}$ est définie par $r^k := F(\xi^k)$ et converge vers 0 lorsque $k \to \infty$. L'itération (2.3.1) entraîne donc *une correction du résidu* à chaque pas.

Exemple 2.3.1 : On considère un opérateur $T \in L(X)$ et l'équation $(T-z)x = f$ avec $z \in \rho(T)$ et $f \in X$ donnés. On peut écrire cette équation sous la forme (2.1.1) si l'on pose $F(x) = (T-z)x-f$. $F : X \to X$ est un opérateur affine bijectif continu dont l'inverse l'est aussi et vient donné par $F^{-1}(x) = R(z)(x+f)$.
Nous allons construire un inverse approché de F autour de la solution (unique) $u = R(z)f$, par l'intermédiaire d'une approximation T_n de T qui satisfait aux conditions :

i) $T_n - z \xrightarrow{s} T-z$

ii) $r_\sigma((T-T_n)R(z)) \to 0$ si $n \to \infty$.

Soit $G : X \to X$ l'opérateur affine défini par $G(x) = R_n(z)(x+f)$. Il est bien défini pour n suffisamment grand, d'après la condition i) ci-dessus. Alors $1-GF = 1-R_n(z)(T-z) = R_n(z)\ (T_n-T)$
1-GF est un opérateur linéaire continu (même si F et G ne sont pas linéaires mais affines). D'après les propriétés 1.2. et 1.1.6 on conclut à l'existence d'un entier n pour lequel $r_\sigma(1-GF) < 1$. L'opérateur G ainsi construit est un inverse approché de F autour de u selon la propriété 2.3.4. L'itération (2.3.1) s'écrit

$$u^0 := R_n(z)f$$

$$u^{k+1} := u^k - R_n(z)\ ((T-z)u^k - f)$$

Cette méthode de raffinement de u^0 fut proposée par Atkinson 1973 dans le cas où T est un opérateur intégral de Fredholm à noyau continu et T_n son approximation de Nyström.

□

Exemple 2.3.2 : On considère de nouveau l'équation $(T-z)x = f \quad z \in \rho(T)$ pour un opérateur $T \in L(X)$ que cette fois-ci l'on suppose compact. Comme dans l'exemple précédent on pose $F(x) = (T-z)x-f$ et l'on considère pour la construction d'un inverse approché de F autour de $u = R(z)f$, une suite d'opérateurs $\{T_n\}$ dans $L(X)$ telle que $T_n - z \xrightarrow{s} T-z$.

D'après la propriété 1.1.14 appliquée aux opérateurs T_n et $R_n(z)$ on a l'identité, pour $z \in \rho(T_n)$, $z \neq 0$:

$$R_n(z) = \frac{1}{z}\ (R_n(z)T_n - 1)$$

à partir de laquelle, en substituant T à T_n, nous sommes amenés à définir cet opérateur que l'on appelera approximation de Brakhage de la résolvante R(z) :

$$R_n^B(z) := \frac{1}{z}\ (R_n(z)T - 1)$$

Remarquons que $\{R_n^B(z)\}$ converge en norme vers $R(z)$:

$$\|R(z)-R_n^B(z)\| = \|R(z)- \frac{1}{z}(R_n(z)T - R(z)\ (T-z))\|$$

$$= \frac{1}{|z|}\|(R(z)-R_n(z))T\|$$

qui tend vers 0 lorsque $n \to \infty$ car $R_n(z) \xrightarrow{p} R(z)$ et T est compact (Cf. propriété 1.2.6.).

On construit un inverse approché de F à l'aide de $R_n^B(z)$ ainsi
$G(x) = R_n^B(z)\ (x+f) \qquad x \in X.$
Nous avons :

$$r_\sigma(1-GF) \leq \|1-GF\| = \frac{1}{|z|}\|R_n(z)\ (T_n-T)T\|$$

Mais $R_n(z)$ est uniformément borné par rapport à n donc pour n assez grand $r_\sigma(1-GF) < 1$ et G devient un inverse approché de F (Cf. propriété 2.3.4).

On remarque que nous sommes aussi sous les hypothèses de la propriété 2.3.1. L'opérateur $V(x) = G(0)+(1-GF)(x)$ vérifie $\|V(x_1)-V(x_2)\| \leq \frac{1}{|z|}\|R_n(z)(T_n-T)T\|.\|x_1-x_2\|$ donc il est contractant sur X pour n suffisamment grand.

L'itération (2.3.1) s'écrit :

$$u^0 := R_n^B(z)f$$

$$u^{k+1} := u^k-R_n^B(z)\ ((T-z)u^k-f)$$

Il s'agit de l'une des versions de la méthode proposée par Brakhage 1960 pour des opérateurs intégraux compacts.

Un autre opérateur qui vérifie les mêmes propriétés que l'opérateur G défini ci-dessus est le suivant

$$G'(x) = R_n(z)f + R_n^B(z)x \qquad x \in X$$

Seul le point de départ u^0 change : $R_n(z)f$ prend la place de $R_n^B(z)f$ (Cf. Ahués et al 1982). □

2.4 EXEMPLE D'UNE METHODE MIXTE

On considère le problème général (2.1.1) dont une solution est notée ξ . Soit ξ^0 une solution approchée et $Q : X \to X$ une projection sur le sous-espace engendré par ξ^0 le long d'un autre sous-espace fermé donné. On suppose en outre que les conditions suivantes sont vérifiées, $J(x)$ étant la dérivée de F en x.

i) $Q\xi = \xi^0$

ii) $\forall x \in Q^{-1}\{\xi^0\} \quad F(x) \in (1-Q)X$

iii) $\forall x \in Q^{-1}\{\xi^0\} \quad J(x)\big|_{(1-Q)X} \in \mathrm{Iso}((1-Q)X)$

iv) $\exists r > 0$ tel que $J : B(\xi,r) \to L(X)$ est lipschitzien.

Si l'on utilise la méthode de Newton on aura à résoudre à l'itération $k+1$, pour le calcul de ξ^{k+1}, le problème linéaire $J(\xi^k)\delta^k = F(\xi^k)$ et ensuite on aura $\xi^{k+1} = \xi^k - \delta^k$.

Puisque la solution ξ de l'équation $F(x) = 0$ vérifie $Q\xi = \xi^0$ on va raffiner ξ^0 de façon à l'améliorer sur $(1-Q)X$, c'est-à-dire, on construira une suite d'approximations $\{\xi^k\}$ telle que

$$\delta^k := \xi^k - \xi^{k+1} \in (1-Q)X \qquad \forall k \geq 0,$$

ce qui veut dire :

$$Q\xi^k = \xi^0 \qquad \forall k \geq 0.$$

De cette façon là l'inversion de $J(\xi^k)$ va être faite dans $(1-Q)X$ ce qui est possible d'après l'hypothèse iii) ci-dessus. Très souvent ce calcul ne peut être fait exactement. On propose de le faire à l'aide d'une méthode de correction du résidu :

On définit sur (1-Q)X les opérateurs suivants

$$K(u) = J(\xi^k)u - F(\xi^k)$$

G, un inverse approché de K autour de δ^k tel que les hypothèses de la propriété 2.3.4 sont vérifiées. Alors

$$\delta_0^k := G(0)$$

$$\delta_{j+1}^k := G(0) + (1-GK)\delta_j^k \qquad j \geq 0$$

converge, lorsque $j \to \infty$, de façon linéaire, vers $\delta^k = (J(\xi^k)\big|_{(1-Q)X})^{-1}F(\xi^k)$.

Supposons qu'il existe une suite $\{\alpha_j\}$ indépendante de k telle que

$$\|\delta_j^k - \delta^k\| \leq \alpha_j \|\xi^k - \xi\| , \qquad \alpha_j \to 0 \text{ si } j \to \infty.$$

Soit $\rho \in]0,r[$ tel que

$$\text{Sup}\{\|(J(x)\big|_{(1-Q)X})^{-1}\| : x \in Q^{-1}\{\xi^0\} \cap B(\xi,\rho)\} \leq \mu < +\infty$$

$$\forall x_1, x_2 \in B(\xi,\rho) \quad \|J(x_1) - J(x_2)\| < \frac{1}{2\mu}$$

et soit ℓ_J une constante de Lipschitz de J sur $B(\xi,r)$.

Si l'on note i(k) le nombre d'itérations que l'on fait à l'itération k+1 pour le calcul approché de δ^k alors la procédure de raffinement de ξ^0 revient à la suivante :

$$\xi^0 \in B(\xi,\rho) \quad \text{donné}$$

$$\xi^{k+1} = \xi^k - \delta_{i(k)}^k \qquad k \geq 0$$

où

$$\delta_{i(k)}^k = \sum_{\ell=0}^{i(k)} (1-GK)^\ell G(0).$$

Or, on trouve facilement la majoration suivante :

$$\|\xi^{k+1}-\xi\| \leq \ell_J \mu\|\xi^k-\xi\|^2 + \alpha_{i(k)}\|\xi^k-\xi\|$$

d'où on conclut ce qui suit :

Si l'on fixe $i(k) = i_0$, constante telle que $\alpha_{i_0} < 1$ alors $\xi^k \to \xi$ de façon linéaire.

Si $i(k) \to \infty$ lorsque $k \to \infty$ alors $\alpha_{i(k)} \to 0$ et $\xi^k \to \xi$ de façon superlinéaire.

Si $\{\frac{\alpha_{i(k)}}{\|\xi^k-\xi\|}\}$ est une suite bornée alors $\xi^k \to \xi$ de façon quadratique.

Du point de vue pratique rappelons que si $i(k) \to \infty$ lorsque $k \to \infty$ alors d'après la propriété 2.1.1. on peut estimer $\|\xi^k-\xi\|$ par $\|\xi^{k+1}-\xi^k\|$ ce qui rend possible la convergence quadratique.

Dans le chapitre suivant on construira des méthodes de ce type-là pour le raffinement d'un vecteur propre approché d'un opérateur compact. Remarquons d'avance les points les plus délicats de la présentation que l'on vient de faire :

- ξ^0 doit être suffisamment proche de ξ : $\xi^0 \in B(\xi,\rho)$
- ξ dépend de ξ^0 via la projection Q
- L'espace (1-Q)X bouge lorsque ξ^0 bouge.
- Il nous faut pour chaque k un inverse approché G de l'opérateur K autour de la solution δ^k du problème K(u) = 0, posé dans (1-Q)X.

2.5 REMARQUES ET COMMENTAIRES BIBLIOGRAPHIQUES

On reprend les notions définies dans la section 2.1. D'une manière plus générale, on dit que si $\{\xi^k\}$ est une suite convergente dont la limite est ξ alors la convergence $\xi^k \to \xi$ est *d'ordre* $p \geq 1$ si $\forall k \geq 0$

$$\|\xi^k-\xi\| \leq c_p(d_p)^{p^k} \qquad (2.5.1)$$

où $c_p > 0$ et $0 < d_p < 1$.

Ceci est le cas si la suite $\{\xi^k\}$ vérifie $\forall k \geq 0$

$$\|\xi^{k+1}-\xi\| \leq \beta_p\|\xi^k-\xi\|^p \qquad (2.5.2)$$

où $0 < \beta_1 < 1$ et $\|\xi^0-\xi\| < (\beta_p)^{1/(1-p)}$ si $p > 1$.

Alors la borne (2.5.1) est satisfaite avec

$$c_1 = \beta_1\|\xi^0-\xi\| \quad , \; d_1 = \beta_1$$

et si $p > 1$ $\quad c_p = (\beta_p)^{1/(1-p)}, \quad d_p = \frac{1}{c_p}\|\xi^0-\xi\|$

Evidemment la convergence linéaire est d'ordre 1 et la convergence quadratique, d'ordre 2. Or, si (2.5.2) est vérifié avec $p > 1$ alors $\xi^k \to \xi$ de façon superlinéaire :

$$\|\xi^{k+1}-\xi\| \leq \alpha_k\|\xi^k-\xi\| \qquad \forall k \geq 0$$

avec

$$\alpha_k = \|\xi^k-\xi\|^{p-1} \to 0 \qquad \text{si } k \to \infty.$$

Si l'on se proposait maintenant la tâche de faire un tour d'horizon sur tout ce qui a été écrit à propos de la méthode de Newton on pourrait difficilement le réussir dans les délais que nous imposent les circonstances actuelles. Il ne s'agit donc pas d'entreprendre un tel travail. Mais on peut dire, pourtant, quelques mots à ce sujet.

Cette méthode fut proposée par Isaac Newton autour de 1669. Définie dans un espace fonctionnel elle est connue aussi sous le nom de méthode de Newton-Kantorovich. Les premiers travaux du lauréat soviétique du prix Nobel remontent aux années 1940. Une présentation très complète est faite dans Kantorovich et Akilov 1964 (pp. 695-749) ; on y trouve notamment des applications à la résolution d'équations dans le champ complexe, d'équations

intégrales non linéaires, d'équations différentielles ordinaires, d'équations différentielles quasi-linéaires aux dérivées partielles de deuxième ordre et aussi à l'étude du problème spectral de la théorie de perturbations.

Parmi les autres ouvrages traitant de ce sujet de façon générale on cite, par exemple, Krasnoselskii et al 1972, Ortega et Rheinboldt 1970 et Rall 1969.

Les hypothèses de Kantorovich ont été très étudiées au long des années. Parmi ces travaux on peut citer Dennis 1969, Potra et Pták 1980 ; l'étude de la convergence de méthodes de type Newton a été faite sous des hypothèses les plus diverses tel qu'en témoignent les articles suivants : Axelsson 1982, Bank et Rose 1981, Dembo et al 1982, Dennis 1971, Dennis et Moré 1974, Potra 1982 et Potra et Pták 1983.

La recherche a été poursuivie aussi dans le but d'appliquer la méthode à la résolution de problèmes d'optimisation : Dennis et Moré 1974, Moré et Sorensen 1982, Sorensen 1982 et Tanabe 1981.

Originellement conçue pour le calcul itératif des racines isolées d'une fonction numérique différentiable, des travaux se poursuivent encore, axés sur cette application et particulièrement pour localiser des zéros de polynômes complexes : Sato 1981, Werner 1982.

Une extension de la notion de dérivée a permis d'appliquer la méthode de Newton dans le contexte des mathématiques dites discrètes : Jiang 1982, Robert 1981.

Sur le principe de Correction du Résidu la référence de base demeure, à notre avis, Stetter 1978. A partir de là beaucoup a été écrit sur ce sujet. Une présentation très académique est faite dans Hemker 1982 a, b. Une comparaison avec les méthodes multigrilles pour la résolution du problème de valeurs propres d'un opérateur intégral compact est faite dans Ahués et Chatelin 1983 ; une analyse de ce principe -et aussi de la méthode de Newton- fondée sur l'expansion asymptotique de l'erreur de discrétisation est faite dans Böhmer 1981 ; l'application de ce principe dans le contexte des méthodes multigrilles pour

résoudre des équations intégrales linéaires -et notamment les itérations d'Atkinson et de Brakhage - est présentée dans Hemker et Schippers 1981. Des présentations illustrées par divers exemples figurent aussi dans Ahués et al 1982, 1983c, et Chatelin 1983.

Sur la méthode mixte proposée dans la section 2.4, on remarque qu'elle peut rentrer dans le cadre des méthodes de type Newton décrites dans Axelsson 1982. Effectivement, on peut définir un opérateur $M_k \in \mathrm{Iso}((1-Q)X)$ et un nombre réel $\rho_k \in]0,1[$ tels que $\delta^k_{i(k)}$ est l'une des solutions $u \in (1-Q)X$ de l'inégalité

$$\|M_k(F(\xi^k)-J(\xi^k)u)\| \leq \rho_k \|M_k\, F(\xi^k)\|$$

Pour se rendre compte de ceci il suffit de construire l'inverse approché G comme suit :

Soit $G^0 \in L((1-Q)X)$ une approximation de $(J(\xi^k)\big|_{(1-Q)X})^{-1}$ au sens qu'elle vérifie

$$r_\sigma([1-G^0J(\xi^k)]\big|_{(1-Q)X}) < 1$$

Autrement dit G^0 est un inverse approché de $J(\xi^k)\big|_{(1-Q)X}$ autour de la solution (unique) 0 du problème homogène :

$$J(\xi^k)\big|_{(1-Q)X}u = 0 \qquad\qquad u \in (1-Q)X.$$

Or, pour le problème non homogène $K(u) = 0$ on construit l'inverse approché G à partir de G^0 par translation :

$$G(u) = G^0(u+F(\xi^k)) \qquad\qquad \forall u \in (1-Q)X$$

et évidemment on a dans l'espace $L((1-Q)X)$ l'égalité

$$1\big|_{(1-Q)X} - GK = (1-G^0J(\xi^k))\big|_{(1-Q)X}$$

L'opérateur M_k que l'on recherche est alors

$$M_k = \left[\sum_{i=0}^{i(k)} U^i\right] G^0$$

où $U \in L((1-Q)X)$ est défini par

$$U = [1-G^0J(\xi^k)]\Big|_{(1-Q)X}$$

Alors

$$\delta^k_{i(k)} = M_k\, F(\xi^k)$$

et

$$\|M_k(F(\xi^k)-J(\xi^k)\delta^k_{i(k)})\| = \|\Big(\sum_{i=i(k)+1}^{\infty} U^i\Big)\, G^0J(\xi^k)M_kF(\xi^k)\|$$

$$= \|(1-G^0J(\xi^k))^{i(k)+1}\, M_kF(\xi^k)\| \le \rho_k\|M_kF(\xi^k)\|$$

où

$$\rho_k = \|[(1-G^0J(\xi^k))\Big|_{(1-Q)X}]^{i(k)+1}\|$$

Si $i(k)$ est suffisamment grand alors $\rho_k \in]0,1[$.

Sur la possibilité d'avoir une convergence $\xi^k \to \xi$ d'ordre 2, on peut remarquer que si l'on a

$$\rho_k \in]0,\varepsilon[\qquad \varepsilon < 1$$

alors

$$\|\delta^k_{i(k)}-\delta^k\| \le \alpha\varepsilon^{i(k)}$$

où

$$\alpha = \mathrm{Sup}\{\|\delta^k_0-\delta^k\| : k \in \mathbb{N}\} \le \alpha_0\, \mathrm{Sup}\{\|\xi^k-\xi\| : k \in \mathbb{N}\} < +\infty$$

donc si $i(k) = 2^{k+1}$ on aura $\forall k \ge 0$:

$$\|\xi^k-\xi\| \le 2\,\mathrm{Max}\{\mu\, \ell_J\|\xi^k-\xi\|^2,\ \alpha\varepsilon^{2^k}\}$$

Si le rayon ρ de la boule de convergence de la méthode est tel que $2\rho\mu\ell_J < 1$ alors on aura $\forall k \geq 0$

$$\|\xi^k - \xi\| \leq \beta_2 (d_2)^{2^k}$$

où $\beta_2 > 0$ et $d_2 = \text{Max}\{2\mu\ell_J \|\xi^0 - \xi\|, \varepsilon\} < 1$

Mais certainement, il se peut qu'une convergence quadratique soit atteinte avec un nombre i(k) beaucoup moins élevé. Si l'on considère que la convergence est superlinéaire lorsque $i(k) \to \infty$ quand $k \to \infty$ alors, en pratique, on pourrait se contenter d'un nombre d'itérations i(k) tel que

$$\|\delta^k_{i(k)} - \delta^k_{i(k)-1}\| < \varepsilon \|F(\xi^k)\| \qquad (2.5.3.)$$

où ε est une constante arbitrairement choisie dans $]0.1[$.

En effet, dans l'espace (1-Q)X on a

$$\delta^k_{j+1} = \delta^k_j - G^0 K(\delta^k_j)$$

$$K(\delta^k_j) - K(\delta^k) = (G^0)^{-1}(\delta^k_j - \delta^k_{j+1})$$

$$\delta^k_j - \delta^k = K^{-1}(K(\delta^k_j)) - K^{-1}(0)$$

Notons $d^k_j := K(\delta^k_j)$ le résidu de l'équation $K(u) = 0$ au point δ^k_j. Supposons que G^0, $(G^0)^{-1}$, K et K^{-1} soient uniformément bornés en k. Alors on a les inégalités suivantes, où c est une constante générique

$$\|\delta^k_{j+1} - \delta^k_j\| \leq c\|d^k_j\|$$

$$\|d^k_j\| \leq c\|\delta^k_{j+1} - \delta^k_j\|$$

$$\|\delta^k_j - \delta^k\| \leq c\|d^k_j\|$$

De même, si l'on note $r^k := F(\xi^k)$ le résidu de l'équation $F(x) = 0$ au point ξ^k alors dans (1-Q)X on a

$$\|\xi^{k+1}-\xi^{k}\| \leq c\|r^{k}\|$$

$$\|r^{k}\| \leq c\|\xi^{k}-\xi\|$$

$$\|\xi^{k}-\xi\| \leq c\|r^{k}\|$$

Ceci montre que (2.5.3) entraîne

$$\|\delta^{k}_{i(k)}-\delta^{k}\| \leq c\varepsilon\|\xi^{k}-\xi\|$$

Ainsi, pour ε suffisamment petit on peut, en pratique, atteindre une convergence $\xi^{k} \to \xi$ quadratique.

Or, de la même façon, le processus itératif qui génère la suite $\{\xi^{k}\}$ peut être terminé avec une condition sur le résidu. ξ^{ℓ} est le dernier itéré, où

$$\ell := \text{Min}\{k \in \mathbb{N} : \|r^{k}\| < \varepsilon_{0}\}$$

et ε_{0} est une précision donnée.

Une analyse plus approfondie de ce sujet est faite dans Bank et Rose 1981.

Disons finalement que -ainsi qu'on le verra sur les exemples étudiés au chapitre 3- on obtient la convergence de $\{\xi^{k}\}$ vers ξ même pour le choix $i(k) = 0 \quad \forall k \geq 0$. Cette convergence sera, pourtant, seulement linéaire. Cela revient à substituer à δ^{k} l'approximation $G^{0}F(\xi^{k})$.

CHAPITRE 3

SUR LE PROBLEME DE VALEURS PROPRES D'UN OPERATEUR LINEAIRE COMPACT

Proposition et convergence de méthodes de raffinement itératif

INTRODUCTION

Quatre familles de méthodes de raffinement itératif d'un vecteur propre approché d'un opérateur compact T sont proposées dans le cadre des méthodes mixtes décrites à la fin du chapitre 2.

Il s'agit donc de méthodes de Newton inexactes, le calcul de l'inverse de la dérivée étant fait numériquement à l'aide d'un nombre fini d'itérations d'une technique de correction du résidu.

L'hypothèse fondamentale est que le vecteur à raffiner est un vecteur propre d'une approximation de T fortement stable autour de la valeur propre exacte. Celle-ci est supposée non nulle, simple et isolée.

La convergence de deux des quatre familles de méthodes découle directement des considérations précédentes ; ce sont les familles de type Brakhage : B1 et B2.

La convergence des autres deux familles -de type Atkinson : A1 et A2- est démontrée sous une hypothèse additionnelle qui concerne la résolvante réduite de l'approximation de T et qui est trivialement vérifiée par les approximations collectivement compactes et, a fortiori, par les approximations uniformes.

Des quatre familles que l'on décrit, on choisit en particulier une méthode de chaque famille : celle qui est la plus simple et pour laquelle la convergence est étudiée de manière plus approfondie.

3.1 POSITION DU PROBLEME

Soit X un espace de Banach complexe. On considère le *problème de valeurs propres* de l'opérateur $T \in L(X)$

$$T\phi = \lambda\phi \qquad \phi \neq 0 \qquad \phi \in X \qquad (3.1.1.)$$

sous les hypothèses suivantes :

H1. T est compact.

H2. $\lambda \neq 0$ est une valeur propre simple et isolée par une courbe Γ_λ.

On s'intéresse au raffinement itératif d'une solution approchée (λ_n, ϕ_n) de (3.1.1) que l'on suppose obtenue en résolvant

$$T_n\phi_n = \lambda_n\phi_n \qquad \phi_n \neq 0 \qquad \phi_n \in X \qquad (3.1.2.)$$

où $\{T_n\}$ est une suite dans $L(X)$ telle que

H3. $T_n - z \xrightarrow{f\,s} T - z$ sur Γ_λ.

D'après la propriété 1.3.5, on peut supposer que λ_n est non nulle, simple et isolée par Γ_λ en tant qu'élément de $\sigma(T_n)$.

On va utiliser la notation suivante (cf. section 1.3) :

$\hat{\phi}$: vecteur propre de T associé à λ et normalisé par $\|\hat{\phi}\| = 1$.

$P = P(T,\lambda)$, $M = PX = M(T,\lambda)$, $S = S(T,\lambda)$.

$\hat{\phi}^*$: vecteur propre de T^* associé à $\bar{\lambda}$ et normalisé par $\langle\hat{\phi},\hat{\phi}^*\rangle = 1$.

λ_n : valeur propre simple, non nulle, de T_n isolée par Γ_λ.

ϕ_n : vecteur propre de T_n associé à λ_n et normalisé par $\|\phi_n\| = 1$.

P_n : projection spectrale de T_n associée à λ_n.

$M_n = P_n X$: sous-espace invariant engendré par ϕ_n.

$S_n(z) = (T_n - z)^{-1}(1-P_n) = R_n(z)(1-P_n)$ $\quad S_n = \lim_{z\to\lambda_n} S_n(z)$.

ϕ_n^* : vecteur propre de T_n^* associé à $\bar{\lambda}_n$ normalisé par $\langle \phi_n, \phi_n^* \rangle = 1$.

$\phi^{(n)}$: vecteur propre de T associé à λ normalisé par $\langle \phi^{(n)}, \phi_n^* \rangle = 1$ (cf. propriété 1.3.1).

$\hat{\phi}^{(n)} := \dfrac{1}{\|\phi^{(n)}\|}\phi^{(n)}$.

$\hat{\phi}^{(n)*}$: vecteur propre de T^* associé à $\bar{\lambda}$ normalisé par $\langle \hat{\phi}^{(n)}, \hat{\phi}^{(n)*} \rangle = 1$.

Le fait que λ et λ_n soient simples implique

$$\bar{B} \cap M = \{e^{2\pi i t}\,\hat{\phi} : 0 \le t \le 1\}.$$

$$\hat{\phi}^{(n)} = e^{2\pi i t_n}\,\hat{\phi} \Rightarrow \hat{\phi}^{(n)*} = e^{2\pi i t_n}\,\hat{\phi}^*.$$

et permet de représenter les projections P et P_n ainsi :

$$Px = \langle x, \hat{\phi}^* \rangle \hat{\phi} = \langle x, \hat{\phi}^{(n)*} \rangle \hat{\phi}^{(n)} \quad \forall x \in X.$$

$$P_n x = \langle x, \phi_n^* \rangle \phi_n \quad \forall x \in X.$$

On note n_0 un entier tel que les constantes et les bornes suivantes sont bien définies (cf. section 1.3)

$$\eta := \mathrm{Sup}\{\|P_n\| : n \ge n_0\}$$

$$\nu := \mathrm{Sup}\{\|T_n\| : n \ge n_0\}$$

$$\gamma := \mathrm{Sup}\{\|S_n\| : n \ge n_0\}$$

$$\ell := \frac{1}{2\pi}\int_{\Gamma_\lambda} |dz|$$

$$c_\lambda := 2\ell\ \mathrm{Sup}\{\|R_n(z)\| : z \in \Gamma_\lambda,\ n \ge n_0\}\mathrm{Sup}\{\|R(z)\| : z \in \Gamma_\lambda\}$$

$$\mathrm{Sup}\{\|(P_n|_M)^{-1}\| : n \ge n_0\} \le 2$$

$$\mathrm{Sup}\{\|\phi^{(n)}\| : n \ge n_0\} \le 2$$

Si $c := \mathrm{Max}\{2n, c_\lambda\}$ alors on suppose que pour $n \geq n_o$ on a (Cf. Propriétés 1.3.2 et 1.3.5) :

$$|\lambda-\lambda_n| \leq c\|(T-T_n)P\|$$

$$\|\phi^{(n)}-\phi_n\| \leq c\|(T-T_n)P\|$$

$$\|\hat{\phi}^{(n)}-\phi^{(n)}\| \leq c\|(T-T_n)P\|$$

$$\|\hat{\phi}^{(n)}-\phi_n\| \leq 2c\|(T-T_n)P\|$$

On s'intéresse à l'approximation numérique de $\phi^{(n)}$ pour n assez grand mais fixe via une méthode de raffinement itératif dont le point de départ est l'approximation ϕ_n. Pour cela, on considère l'opérateur non linéaire $F^{(n)} : X \to X$ défini pour $n \geq n_o$ par

$$F^{(n)}(x) = Tx-\langle Tx,\phi_n^*\rangle x \qquad \forall x \in X. \qquad (3.1.3.)$$

dont la dérivée en x est donnée par

$$J^{(n)}(x)u = Tu-\langle Tx,\phi_n^*\rangle u-\langle Tu,\phi_n^*\rangle x$$

et dont $\phi^{(n)}$ est une racine isolée (cf. exemple 2.2.1 et propriété 1.4.1.)

Lemme 3.1.1

Si $x \in P_n^{-1}\{\phi_n\}$ *alors*

i) $(1-P_n)X$ *est invariant par* $J^{(n)}(x)$

ii) $F^{(n)}(x) \in (1-P_n)X$.

Preuve : Si $P_n x = \phi_n$ alors

i) Etant donné $u \in (1-P_n)X$,

$$P_n J^{(n)}(x)u = P_n Tu-\langle Tu,\phi_n^*\rangle P_n x =$$

$$\langle Tu,\phi_n^*\rangle\phi_n-\langle Tu,\phi_n^*\rangle\phi_n = 0$$

ii) $P_n F^{(n)}(x) = P_n Tx - \langle Tx, \phi_n^* \rangle P_n x =$

$\langle Tx, \phi_n^* \rangle \phi_n - \langle Tx, \phi_n^* \rangle \phi_n = 0$

□

Soit $L = T - \lambda(1+P) \in L(X)$. Cet opérateur est évidemment indépendant de n, pourtant il admet, pour $n \geq n_o$, la représentation suivante :

$$Lu = Tu - \lambda u - \langle Tu, \hat{\phi}^{(n)*} \rangle \hat{\phi}^{(n)} \qquad \forall u \in X.$$

Lemme 3.1.2

$$L \in \mathrm{Iso}(X)$$

Preuve : $T-\lambda P$ étant compact, il suffit de vérifier que L est injectif, d'après la propriété 1.1.23. Mais cela découle de la propriété 1.1.22. □

Lemme 3.1.3

Il existe des constantes θ_1 *et* θ_2 *telles que pour* $n \geq n_o$:

$$\|L - J^{(n)}(x)\| \leq \theta_1 \|x - \hat{\phi}^{(n)}\| + \theta_2 \|(T-T_n)P\| + \|(P-P_n)T\|$$

Preuve :

$$(L - J^{(n)}(x))u = (\langle Tx, \phi_n^* \rangle - \lambda)u +$$

$$+ \langle Tu, \phi_n^* \rangle (x - \hat{\phi}^{(n)}) + \langle Tu, \phi_n^* - \hat{\phi}^{(n)*} \rangle \hat{\phi}^{(n)} =$$

$$= \langle T(x - \hat{\phi}^{(n)} + \hat{\phi}^{(n)} - \phi^{(n)}), \phi_n^* \rangle u + \langle Tu, \phi_n^* \rangle (x - \hat{\phi}^{(n)}) +$$

$$+ \langle Tu, \phi_n^* \rangle (\hat{\phi}^{(n)} - \phi_n) + (P_n - P)Tu.$$

Donc, si l'on définit

$$\theta_1 := 2n\|T\|, \quad \theta_2 := 3cn\|T\|$$

on a l'inégalité que l'on recherchait. □

Les constantes θ_1 et θ_2 seront utilisées à plusieurs reprises au cours des sections suivantes.

Soit $\mu := 1+\|L^{-1}\|$. D'après la propriété 1.1.12 il existe un réel $\delta \in]0, \frac{1}{\mu}[$ tel que $\forall A \in L(X)$

$$\|L-A\| < \delta \Rightarrow \| L^{-1}-A^{-1}\| < 1$$

l'existence de A^{-1} étant assurée du fait que $\delta < \frac{1}{\|L^{-1}\|}$ (Cf. propriété 1.1.11).

Dorénavant on fixe δ et la constante ω du lemme suivant :

Lemme 3.1.4

Il existe une constante ω telle que

$$\mathrm{Sup}\{\|J^{(n)}(x)\| : x \in B(\hat{\phi}^{(n)},\delta)\} \leq \omega$$

Preuve : Il suffit de prendre :

$\omega := \|T\|+\theta_1(1+\delta)$. □

Nous sommes maintenant bien placés pour définir une méthode de raffinement itératif de ϕ_n, convergeant vers $\phi^{(n)}$, du type de celle décrite dans la section 2.4.

3.2 UNE PREMIERE FAMILLE DE METHODES

On considère la famille de méthodes suivante, caractérisée par le paramètre i(k) :

$$\left.\begin{array}{l} \phi^0 := \phi_n \\ \phi^{k+1} := \phi^k - \delta^k_{i(k)} \qquad k \geq 0 \\ \text{où} \\ \delta^k_{i(k)} := \sum_{\ell=0}^{i(k)} (S_n H_n(\phi^k))^{\ell} S_n F^{(n)}(\phi^k) \\ \text{et} \\ H_n(x) := (T_n - \lambda_n - J^{(n)}(x))\big|_{(1-P_n)X} \end{array}\right\} \quad (A1,i(k))$$

On remarque que $\delta^k_{i(k)}$ est le dernier itéré du processus suivant

$$\delta^k_0 := S_n F^{(n)}(\phi^k)$$

$$\delta^k_{j+1} := \delta^k_0 + (1 - S_n J^{(n)}(\phi^k))\delta^k_j \qquad 0 \le j \le i(k)-1$$

qui peut être interprété comme une méthode de correction du résidu pour la résolution numérique de l'équation

$$K(u) := J^{(n)}(\phi^k)u - F^{(n)}(\phi^k) = 0$$

où l'on a pris $G(u) := S_n(u+F^{(n)}(\phi^k))$ comme un inverse approché de K, autour de la solution exacte $\delta^k := J^{(n)}(\phi^k)^{-1}F^{(n)}(\phi^k)$ dont l'existence découle du lemme 3.1.3 si n est assez grand et ϕ^k suffisamment proche de $\hat{\phi}^{(n)}$. Mais ceci concerne la convergence de la méthode.

Lemme 3.2.1

La suite $\{\phi^k\}$ *définie par la méthode* (A1,i(k)) *vérifie :*

$$P_n\phi^k = \phi_n \qquad \forall k \ge 0.$$

Preuve : La suite $\{\delta^k_{i(k)}\}$ vérifie $P_n\delta^k_{i(k)} = 0$ et $P_n\phi^0 = P_n\phi_n = \phi_n$.

□

La convergence de la méthode (A1,i(k)) sera démontrée sous les hypothèses H1, H2, H3 et

H4. Il existe un entier q tel que $\lim_{n\to\infty} \|(S_n(T-T_n))^q\| = 0$.

Lemme 3.2.2

La suivante est une condition suffisante pour H3 *et* H4 *:*

$T_n \xrightarrow{p} T$ *et il existe un entier* p *tel que*

$$\lim_{n\to\infty} \mathrm{Sup}\{\|((T-T_n)R(z))^p\| : z \in \Gamma_\lambda\} = 0$$

Cette condition est vérifiée si $T_n \xrightarrow{cc} T$ *ou* $T_n \xrightarrow{\|\|} T$.

Preuve :

i) Elle est suffisante pour H3 :
On définit la fonction

$$t \in [0,1] \to T_n(t) := T-t(T-T_n) \in L(X).$$

La condition du lemme 3.2.2 implique

$$\mathrm{Sup}\{r_\sigma(t(T-T_n)R(z)) : z \in \Gamma_\lambda\} < 1$$

pour tout n suffisamment grand. Donc $1-t(T-T_n)R(z)$ admet un inverse dans $L(X)$.
L'identité

$$T_n(t)-z = (1-t(T-T_n)R(z))(T-z)$$

montre que pour $z \in \Gamma_\lambda \subset \rho(T)$ et n assez grand $z \in \rho(T_n(t))$ et

$$(T_n(t)-z)^{-1} = R(z) \sum_{j=0}^{\infty} t^j ((T-T_n)R(z))^j$$

où la série converge dans $L(X)$.
L'application

$$t \in [0,1] \to P_n(t) := -\frac{1}{2\pi i} \int_{\Gamma_\lambda} (T_n(t)-z)^{-1} dz \in L(X)$$

est continue et $P_n(t)$ est une projection donc dim $P_n(t)X$ est constante pour $t \in [0,1]$.
Ceci montre que dim $P_n X$ = dim PX car

$$P_n = P_n(1) \text{ et } P = P_n(0).$$

D'autre part on établit

$$R_n(z) = R(z)\, Y(n,z,p) \sum_{j=0}^{\infty} [(T-T_n)R(z)]^{pj}$$

où

$$Y(n,z,p) = \sum_{j=0}^{p-1} ((T-T_n)R(z))^j$$

Etant donné $\epsilon \in]0,1[$ on a pour tout n assez grand

$$\text{Sup}\{\|((T-T_n)R_n(z))^p\| : z \in \Gamma_\lambda\} < \epsilon$$

donc $R_n(z)$ est uniformément borné par rapport à n et à $z \in \Gamma_\lambda$.
On en déduit les convergences

$$R_n(z) \xrightarrow{p} R(z) \qquad \forall z \in \Gamma_\lambda$$

$$P_n \xrightarrow{p} P$$

Ainsi, H3 est vérifié.

ii) Elle est suffisante pour H4 :
L'identité

$$(1-(T-T_n)R(z))^{-1} = (T-z)R_n(z)$$

montre qu'il existe une constante c_o telle que

$$\text{Sup}\{\|(1-(T-T_n)R(z))^{-1}\| : z \in \Gamma_\lambda\} \leq c_o$$

Comme $(T-T_n)R(z)$ et $(1-(T-T_n)R(z))^{-1}$ commutent et vérifient

$$(T-T_n)R_n(z) = (1-(T-T_n)R(z))^{-1}\,(T-T_n)R(z)$$

on a

$$\|((T-T_n)R_n(z))^p\| \leq (c_o)^p \|((T-T_n)R(z))^p\|$$

qui tend vers zéro si $n \to \infty$.

D'autre part, il existe c(p) -dépendant de p mais indépendant de n- tel que

$$\|((T-T_n)S_n(z))^p\| \le \|((T-T_n)R_n(z))^p\| + c(p)\|(T-T_n)P_n\|$$

D'après le Principe du Maximum on a

$$\|((T-T_n)S_n)^p\| \le \text{Max}\{\|((T-T_n)S_n(z))^p\| : z \in \Gamma_\lambda\}$$

Donc H4 est vérifié avec q = p+1 .

Ce qui reste est trivial.

□

Lemme 3.2.3

Il existe $r_o > 0$ *tel que pour* n *assez grand*

$$\text{Sup}\{r_\sigma(S_nH_n(x)) : x \in B(\hat{\phi}^{(n)}, r_o) \cap P_n^{-1}\{\phi_n\}\} < 1$$

Preuve : Premièrement on remarque que $S_nH_n(x) \in L((1-P_n)X)$ et que $\forall u \in (1-P_n)X$, $\forall x \in X$ tel que $P_nx = \phi_n$ on a :

$$H_n(x)u = (T_n-T)u-\lambda_n u+\langle Tx,\phi_n^*\rangle u-\langle(T_n-T)u,\phi_n^*\rangle x$$

$$S_nH_n(x)u = S_n(T_n-T)u-\lambda_nS_nu+\langle T(x-\hat{\phi}^{(n)}+\hat{\phi}^{(n)}-\phi^{(n)}+\phi^{(n)}),\phi_n^*\rangle S_nu$$

$$+\langle(T_n-T)u,\phi_n^*\rangle S_n(x-\hat{\phi}^{(n)}+\hat{\phi}^{(n)}-\phi_n)$$

$$= (D_n+Q_n(x))u$$

Avec :

$$D_nu = S_n(T_n-T)u$$

$$Q_n(x)u = (\lambda-\lambda_n)S_nu+\langle T(x-\hat{\phi}^{(n)}),\phi_n^*\rangle S_nu +$$

$$+\lambda\langle\hat{\phi}^{(n)}-\phi^{(n)},\phi_n^*\rangle S_nu +\langle(T_n-T)u,\phi_n^*\rangle S_n(x-\hat{\phi}^{(n)}) +$$

$$+\langle(T_n-T)u,\phi_n^*\rangle S_n(\hat{\phi}^{(n)}-\phi_n).$$

On définit les constantes

$$\gamma_1 := \gamma(\nu+\|T\|)$$

$$\gamma_2 := \gamma c(1+|\lambda|n+2n(\nu+\|T\|))$$

$$\gamma_3 := \gamma n(\nu+2\|T\|)$$

et l'on a :

$$\|D_n\| \leq \gamma_1$$

$$\|Q_n(x)\| \leq \gamma_2\|(T-T_n)P\|+\gamma_3\|x-\hat{\phi}^{(n)}\|$$

D'après les hypothèses H3 et H4 il existe $n_1 \geq n_o$ et q tels que

$$\|D_n^q\| < \frac{1}{4} \qquad \forall n \geq n_1$$

$$\|(T-T_n)P\| < \frac{\varepsilon_q}{2\gamma_2} \qquad \forall n \geq n_1$$

où $\varepsilon_q := \text{Min}\{\frac{1}{4}, (\sum_{j=1}^{q} \binom{q}{j} \frac{\gamma_1^{q-j}}{4^{j-2}})^{-1}\}$

donc

$$\|Q_n(x)\| < \frac{\varepsilon_q}{2} + \gamma_3\|x-\hat{\phi}^{(n)}\|$$

Si l'on prend $r_o := \frac{\varepsilon_q}{2\gamma_3}$ alors on aura

$$\|(D_n+Q_n(x))^q\| < \frac{1}{2} \qquad \forall n \geq n_1 \quad \forall x \in B(\hat{\phi}^{(n)}, r_o)$$

d'où le résultat.

□

Théorème 3.2.4

Convergence de (A1,i(k)) *sous les hypothèses* H1, H2, H3 *et* H4.
Si n *est suffisamment grand et si pour tout* $k \in \mathbb{N}$, i(k) *est suffisamment grand alors* $\phi^k \to \phi^{(n)}$ *lorsque* $k \to \infty$.
La convergence est linéaire si i(k) *est constante ; superlinéaire si* $i(k) \to \infty$ *lorsque* $k \to \infty$ *et quadratique si la suite* $\{\nu_k\}$ *définie par*

$$\nu_k = \frac{\|(S_n H_n(\phi^k))^{i(k)+1}\|}{\|\phi^k - \phi^{(n)}\|}$$

est bornée par rapport à k.

Preuve : Soit $r_1 := \text{Min}\{\delta, r_o, \frac{\delta}{2\theta_1}, \frac{1}{8\theta_1\mu}\}$ et $r = \frac{r_1}{2}$.

Il existe $n_1 \geq n_o$ tel que $\forall n \geq n_1$ on a

$$\text{Max}\{\theta_2\|(T-T_n)P\| + \|(P-P_n)T\|,\ 2c\|(T-T_n)P\|\} < \frac{r_1}{4}$$

Donc, $\forall n \geq n_1$ et $\forall x \in B(\hat{\phi}^{(n)}, r_1)$:

$$\|L - J^{(n)}(x)\| < \frac{\delta}{2} + \frac{\delta}{4} < \delta$$

d'après le lemme 3.1.3. Ceci entraîne $J^{(n)}(x) \in \text{Iso}(X)$ et

$$\text{Sup}\{\|(J^{(n)}(x))^{-1}\| : x \in B(\hat{\phi}^{(n)}, r_1)\} \leq 1 + \|L^{-1}\| = \mu$$

à cause du choix de δ .

D'autre part, θ_1 est une constante de Lipschitz de $J^{(n)}$ sur $B(\hat{\phi}^{(n)}, r_1)$ et l'on a : $\forall x_1, x_2 \in B(\hat{\phi}^{(n)}, r_1)$:

$$\|J^{(n)}(x_1) - J^{(n)}(x_2)\| \leq \theta_1\|x_1 - x_2\| < 2\theta_1 r_1 < \frac{1}{4\mu}$$

Mais

$$\|\hat{\phi}^{(n)} - \phi^{(n)}\| \leq c\|(T-T_n)P\| \leq \frac{r_1}{8}$$

et $$\|\hat{\phi}^{(n)} - \phi_n\| \leq 2c\|(T-T_n)P\| \leq \frac{r_1}{4}$$

donc

$$\phi_n \in B(\phi^{(n)}, r) \subseteq B(\hat{\phi}^{(n)}, r_1)$$

Le lemme 3.2.3. montre l'existence d'un entier i_o tel que $\forall j \geq i_o$

$$\mathrm{Sup}\{\|(S_nH_n(x))^{j+1}\| : x \in B(\hat{\phi}^{(n)},r_1) \cap P_n^{-1}\{\phi_n\}\} < \frac{1}{2\mu\omega}$$

Or, si $\phi^k \in B(\phi^{(n)},r)$ et $\delta^k = (J^{(n)}(\phi^k))^{-1}F^{(n)}(\phi^k)$ alors

$$\delta^k - \delta^k_j = \sum_{\ell=j+1}^{\infty} (S_nH_n(\phi^k))^{\ell} S_nF^{(n)}(\phi^k)$$

$$= (S_nH_n(\phi^k))^{j+1}(J^{(n)}(\phi^k))^{-1} \int_0^1 J^{(n)}(\phi^k(t))(\phi^k-\phi^{(n)})dt$$

où

$$\phi^k(t) = \phi^{(n)}+t(\phi^k-\phi^{(n)}) \in B(\hat{\phi}^{(n)},r_1) \qquad \forall t \in [0,1].$$

Donc :

$$\|\delta^k-\delta^k_j\| \leq \|(S_nH_n(\phi^k))^{j+1}\| \, \|(J^{(n)}(\phi^k))^{-1}\| \, \|\phi^k - \phi^{(n)}\|\omega$$

$$< \frac{1}{2}\|\phi^k-\phi^{(n)}\|.$$

Finalement :

$$\phi^{k+1}-\phi^{(n)} = \phi^k-\delta^k_{i(k)}-\phi^{(n)} = \phi^k-\delta^k-\phi^{(n)}+\delta^k-\delta^k_{i(k)}$$

$$= (J^{(n)}(\phi^k))^{-1}\int_0^1 [J^{(n)}(\phi^k(t))-J^{(n)}(\phi^k)](\phi^k-\phi^{(n)})dt+\delta^k-\delta^k_{i(k)}$$

et d'après les majorations précédentes,

$$\|\phi^{k+1}-\phi^{(n)}\| \leq (\mu\,\mathrm{Sup}\{\|J^{(n)}(\phi^k(t))-J^{(n)}(\phi^k)\| : t \in [0,1]\}+$$

$$+\frac{1}{2})\|\phi^k-\phi^{(n)}\| \leq \frac{3}{4}\|\phi^k-\phi^{(n)}\|$$

ce qui montre, d'une part, que $\phi^{k+1} \in B(\phi^{(n)},r)$ lorsque $\phi^k \in B(\phi^{(n)},r)$ -et l'on sait déjà que $\phi^0 = \phi_n \in B(\phi^{(n)},r)$- et d'autre part, que $\phi^k \to \phi^{(n)}$ lorsque $k\to\infty$ si $i(k)$ vérifie $i(k) \geq i_o$ $\forall k > 0$ et si n est assez grand.

Mais une majoration plus fine est valable du fait que $J^{(n)}$ est contractant sur $B(\phi^{(n)},r)$:

$$\|\phi^{k+1}-\phi^{(n)}\| \leq \theta_1\mu\|\phi^k-\phi^{(n)}\|^2+\alpha_{i(k)}\|\phi^k-\phi^{(n)}\|$$

où $$\alpha_{i(k)} := \mu\,\omega\,\|(S_nH_n(\phi^k))^{i(k)+1}\|.$$

Ceci démontre les assertions finales du théorème. □

Cette première famille de méthodes répond aux caractéristiques du schéma proposé dans la section 2.4, le rôle de la projection Q étant joué par P_n. Nous remarquons que la preuve du théorème précédent montre que l'opérateur

$$G : (1-P_n)X \to (1-P_n)X$$

défini par $G(u) = S_n(u+F^{(n)}(\phi^k))$ $\quad \forall u \in (1-P_n)X,$

est un inverse approché de l'opérateur

$$K : (1-P_n)X \to (1-P_n)X$$

défini par $K(u) = J^{(n)}(\phi^k)u - F^{(n)}(\phi^k)$, $\forall u \in (1-P_n)X$, autour de la solution (unique) $\delta^k = (J^{(n)}(\phi^k))^{-1}F^{(n)}(\phi^k)$ du problème $K(u) = 0$. En fait, sur l'espace $(1-P_n)X$ on a

$$1\big|_{(1-P_n)X} - GK = S_nH_n(\phi^k)$$

donc les hypothèses de la propriété 2.3.4 sont satisfaites.

Il est utile de signaler aussi que la façon de traiter numériquement l'équation $K(u) = 0$ est du même type que la méthode d'Atkinson (Cf. Exemple 2.3.1). En effet, si dans l'expression de $J^{(n)}(x)$ on substitue T_n à T on obtient une dérivée approchée, que l'on note $J_n(x)$, et qui vérifie

$$J_n(x)\big|_{(1-P_n)X} = (T_n-\lambda_n)\big|_{(1-P_n)X} \qquad \text{si} \quad P_nx = \phi_n.$$

La solution de l'équation approchée :

$$J_n(\phi^k)\big|_{(1-P_n)X}\, u_n = F^{(n)}(\phi^k), \qquad u_n \in (1-P_n)X$$

est $u_n = S_n F^{(n)}(\phi^k)$. Le choix de l'inverse approché G découle directement de ces considérations.

En ce qui concerne l'approximation numérique de la valeur propre λ nous sommes amenés à définir la suite $\{\lambda^k\}$ par

$$\left.\begin{aligned} \lambda^0 &:= \lambda_n \\ \lambda^{k+1} &:= \langle T\phi^k, \phi_n^* \rangle \qquad k \geq 0 \end{aligned}\right\} \qquad (3.2.1.)$$

Théorème 3.2.5

Si $\phi^k \to \phi^{(n)}$ *lorsque* $k \to \infty$ *alors* $\lambda^k \to \lambda$ *lorsque* $k \to \infty$.

Preuve : Il suffit de considérer la borne
$|\lambda-\lambda^{k+1}| \leq \eta \|T\| \, \|\phi^k-\phi^{(n)}\|$.

□

Certainement, la méthode la plus simple de la famille (A1,i(k)) est celle qui consiste à prendre $i(k) = 0 \quad \forall k \geq 0$.
Ceci revient à l'itération suivante :

$$\left.\begin{aligned} \phi^0 &:= \phi_n \\ \phi^{k+1} &:= \phi^k - S_n(T\phi^k - \lambda^{k+1}\phi^k) \qquad k \geq 0 \end{aligned}\right\} \qquad (A1,0)$$

Théorème 3.2.6

Si $T_n \xrightarrow{\|\ \|} T$ *alors la suite* $\{\phi^k\}$ *définie par* (A1,0) *converge vers* $\phi^{(n)}$ *pour* n *suffisamment grand. De plus il existe des constantes* β *et* β_o *telles que* $\forall k \geq 0$:

$$\mathrm{Max}\{|\lambda-\lambda^k|, \|\phi^k-\phi^{(n)}\|\} \leq \beta(\mu_n)^{k+1} \quad \text{où} \quad \mu_n = \beta_o \|T-T_n\|$$

Preuve : D'après la preuve du théorème 3.2.4 on aura la borne suivante

$$\|\phi^{k+1}-\phi^{(n)}\| \leq \theta_1\mu\|\phi^k-\phi^{(n)}\|^2 + \alpha_o\|\phi^k-\phi^{(n)}\|$$

où

$$\alpha_o = \mu\omega\|S_nH_n(\phi^k)\| \leq \mu\omega\text{Sup}\{\|S_nH_n(x)\| : x \in B(\hat{\phi}^{(n)},r_1), P_nx = \phi_n\}$$

Or, la preuve du lemme 3.2.3 nous montre l'existence de constantes β_1 et β_2 telles que $\forall x \in B(\hat{\phi}^{(n)},r_1) \cap P_n^{-1}\{\phi_n\}$:

$$\|S_nH_n(x)\| \leq \beta_1\|x-\hat{\phi}^{(n)}\|+\beta_2\|T-T_n\|$$

ce qui démontre le théorème par induction. Dans le cas k = 0 la borne découle des propriétés 1.3.2 et 1.3.5 et pour la suite $\{\lambda^k\}$ on vérifie aisément l'identité

$$\lambda - \lambda^{k+1} = \langle(T_n-T)(\phi^k-\phi^{(n)}),\phi_n^*\rangle.$$

□

On peut montrer aussi la convergence de (A1,0) sous l'hypothèse plus faible $T_n \xrightarrow{cc} T$ mais elle n'est plus une convergence dominée de façon monotone :

Théorème 3.2.7

Si $T_n \xrightarrow{cc} T$ *alors la suite* $\{\phi^k\}$ *définie par* (A1,0) *converge vers* $\phi^{(n)}$ *pour* n *suffisamment grand. De plus, il existe des constantes* β *et* β_o *telles que* $\forall k \geq 0$:

$$\text{Max}\{|\lambda^{2k}-\lambda|, |\lambda^{2k+1}-\lambda|, \|\phi^{2k}-\phi^{(n)}\|, \|\phi^{2k+1}-\phi^{(n)}\|\} \leq \beta\eta_n(\eta_n+\epsilon_n)^k$$

où

$\eta_n=\beta_o\|(T-T_n)P\|$ *et* $\epsilon_n = \beta_o\|(T-T_n)S_n(T-T_n)\|$.

Preuve : Le résultat est une conséquence des identités suivantes :

$$\lambda^{2k+1}-\lambda = \langle T(\phi^{2k}-\phi(n)),\phi_n^*\rangle$$

$$\phi^{2k+1}-\phi(n) = (\lambda^{2k+1}-\lambda)S_n\phi^{2k}+(\lambda-\lambda_n)S_n(\phi^{2k}-\phi(n)) +$$

$$+ S_n(T_n-T)(\phi^{2k}-\phi(n))$$

$$\lambda-\lambda^{2k+2} = \langle (T_n-T)(\phi^{2k+1}-\phi(n)),\phi_n^*\rangle$$

$$(T_n-T)(\phi^{2k+1}-\phi(n)) = (\lambda^{2k+1}-\lambda)(T_n-T)S_n(\phi^{2k}-\phi(n)+\phi(n)-\phi_n) +$$

$$+ (\lambda-\lambda_n)(T_n-T)S_n(\phi^{2k}-\phi(n)) +$$

$$+ (T_n-T)S_n(T_n-T)(\phi^{2k}-\phi(n)).$$

$$\phi^{2k+2}-\phi(n) = (\lambda^{2k+2}-\lambda)S_n\phi^{2k+1}+(\lambda-\lambda_n)(\lambda^{2k+1}-\lambda)S_n^2(\phi^{2k}-\phi(n))+$$

$$+ (\lambda-\lambda_n)(\lambda^{2k+1}-\lambda)S_n^2(\phi(n)-\phi_n) +$$

$$+ (\lambda-\lambda_n)^2S_n^2(\phi^{2k}-\phi(n)) +$$

$$+ (\lambda-\lambda_n)S_n^2(T_n-T)(\phi^{2k}-\phi(n)) +$$

$$+ (\lambda^{2k+1}-\lambda)S_n(T_n-T)S_n(\phi^{2k}-\phi(n)) +$$

$$+ (\lambda^{2k+1}-\lambda)S_n(T_n-T)S_n(\phi(n)-\phi_n) +$$

$$+ (\lambda-\lambda_n)S_n(T_n-T)S_n(\phi^{2k}-\phi(n)) +$$

$$+ S_n(T_n-T)S_n(T_n-T)(\phi^{2k}-\phi(n)).$$

Celles-ci permettent d'achever la démonstration par induction. On déduit le cas k = 0 des propriétés 1.3.2 et 1.3.5. Remarquons que $\eta_n \to 0$ et $\varepsilon_n \to 0$ si $n \to \infty$ d'après les propriétés 1.2.5 et 1.2.6.

□

Les théorèmes 3.2.6 et 3.2.7 ont un intérêt du fait qu'en pratique les approximations T_n que l'on considère satisfont à ses hypothèses (Cf. Exemples 1.2.1. et 1.2.2.). Cependant, sous les hypothèses H1, H2, H3, H4, on a, selon la preuve du théorème 3.2.4, l'existence de trois constantes ω_1, ω_2 et ω_3 telles que pour $x \in B(\hat{\phi}^{(n)}, r_1) \cap P_n^{-1}\{\phi_n\}$:

$$\|(S_n H_n(x))^q\| \leq \omega_1 \|x - \hat{\phi}^{(n)}\| + \omega_2 \|(S_n(T_n - T))^q\| +$$

$$\omega_3 \|(T - T_n)P\|$$

où q est un entier tel que $\|(S_n(T_n - T))^q\| < 1$.

Nous pouvons ainsi établir pour la méthode (A1,0) la borne

$$\text{Max}\{|\lambda - \lambda^{qk+j}|, \|\phi^{qk+j} - \phi^{(n)}\| : j = 0, \ldots, q-1\} \leq \beta\ \eta_n(\eta_n + \varepsilon_{q,n})^k$$

où $\varepsilon_{q,n} = \beta_o \|(S_n(T_n - T))^q\|$. Pour n assez grand nous avons $\eta_n + \varepsilon_{q,n} < 1$ d'où la convergence. Nous ne ferons pas les détails, les cas les plus intéressants étant donnés par les théorèmes 3.2.6 et 3.2.7.

Nous présentons ensuite une autre famille de méthodes. Il est utile de revenir sur l'exemple 2.3.2.

3.3 UNE DEUXIEME FAMILLE DE METHODES

On considère maintenant une autre famille de méthodes qui est aussi caractérisée par le paramètre i(k) :

$$\left.\begin{aligned} &\phi^0 := \phi_n \\ &\phi^{k+1} := \phi^k - \delta^k_{i(k)} \qquad k \geq 0 \\ &\text{où} \qquad \delta^k_{i(k)} := \sum_{\ell=0}^{i(k)} (1-S^B_n J^{(n)}(\phi^k))^\ell S^B_n F^{(n)}(\phi^k) \\ &\text{et} \qquad S^B_n := \frac{1}{\lambda_n}(S_n T - 1 + P_n) \end{aligned}\right\} \quad (B1,i(k))$$

On remarque que $\delta^k_{i(k)}$ est le dernier itéré du procédé suivant :

$$\delta^k_0 := S^B_n F^{(n)}(\phi^k)$$

$$\delta^k_{j+1} := \delta^k_0 + (1-S^B_n J^{(n)}(\phi^k))\delta^k_j \qquad 0 \leq j \leq i(k)-1$$

qui répond aux caractéristiques d'une méthode de correction du résidu appliquée sur le problème $K(u) = 0$ où l'on a défini comme inverse approché autour de $\delta^k = (J^{(n)}(\phi^k))^{-1} F^{(n)}(\phi^k)$ l'opérateur G^B défini par $G^B(u) = S^B_n(u+F^{(n)}(\phi^k))$.

On démontre la convergence sous les hypothèses H1, H2 et H3.

Pour $n \geq n_0$ et $x \in P_n^{-1}\{\phi_n\}$ on définit l'opérateur $U_n(x) : (1-P_n)X \to (1-P_n)X$ par

$$U_n(x) := (1-S^B_n J^{(n)}(x))\big|_{(1-P_n)X}$$

Lemme 3.3.1

La suite $\{\phi^k\}$ *définie par la méthode* (B1,i(k)) *vérifie*

$$P_n\phi^k = \phi_n \qquad \forall k \geq 0$$

Preuve : Comme dans le Lemme 3.2.1 ceci est une conséquence du fait que la suite $\{\delta^k_{i(k)}\}$ vérifie $P_n\delta^k_{i(k)} = 0 \quad \forall k \geq 0$ et $P_n\phi^0 = \phi_n$. □

Lemme 3.3.2

$$\|(T-T_n)P\| \leq \frac{\|P\|}{|\lambda|}\|(T-T_n)T\|$$

Preuve : $TP = PT = \lambda P$ car λ est simple. Comme $\lambda \neq 0$ on écrit $P = \frac{1}{\lambda}TP$. □

Lemme 3.3.3

Il existe des constantes σ_1 *et* σ_2 *telles que pour tout* n *suffisamment grand et* $x \in P_n^{-1}\{\phi_n\}$:

$$\|U_n(x)\| \leq \sigma_1\|(T-T_n)T\| + \sigma_2\|x-\hat{\phi}^{(n)}\|.$$

Preuve : Soit $n \geq n_0$, $x \in P_n^{-1}\{\phi_n\}$ et $u \in (1-P_n)X$.

$$U_n(x)u = S_n[T_nu - \lambda_n u - \frac{1}{\lambda_n}T^2u + \frac{1}{\lambda_n}\langle Tu,\phi_n^*\rangle Tx +$$

$$+ \frac{1}{\lambda_n}\langle Tx,\phi_n^*\rangle Tu + \frac{1}{\lambda_n}T_nTu - \frac{1}{\lambda_n}\langle Tu,\phi_n^*\rangle T_nx -$$

$$\frac{1}{\lambda_n}\langle Tx,\phi_n^*\rangle T_nu - Tu + \langle Tu,\phi_n^*\rangle x + \langle Tx,\phi_n^*\rangle u] =$$

$$= S_n[(T_n-T)u + \frac{1}{\lambda_n}(T_n-T)Tu + \frac{1}{\lambda_n}\langle Tu,\phi_n^*\rangle(T-T_n)(x-\hat{\phi}^{(n)}) +$$

$$+ \frac{1}{\lambda_n}\langle Tu,\phi_n^*\rangle(T-T_n)P\hat{\phi}^{(n)} + \frac{1}{\lambda_n}\langle T(x-\hat{\phi}^{(n)}+\hat{\phi}^{(n)}-\phi^{(n)}),\phi_n^*\rangle(T-T_n)u +$$

$$+ \frac{\lambda}{\lambda_n}(T-T_n)u + \langle Tu,\phi_n^*\rangle(x-\hat{\phi}^{(n)}+\hat{\phi}^{(n)}-\phi_n) +$$

$$+ \langle Tu,\phi_n^*\rangle\phi_n + \langle T(x-\hat{\phi}^{(n)}+\hat{\phi}^{(n)}-\phi^{(n)}),\phi_n^*\rangle u + (\lambda-\lambda_n)u] =$$

$$= \frac{\lambda-\lambda_n}{\lambda_n}S_n(T_n-T)u + \frac{1}{\lambda_n}S_n(T_n-T)Tu +$$

$$+ \frac{1}{\lambda_n}\langle Tu,\phi_n^*\rangle S_n(T-T_n)(x-\hat{\phi}^{(n)}) + \frac{1}{\lambda_n}\langle Tu,\phi_n^*\rangle S_n(T-T_n)P\hat{\phi}^{(n)} +$$

$$+ \frac{1}{\lambda_n} \langle T(x-\hat{\phi}^{(n)}), \phi_n^* \rangle S_n(T-T_n)u + \frac{\lambda}{\lambda_n} \langle \hat{\phi}^{(n)} - \phi^{(n)}, \phi_n^* \rangle S_n(T-T_n)u +$$

$$+ \langle Tu, \phi_n^* \rangle S_n(x-\hat{\phi}^{(n)}) + \langle Tu, \phi_n^* \rangle S_n(\hat{\phi}^{(n)} - \phi_n) +$$

$$+ \langle T(x-\hat{\phi}^{(n)}), \phi_n^* \rangle S_n u + \lambda \langle \hat{\phi}^{(n)} - \phi^{(n)}, \phi_n^* \rangle S_n u +$$

$$+ (\lambda - \lambda_n) S_n u.$$

Si l'on définit les constantes

$$\sigma_1 := \frac{\|P\|}{|\lambda|} \left(\frac{2\gamma}{|\lambda|} \left((\nu + \|T\|)(c+\eta\|T\| + |\lambda| c\eta) + |\lambda| \eta c \|T\| \right) + \right.$$

$$\left. + c\gamma(|\lambda|\eta + 1) \right) + \frac{2\gamma}{|\lambda|}$$

$$\sigma_2 := \frac{4\gamma\eta\|T\|}{|\lambda|} \left(\nu + \|T\| + \frac{|\lambda|}{2} \right)$$

et si n est assez grand pour que $|\lambda_n| > \frac{|\lambda|}{2}$ alors on a le résultat.

□

Nous sommes en position d'établir un théorème de convergence pour la famille de méthodes (B1,i(k)).

Théorème 3.3.4

Convergence de (B1,i(k)) *sous les hypothèses* H1, H2 *et* H3.
Si n *est suffisamment grand alors* $\phi^k \to \phi^{(n)}$ *lorsque* $k \to \infty$
pour toute valeur du paramètre entier i(k).
La convergence est linéaire si i(k) *est constante ;*
superlinéaire si $i(k) \to \infty$ *lorsque* $k \to \infty$ *et quadratique si la suite* $\{\nu_k\}$ *définie par*

$$\nu_k := \frac{\|U_n(\phi^k)\|^{i(k)+1}}{\|\phi^k - \phi^{(n)}\|}$$

est bornée par rapport à k.

Preuve : La démonstration suit le même chemin que celle du Théorème 3.2.4. Si l'on définit

$$\rho_1 := \text{Min}\{\frac{1}{\omega\mu}, \frac{1}{8\theta_1\mu}, \frac{\delta}{2\theta_1}, \frac{1}{4\sigma_2}, \frac{1}{4\omega\mu\sigma_2}, \delta\} \quad \text{et} \quad \rho := \frac{\rho_1}{2}$$

et si n est assez grand pour que

$$\text{Max}\{\sigma_1 \|(T-T_n)T\|, 2c\|(T-T_n)P\|\} < \frac{\rho_1}{4}$$

alors :

$$J^{(n)}(x) \in \text{Iso}(X) \qquad \forall x \in B(\hat{\phi}^{(n)}, \rho_1)$$

$$\text{Sup}\{\|(J^{(n)}(x))^{-1}\| : x \in B(\hat{\phi}^{(n)}, \rho_1)\} < \mu$$

$$\|J^{(n)}(x_1)-J^{(n)}(x_2)\| \leq \theta_1\|x_1-x_2\| < \frac{1}{4\mu}$$

$$\phi_n \in B(\phi^{(n)}, \rho) \subseteq B(\hat{\phi}^{(n)}, \rho_1)$$

$$\|U_n(x)\| \leq \frac{1}{2}\,\text{Min}\{1, \frac{1}{\omega\mu}\} \quad \forall x \in B(\hat{\phi}^{(n)}, \rho_1) \cap P_n^{-1}\{\phi_n\}.$$

Ceci montre que si $\phi^k \in B(\phi^{(n)}, \rho)$ et $\delta^k = (J^{(n)}(\phi^k))^{-1}F^{(n)}(\phi^k)$ alors :

$$\delta^k-\delta^k_j = \sum_{\ell=j+1}^{\infty} (U_n(\phi^k))^{\ell} S_n^{B} F^{(n)}(\phi^k) =$$

$$= (U_n(\phi^k))^{j+1}(J^{(n)}(\phi^k))^{-1}F^{(n)}(\phi^k)$$

$$\|\delta^k-\delta^k_j\| \leq \frac{1}{2^{j+1}}\,\text{Min}\{1, \frac{1}{(\omega\mu)^{j+1}}\}\omega\mu\|\phi^k-\phi^{(n)}\|$$

$$\phi^{k+1}-\phi^{(n)} = \phi^k-\delta^k-\phi^{(n)}+\delta^k-\delta^k_{i(k)}$$

$$\|\phi^{k+1}-\phi^{(n)}\| \leq \|\phi^k-\delta^k-\phi^{(n)}\| + \|\delta^k-\delta^k_{i(k)}\|$$

Ainsi que l'on a fait dans la preuve du Théorème 3.2.4. :

$$\|\phi^k-\delta^k-\phi^{(n)}\| \leq \frac{1}{4}\|\phi^k-\phi^{(n)}\|$$

Donc

$$\|\phi^{k+1}-\phi^{(n)}\| \leq \frac{3}{4}\|\phi^{k}-\phi^{(n)}\| \quad \text{pour toute valeur de } i(k).$$

Ainsi, $\phi^{k+1} \in B(\phi^{(n)},\rho)$ et $\phi^{k} \to \phi^{(n)}$ si $k \to \infty$.

Mais, on a la majoration

$$\|\phi^{k+1}-\phi^{(n)}\| \leq \theta_1\mu\|\phi^{k}-\phi^{(n)}\|^2 + \alpha_{i(k)}\|\phi^{k}-\phi^{(n)}\|$$

avec

$$\alpha_{i(k)} = \frac{\omega\mu}{2^{i(k)+1}} \ \text{Min}\{1, \frac{1}{(\omega\mu)^{i(k)+1}}\}$$

ce qui démontre les affirmations finales du théorème concernant la convergence linéaire, superlinéaire ou quadratique de $\{\phi^k\}$ vers $\phi^{(n)}$.

□

De même que (A1,i(k)), la famille (B1,i(k)) est du type de la méthode présentée dans la section 2.4, la place de la projection Q étant prise par la projection spectrale P_n.
En effet, le théorème 3.3.4 démontre que l'opérateur

$$G^B : (1-P_n)X \to (1-P_n)X$$

défini par $G^B(u) = S_n^B(u+F^{(n)}(\phi^k))$ $\forall u \in (1-P_n)X$ est un inverse approché de l'opérateur K autour de la solution δ^k du problème K(u) = 0, posé dans $(1-P_n)X$, car

$$1\big|_{(1-P_n)X} - G^B K = U_n(\phi^k)$$

et nous avons prouvé que nous sommes dans les hypothèses de la propriété 2.3.4 aussi bien que dans celles de la propriété 2.3.1 car
$r_\sigma(U_n(\phi^k)) \leq \|U_n(\phi^k)\|$.

Remarquons que la démarche suivie pour la construction de l'inverse approché G^B est tout à fait comparable à celle de la méthode de Brakhage présentée dans l'exemple 2.3.2. à propos de l'équation (T-z)x-f = 0.

Effectivement S_n vérifie l'identité (Cf. Propriété 1.1.20 appliquée à T_n) :

$$S_n = \frac{1}{\lambda_n}(S_n T_n - 1 + P_n)$$

donc si l'on substitue T à T_n -qui est la même technique que nous avons employée pour construire $R_n^B(z)$ à partir de $R_n(z)$ lors de l'exemple 2.3.2- nous sommes amenés à

$$S_n^B := \frac{1}{\lambda_n}(S_n T - 1 + P_n)$$

et la définition de G^B s'inscrit tout naturellement à partir de l'opérateur G de la section précédente.

L'approximation de la valeur propre λ est faite comme précédemment avec la suite $\{\lambda^k\}$ définie par (3.2.1.).

Quant à la méthode la plus simple de cette deuxième famille elle est bien évidemment la suivante :

$$\left.\begin{aligned} \phi^0 &:= \phi_n \\ \phi^{k+1} &:= \phi^k - S_n^B(T\phi^k - \lambda^{k+1}\phi^k) \quad k \geq 0 \end{aligned}\right\} (B1,0)$$

qui correspond au choix $i(k) = 0 \quad \forall k \geq 0$.

Théorème 3.3.5

Pour n *assez grand et sous les hypothèses générales H1, H2, H3 de ce chapitre, la suite* $\{\phi^k\}$ *définie par* (B1,0) *converge vers* $\phi^{(n)}$. *De plus il existe des constantes* β *et* β_o *telles que* $\forall k \geq 0$:

$$\text{Max}\{|\lambda - \lambda^{k+1}|, \| \phi^k - \phi^{(n)} \| \} \leq \beta(\gamma_n)^{k+1} \quad \text{où} \quad \gamma_n = \beta_o \|(T - T_n)T\|$$

Preuve : La convergence $\phi^k \to \phi^{(n)}$ est un corollaire trivial du théorème 3.3.4. Quant à la borne d'erreur, pour k = 0 elle découle du lemme 3.3.2 et des propriétés 1.3.2 et 1.3.5, l'induction s'achevant si l'on tient compte de l'inégalité

$$\|\phi^{k+1}-\phi^{(n)}\| \leq \theta_1\mu\|\phi^k-\phi^{(n)}\|^2+\omega\mu\|U_n(\phi^k)\|\,\|\phi^k-\phi^{(n)}\|$$

et du lemme 3.3.3 pour borner $\|U_n(\phi^k)\|$.

□

3.4 UNE AUTRE FORMULATION NON LINEAIRE

Dans les deux sections précédentes on a vu qu'il est possible de construire diverses méthodes, dont la convergence est mathématiquement fondée, pour la résolution numérique du problème non linéaire

$$F^{(n)}(x) = 0 \qquad x \neq 0 \quad x \in X$$

où l'opérateur $F^{(n)}$ défini par (3.1.3.) est, pour ainsi dire, une réalisation possible de celui présenté dans l'exemple 2.2.1, la place de la fonctionnelle ψ étant occupée par ϕ_n^*.

Il est donc clair que l'on peut aussi bien développer les mêmes idées que précédemment -et en particulier deux nouvelles familles de méthodes- en partant de l'opérateur $\tilde{F}^{(n)} : B(\hat{\phi}^{(n)},r) \to X$ défini pour $n \geq n_0$ et r assez petit par

$$\tilde{F}^{(n)}(x) = x - \frac{1}{\langle Tx,\phi_n^*\rangle}\,Tx \qquad \forall x \in B(\hat{\phi}^{(n)},r) \qquad (3.4.1.)$$

et qui peut être conçu comme une réalisation de l'opérateur $\tilde{F}$ de l'exemple 2.2.2. le rôle de ψ étant joué à nouveau par la fonctionnelle ϕ_n^*.

On introduit les notations

$$\tilde{J}^{(n)}(x) := D_x\,\tilde{F}^{(n)} \qquad x \in B(\hat{\phi}^{(n)},r)$$

$$\tilde{S}_n := -\lambda_n S_n \,.$$

$$\tilde{S}_n^B := 1-P_n-S_nT$$

qui permettent d'écrire les deux familles de méthodes auxquelles on aboutit ainsi :

$$\left.\begin{aligned} &\phi^0 := \phi_n \\ &\phi^{k+1} := \phi^k - \delta^k_{i(k)} \qquad k \geq 0 \\ &\delta^k_{i(k)} := \sum_{\ell=0}^{i(k)} (1-\tilde{S}_n \tilde{J}^{(n)}(\phi^k))^\ell \tilde{S}_n \tilde{F}^{(n)}(\phi^k) \end{aligned}\right\} \quad (A2,i(k))$$

$$\left.\begin{aligned} &\phi^0 := \phi_n \\ &\phi^{k+1} := \phi^k - \delta^k_{i(k)} \qquad k \geq 0 \\ &\delta^k_{i(k)} := \sum_{\ell=0}^{i(k)} (1-\tilde{S}^B_n \tilde{J}^{(n)}(\phi^k))^\ell \tilde{S}^B_n \tilde{F}^{(n)}(\phi^k) \end{aligned}\right\} \quad (B2,i(k))$$

dont la convergence se démontre comme dans les théorèmes 3.2.4. et 3.3.4 respectivement. Bien entendu, tous les rayons des boules qu'y interviennent doivent être pris inférieurs à r.

On reconnaît de nouveau le traitement de l'inversion de la dérivée $\tilde{J}^{(n)}(\phi^k)$ à l'aide de i(k) itérations portées sur deux méthodes de correction du résidu définies pour la résolution du problème $\tilde{K}(u) = 0$ où l'opérateur $\tilde{K} : (1-P_n)X \to (1-P_n)X$ est défini par

$$\tilde{K}(u) = \tilde{J}^{(n)}(\phi^k)u - \tilde{F}^{(n)}(\phi^k) \qquad \forall u \in (1-P_n)X$$

Les inverses approchés associés à ces deux méthodes sont

$\tilde{G} : (1-P_n)X \to (1-P_n)X$ défini par

$\tilde{G}(u) = \tilde{S}_n(u+\tilde{F}^{(n)}(\phi^k))$ $\forall u \in (1-P_n)X$

et

$\tilde{G}^B : (1-P_n)X \to (1-P_n)X$ défini par

$\tilde{G}^B(u) = \tilde{S}^B_n(u+\tilde{F}^{(n)}(\phi^k))$ $\forall u \in (1-P_n)X$

La famille (A2,i(k)) correspond à l'utilisation de $\tilde{G}$ et la famille (B2,i(k)) à l'utilisation de $\tilde{G}^B$.

On peut, comme dans la section précédente, motiver la construction de $\tilde{G}^B$ à partir de $\tilde{G}$ si l'on tient compte de l'identité

$$\tilde{S}_n = S_n T_n - 1 + P_n$$

où l'on substitue T à T_n pour obtenir ainsi $\tilde{S}_n^B$.

Pour ce qui est de l'approximation de la valeur propre λ on considère comme précédemment la suite $\{\lambda^k\}$ définie par (3.2.1.).

Les méthodes les plus simples des familles présentées ci-dessus correspondent au cas $i(k) = 0 \ \forall k$.

En utilisant la notation (3.2.1.) elles s'écrivent, respectivement,

$$\left.\begin{aligned} \phi^0 &:= \phi_n \\ \phi^{k+1} &:= \phi^k + \lambda_n S_n(\phi^k - \frac{1}{\lambda^{k+1}} T\phi^k) \quad k \geq 0 \end{aligned}\right\} \quad (A2,0)$$

et

$$\left.\begin{aligned} \phi^0 &:= \phi_n \\ \phi^{k+1} &:= \frac{1}{\lambda^{k+1}} T\phi^k + S_n T(\phi^k - \frac{1}{\lambda^{k+1}} T\phi^k) \quad k \geq 0 \end{aligned}\right\} \quad (B2,0)$$

On peut démontrer la convergence par des théorèmes tout à fait analogues aux théorèmes 3.2.6, 3.2.7 et 3.3.5, notamment, sous les mêmes hypothèses que ceux-ci.

3.5 REMARQUES ET COMMENTAIRES BIBLIOGRAPHIQUES

A propos des méthodes appelées ici les plus simples il y a beaucoup de commentaires à faire.

Tout d'abord disons qu'elles peuvent être interprétées comme étant des méthodes de correction du résidu, (Cf. Ahués et al. 1982). Effectivement pour ce qui est des méthodes (A1,0) et (A2,0) on considère l'hypothèse $T_n \xrightarrow{\|\|\|\|} T$ ce qui nous permet de montrer que les opérateurs G_0 et $\tilde{G}_0$ définis sur X par

$$G_0(x) = \phi_n + (S_n - \frac{1}{\lambda_n} P_n)x \qquad \forall x \in X$$

$$\tilde{G}_o(x) = \phi_n + (P_n - \lambda_n S_n)x \qquad \forall x \in X$$

sont, respectivement, un inverse approché de $F^{(n)}$ et un inverse approché de $\tilde{F}^{(n)}$, tous deux autour de $\phi^{(n)}$, et pour n suffisamment grand.

En ce qui concerne les méthodes (B1,0) et (B2,0) on considère l'hypothèse générale $T_n - z \xrightarrow{fs} T - z$ sur Γ_λ qui démontre que les opérateurs G_o^B et $\tilde{G}_o^B$ définis sur X par

$$G_o^B(x) = \phi_n + \frac{1}{\lambda_n}(S_n T - 1)x \qquad \forall x \in X$$

$$\tilde{G}_o^B(x) = \phi_n + (1 - S_n T)x \qquad \forall x \in X$$

sont, respectivement, un inverse approché de $F^{(n)}$ et un inverse approché de $\tilde{F}^{(n)}$, tous deux autour de $\phi^{(n)}$ et pour n assez grand.

Deuxièmement, remarquons que ces quatre méthodes plus simples peuvent être écrites comme suit :

$$\phi^{k+1} := (1-w_k)\phi^k + w_k\zeta^k - \frac{1}{\lambda_n} S_n T_n r^k \qquad \text{(A1,0)}$$

$$\phi^{k+1} := (1-w_k)\phi^k + w_k\zeta^k - \frac{1}{\lambda_n} S_n T r^k \qquad \text{(B1,0)}$$

$$\phi^{k+1} := \zeta^k + S_n T_n \tilde{r}^k \qquad \text{(A2,0)}$$

$$\phi^{k+1} := \zeta^k + S_n T \tilde{r}^k \qquad \text{(B2,0)}$$

où

$$\phi^0 := \phi_n, \quad \lambda^{k+1} := \langle T\phi^k, \phi_n^* \rangle,$$

$$\zeta^k := \frac{1}{\lambda^{k+1}} T\phi^k, \qquad w_k := \frac{\lambda^{k+1}}{\lambda_n}$$

$$r^k := F^{(n)}(\phi^k), \qquad \tilde{r}^k := \tilde{F}^{(n)}(\phi^k)$$

On peut identifier ζ^k au résultat d'un pas d'une itération de type puissance itérée à partir de ϕ^k ; r^k et $\tilde{r}^k$ sont les résidus en ϕ^k des équations $F^{(n)}(x) = 0$ et $\tilde{F}^{(n)}(x) = 0$ respectivement ; finalement, w_k joue le rôle d'un paramètre de relaxation variable (Cf. Kaspar 1982).

Très souvent, en pratique, l'approximation T_n de l'opérateur compact T est aussi un opérateur compact. Si l'on pense à un opérateur intégral T, à noyau continu, de type Fredholm défini sur l'espace $X = C[0,1]$, l'image par T d'une fonction $x \in X$ est une fonction Tx généralement plus lisse que x. Supposons que cela est vrai et qu'il en est de même pour l'approximation T_n. Alors les méthodes dont on parle admettent l'interprétation suivante :

Il s'agit de méthodes qui font soit une itération de type puissance itérée, soit celle-ci combinée avec une relaxation à paramètre variable, et en plus, une correction de l'itéré ainsi obtenu combinée à un lissage du résidu, ce lissage étant fait soit par l'approximation T_n de T, soit par l'opérateur T lui-même.

La terminologie que l'on vient d'utiliser, bien qu'elle ne soit pas très précise, est pourtant très fréquente dans le domaine -actuellement en plein développement- des techniques dites de multigrilles, et desquelles nos méthodes se rapprochent un peu du fait, au moins, qu'elles utilisent, en pratique, deux niveaux de discrétisation : l'un pour l'approximation T_n et l'autre pour la représentation de T lors des calculs numériques (Cf. Ahués et Chatelin 1983).

Au sujet des méthodes multigrille, le lecteur intéressé peut consulter, par exemple, Hackbusch 1979, 1981 a, 1981 b ; Hemker et Schippers 1981 et Nicolaïdes 1979.

Du point de vue de la méthode de Newton citons le travail de Cubillos 1980 qui propose deux itérations comparables aux méthodes (A1,0) et (B1,0). Il considère l'espace produit $Z = X \times C$ où $X = C[0,1]$ est muni de la norme du max. La norme de Z est définie par

$$\|(x,\mu)\|_Z := \|x\|+|\mu| \qquad (x,\mu) \in Z$$

où $\|\ \|$ est la norme de X. Le problème de valeurs propres est mis sous la forme non linéaire suivante :

$$A(\phi,\lambda) := ((T-\lambda)\phi, \tfrac{1}{2}((\phi,\phi)_2-1)) = (0,0) \qquad (3.5.1)$$

où $(\ ,\ )_2$ est le produit scalaire réel défini par

$$(x,y)_2 = \int_0^1 x(t)y(t)dt$$

Le problème (3.5.1) est résolu numériquement par deux méthodes de type Newton à dérivée approchée. L'inverse de la dérivée de A est approchée respectivement par deux opérateurs qui ne dépendent pas de l'itération. Ces deux approximations vérifient entre elles une relation comparable à celle qui existe entre S_n et S_n^B. La convergence de ces deux méthodes est linéaire mais il faut remarquer que les résultats numériques ne sont pas satisfaisants. D'autre part, les notions de projection spectrale et de convergence fortement stable n'y sont pas considérées de manière explicite. Finalement, on remarque que dans ce travail X est considéré en tant qu'espace vectoriel sur $\mathbb{R}$ ce qui n'est pas l'habitude lorsque l'on étudie des problèmes spectraux.

Aussi dans l'optique des méthodes de type Newton on remarque que l'approximation δ_o^k fournie par (A1,0) s'inscrit dans le cadre des approximations étudiées dans Axelsson 1982, du fait que δ_o^k vérifie l'inégalité

$$\|M_k[F^{(n)}(\phi^k)-J^{(n)}(\phi^k)\delta_o^k]\| \le \rho_k\|M_kF^{(n)}(\phi^k)\|$$

pour un certain opérateur régulier M_k et un certain nombre $\rho_k \in]0,1[$.

Effectivement, dans notre cas $M_k = S_n|_{(1-P_n)X}$ et ρ_k est une constante de Lipschitz de l'opérateur $1|_{(1-P_n)X} - S^nJ^{(n)}(\phi^k)|_{(1-P_n)X}$ laquelle peut être choisie dans l'intervalle $]0,1[$ si $T_n \xrightarrow{\|\ \|} T$ et pour n assez grand. Sous cette hypothèse une analyse similaire est valable pour la méthode (A2,0). Et il en est de même pour (B1,0) et (B2,0) si l'on ne considère que l'hypothèse générale $T_n-z \xrightarrow{fs} T-z$ sur Γ_λ .

La méthode (A1,0) est voisine d'une méthode proposée par Stewart 1973 pour le cas de matrices. Elle a été aussi étudiée dans le contexte des méthodes multigrilles (Cf. Ahués et Chatelin 1983).

La méthode (A2,0) constitue une généralisation d'une méthode due à Lin Qun 1982 pour des opérateurs différentiels elliptiques à résolvante compacte et sous l'hypothèse de convergence en norme:

Les méthodes (B1,0) et (B2,0) ont été suggérées dans Ahués et al 1983 b,c.

Une autre méthode de raffinement itératif, la méthode de Chatelin, peut être dérivée à partir de la théorie de perturbations d'opérateurs linéaires (Cf. Kato 1966) moyennant une généralisation des résultats de Rayleigh-Schrödinger pour le calcul des développements en série des éléments propres. Ce travail est présenté -quoique dans deux contextes topologiques quelque peu différents- dans Lemordant 1980 et dans Chatelin 1983. La méthode en question s'écrit

$$\left.\begin{array}{lll} \nu^0 := \lambda_n \qquad \nu^k := \langle (T-T_n)\eta^{k-1}, \phi_n^* \rangle & & k \geq 1 \\ \eta^0 := \phi_n \qquad \eta^k := S_n[(T_n-T)\eta^{k-1} + \sum_{i=1}^{k} \nu^i \eta^{k-i}] & & k \geq 1 \\ \lambda^k := \sum_{i=0}^{k} \nu^i, \qquad \phi^k := \sum_{i=0}^{k} \eta^i & & k \geq 0 \end{array}\right\} \quad (3.5.2.)$$

On remarque que le calcul de ϕ^k nécessite des itérés $\phi^0,\ldots,\phi^{k-1}$ ce qui rapproche cette méthode des familles (A1,i(k)), (A2,i(k)), (B1,i(k)) et (B2,i(k)), tout au moins pour le choix

$$i(k) = k+j_0 \qquad (j_0 \geq 0 \text{ fixe})$$

qui garantit pour ces quatre familles une convergence superlinéaire. Cependant, la convergence de la méthode (3.5.2) est linéaire -aussi bien en théorie qu'en pratique. A priori on établit (Cf. Chatelin 1983, pp. 258-262 et 298-302) :

$$|\lambda-\lambda^k| \leq c' \frac{q_o^{k+1}}{1-q_o}$$

$$\|\phi^k-\phi^{(n)}\| \leq c'' \frac{q^{k+1}}{1-q}$$

où q_o, q sont des réels tels que

$$\text{Max}\{r_\sigma((T-T_n)R(T,z)) : z \in \Gamma_\lambda\} < q_o < q < 1.$$

Une méthode similaire à (3.5.2) peut être construite pour l'approximation de valeurs propres non simples et du sous-espace invariant associé ou bien pour la localisation de ce qu'on appelle *un groupe de valeurs propres* (Cf. Chatelin 1983 pp. 302-319, Lemordant 1980 pp. 109-130). Une étude de la convergence de cette méthode ainsi que son application à l'équation de Schrödinger sont faites dans Kulkarni 1982.

Les méthodes proposées dans cette thèse peuvent être généralisées au cas d'une valeur propre de multiplicité $m > 1$.

Supposons, pour simplifier, que $m = 2$.

On considère l'espace $X \times X$ muni de la norme produit

$$\|\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\| = (\|x_1\|^2+\|x_2\|^2)^{1/2} \qquad x_1 \in X, \quad x_2 \in X$$

Ce choix entraîne

$$X^* \times X^* = (X \times X)^*$$

en tant qu'espaces de Banach, c'est-à-dire : chaque élément $\psi = (\psi_1,\psi_2) \in X^* \times X^*$ définit une fonctionnelle, qu'on note aussi ψ, dans $(X \times X)^*$ par

$$\langle \psi, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rangle = \langle \psi_1, x_1 \rangle + \langle \psi_2, x_2 \rangle \qquad (3.5.3.)$$

et réciproquement, tout élément ψ de $(X \times X)^*$ définit une paire d'éléments ψ_1,ψ_2 de X^* tels que (3.5.3) est vérifié. En outre,

$$|\psi| = (|\psi_1|^2 + |\psi_2|^2)^{1/2}$$

(Cf. Kato 1966, p. 164).

Un opérateur H dans $L(X \times X)$ est caractérisé par quatre opérateurs $H_{ij} \in L(X)$ $\quad i = 1,2 \; ; \; j = 1,2$:

$$H\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{bmatrix} H_{11}x_1 + H_{12}x_2 \\ H_{21}x_1 + H_{22}x_2 \end{bmatrix}$$

Soit $\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}$ une matrice complexe.

Rappelons que sa *trace* est définie par

$$\mathrm{tr}\,\alpha = \alpha_{11} + \alpha_{22}$$

On définit la loi de composition externe

$$\alpha x := \begin{bmatrix} \alpha_{11}x_1 + \alpha_{12}x_2 \\ \alpha_{21}x_1 + \alpha_{22}x_2 \end{bmatrix}, \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Etant donné $\psi = (\psi_1, \psi_2) \in (X \times X)^*$ on définit la matrice complexe, dite *matrice de Gram :*

$$g(\psi, x) := \begin{bmatrix} \langle \psi_1, x_1 \rangle & \langle \psi_1, x_2 \rangle \\ \langle \psi_2, x_1 \rangle & \langle \psi_2, x_2 \rangle \end{bmatrix} \qquad (3.5.4.)$$

Notons $M_2(C)$ l'espace des matrices complexes de taille 2×2. L'application

$$g : (X \times X)^* \times (X \times X) \to M_2(C)$$

définie par (3.5.4.) vérifie

$$g(\psi,\alpha x) = g(\psi,x)\alpha^H$$

$$\langle\psi,x\rangle = \operatorname{tr} g(\psi,x)$$

pour tout $\alpha \in M_2(C)$, $\psi \in (X \times X)^*$, $x \in X \times X$.
On définit

$$g(x,\psi) := g(\psi,x)^H$$

de façon à ce qu'on ait

$$g(\alpha x,\psi) = \alpha g(x,\psi)$$

Soit $T \in L(X)$. On définit l'opérateur d(T) dans $L(X \times X)$, qu'on appelle *opérateur diagonal* associé à T, par

$$d(T) = \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix}$$

On suppose T compact. Soit $\lambda \in Q\sigma(T)$ une valeur propre de multiplicité m = 2.

Etant donné $\{T_n\}$, approximation fortement stable de T autour de λ, il existe, pour n assez grand, une paire μ_n', μ_n'' de valeurs propres de T_n qui approchent λ . Eventuellement $\mu_n' = \mu_n''$.

Soit $\{\phi_n', \phi_n''\}$ une base du sous-espace invariant de T_n associé à la partie $\{\mu_n', \mu_n''\}$ de son spectre. On note P_n la projection spectrale associée. On peut représenter P_n à l'aide de deux vecteurs ψ_n', ψ_n'' du sous-espace $P_n^* X^*$ normalisés par

$$\langle\phi_n',\psi_n'\rangle = \langle\phi_n'',\psi_n''\rangle = 1$$

$$\langle\phi_n',\psi_n''\rangle = \langle\phi_n'',\psi_n'\rangle = 0 \qquad (3.5.5.)$$

En effet

$$P_n x = \langle x,\psi_n'\rangle\phi_n' + \langle x,\psi_n''\rangle\phi_n''$$

Remarquons qu'on peut identifier la base de $P_n X$ à un élément $\phi_n = \begin{pmatrix} \phi_n' \\ \phi_n'' \end{pmatrix}$ de $X \times X$ et la base de $P_n^* X^*$ à un élément $\psi_n = (\psi_n', \psi_n'')$ de $(X \times X)^*$. Alors les conditions (3.5.5) s'écrivent :

$$g(\phi_n, \psi_n) = I_2 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

En outre, il existe $\phi^{(n)} = \begin{pmatrix} \phi_1^{(n)} \\ \phi_2^{(n)} \end{pmatrix} \in X \times X$ tel que $\{\phi_1^{(n)}, \phi_2^{(n)}\}$ est une base du sous-espace invariant de T associé à λ et tel que

$$g(\phi^{(n)}, \psi_n) = I_2$$

A l'aide de l'opérateur diagonal d(T) on peut exprimer l'invariance par T du sous-espace engendré par $\{\phi_1^{(n)}, \phi_2^{(n)}\}$ ainsi :

Il existe $\alpha \in M_2(C)$ telle que

$$d(T)\phi^{(n)} = \alpha\phi^{(n)}$$

Cette matrice α est facile à déterminer

$$\alpha = g(d(T)\phi^{(n)}, \psi_n)$$

Ceci veut dire que $\phi^{(n)}$ est une solution de l'équation non linéaire posée dans $X \times X$:

$$A(x) := d(T)x - g(d(T)x, \psi_n)x = 0$$

La généralisation de la méthode (A1,0) est la suivante :

$$\phi^0 := \phi_n$$

$$\phi^{k+1} := \phi^k - \Sigma_n A(\phi^k)$$

où l'opérateur $\Sigma_n \in L(X \times X)$ est défini comme suit :

Etant donné $u \in X \times X$, $v := \Sigma_n u$ est la solution de

$$d(1-P_n)(d(T_n) - \mu_n)v = d(1-P_n)u$$
$$d(P_n)v = 0 \qquad (3.5.6)$$

où

$$\mu_n := g(d(T_n)\phi_n, \psi_n)$$

On remarque que :

a) Si $\mu_n' \neq \mu_n''$ alors μ_n' et μ_n'' sont des valeurs propres simples de T_n. Si l'on choisit ϕ_n' et ϕ_n'' comme les vecteurs propres de T_n respectivement associés à μ_n' et μ_n'' et si l'on choisit ψ_n' et ψ_n'' comme les vecteurs propres de T_n^* respectivement associés à $\bar{\mu}_n'$ et $\bar{\mu}_n''$, alors

$$\mu_n = \begin{bmatrix} \mu_n' & 0 \\ 0 & \mu_n'' \end{bmatrix}$$

En général, il existe une matrice $\beta \in M_2(C)$ telle que

$$\beta^H\beta = I_2, \qquad \beta^H\mu_n\beta = \begin{bmatrix} \mu_n' & \delta \\ 0 & \mu_n'' \end{bmatrix}$$

où δ n'est pas nécessairement nul.

b) Puisque $\lambda \neq 0$, on a pour n assez grand,

$$\mu_n' \neq 0 \quad , \mu_n'' \neq 0$$

donc la matrice μ_n est régulière. Ceci permet de généraliser la méthode (B1,0). Il en est de même pour les méthodes (A2,0) et (B2,0).

c) Comme approximation de la valeur propre λ on peut considérer le *quotient de Rayleigh généralisé* (Cf. Chatelin 1983) définit par

$$\lambda^{k+1} := \frac{1}{2} \langle d(T)\phi^k, \psi_n \rangle$$

C'est-à-dire :

$$\lambda^{k+1} = \frac{1}{2}\left(\langle T\phi_1^k, \psi_n'\rangle + \langle T\phi_2^k, \psi_n''\rangle\right) = \frac{1}{2}\,\mathrm{tr}\alpha^{k+1}$$

où

$$\phi^k = \begin{pmatrix} \phi_1^k \\ \phi_2^k \end{pmatrix} \quad \text{et } \alpha^{k+1} := g(d(T)\phi^k, \psi_n)$$

d) La résolution de (3.5.6) est un problème plus difficile que le calcul de la résolvante réduite S_n dans le cas d'une valeur propre simple. On peut dire que Σ_n est une *résolvante réduite généralisée*. En termes de matrices le système (3.5.6) nous amène à résoudre des équations du type

$$BV - VD = E \qquad (3.5.7.)$$

où l'inconnue V est une matrice de taille $N \times 2$, E l'est aussi, B est de taille $N \times N$ et D est une matrice de taille 2×2 ; N est la dimension de l'espace où l'on pose le problème de valeurs propres. L'équation (3.5.7) est étudiée, par exemple, dans Davis et Kahan 1970, Stewart 1973, Varah 1970, Bartels et Stewart 1972 et Golub, Nash et Van Loan 1979.

CHAPITRE 4

SUR LA DISCRETISATION D'OPERATEURS INTEGRAUX DE TYPE FREDHOLM :

Représentation matricielle des méthodes de projection et de quadrature approchée

INTRODUCTION

On considère deux classes de discrétisation d'opérateurs intégraux de type Fredholm définis sur C[0,1] et à noyau continu : la discrétisation *par projection* et celle *par quadrature approchée*. A la première classe appartiennent les approximations de *Galerkin*, de *projection* proprement dite et de *Sloan* tandis que dans la deuxième classe on retrouve les approximations de *Fredholm* et de *Nyström*.

Les discrétisations par projection seront construites en utilisant comme sous-espace d'approximation l'espace des *fonctions continues, linéaires par morceaux sur une partition uniforme* de l'intervalle [0,1] à n points et comme famille de projections les *opérateurs d'interpolation* respectifs aux noeuds de la partition, définis sur l'espace C[0,1]. Les discrétisations par quadrature approchée seront construites en utilisant la *règle du trapèze sur une partition uniforme* de [0,1] à n points.

Seules les approximations de Galerkin -qui correspond dans ce cas à la méthode dite *de collocation* (aux noeuds de la partition)- et celle de Fredholm constituent une *discrétisation complète* qui admet par tant une représentation matricielle *carrée* de taille n. Cependant, les autres approximations peuvent être mises en oeuvre sur un ordinateur moyennant une discrétisation de la variable qui reste continue beaucoup plus fine que celle qui définit l'ordre n de l'approximation.

En utilisant une quadrature par trapèzes sur cette partition fine -qui aura disons N points : $N >> n$- on définit une troisième classe d'approximations de T que l'on pourrait nommer *approximation par projection d'ordre bas à quadrature d'ordre élevé*, l'ordre de l'approximation étant, évidemment, celui de la projection (et non pas celui de la quadrature !).

On présente dans ce chapitre, d'une manière assez détaillée, la représentation matricielle en dimension N de ces trois classes d'approximations T_n d'ordre n d'un opérateur intégral T de type Fredholm défini sur C[0,1] et à noyau continu. On montre les relations qui existent entre

les opérateurs résolvante, projection spectrale et résolvante réduite associés à ces matrices carrées de taille N et ceux qui correspondent à une certaine matrice carrée de taille n qui leur est associée par l'intermédiaire de deux applications, dites prolongement et restriction, lesquelles relient la partition fine à N points -qui sert à décrire la variable continue et dont le caractère est auxiliaire- à la partition fondamentale à n points qui définit l'ordre de l'approximation considérée.

Dans ce contexte, l'opérateur T est lui-même représenté par une matrice carrée de taille N.

Le fait d'avoir une formulation matricielle nous permet, finalement, de faire une étude du coût des itérations proposées dans le chapitre précédent.

4.1 GENERALITES

On considère un noyau $\kappa : [0,1] \times [0,1] \to C$ continu et l'opérateur intégral T sur $X = C[0,1]$ qui lui est associé :

$$(Tx)(t) = \int_0^1 \kappa(t,s)x(s)ds, \qquad 0 \le t \le 1,\ x \in X \qquad (4.1.1.)$$

On a vu dans l'exemple 1.1.3 que $T \in L(X)$ et que T est compact.

Soit $\pi_n : X \to X$ la projection d'interpolation continue linéaire par morceaux définie dans l'exemple 1.2.2.

$$\pi_n x = \sum_{i=1}^{n} \langle x, e_i^{(n)*} \rangle e_i^{(n)} \qquad \forall x \in X$$

où $\{e_i^{(n)}\}$ est la base de fonctions chapeaux de l'espace $X_n = \pi_n X$ associée à la partition de [0,1] induite par les points $t_i^{(n)} = \frac{i-1}{n-1}$ $i = 1,\ldots,n$, et $\{e_i^{(n)*}\}$ est la base adjointe de l'espace X_n^* :

$$\langle x, e_i^{(n)*} \rangle = x(t_i^{(n)}) \qquad i = 1,\ldots,n,\ x \in X.$$

Les approximations T_n^P, T_n^S et T_n^G de l'exemple 1.2.1 prennent la forme

$$(T_n^P x)(t) = \sum_{i=1}^{n} \int_0^1 \kappa(t_i^{(n)},s)x(s)ds\ e_i^{(n)}(t)$$

$$(T_n^S x)(t) = \sum_{i=1}^{n} x(t_i^{(n)}) \int_0^1 \kappa(t,s)\ e_i^{(n)}(s)ds$$

$$(T_n^G x)(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} x(t_i^{(n)}) \int_0^1 \kappa(t_j^{(n)},s)e_i^{(n)}(s)ds\ e_j^{(n)}(t)$$

On remarque que si l'on substitue à $\int_0^1 \kappa(t,s)e_i^{(n)}(s)ds$ son approximation par quadrature de trapèzes associée aux points $\{t_i^{(n)}\}$ alors T_n^S devient l'approximation T_n^N et T_n^G devient T_n^F (Cf. exemple 1.2.2).

Or, de ces cinq approximations de T, seules T_n^G et T_n^F constituent ce que l'on appelle une *discrétisation complète :* en t et en s. Du point de vue du calcul numérique sur un ordinateur on ne peut traiter la discrétisation de T que par une représentation matricielle. Pour cette raison, la variable qui reste continue dans T_n^P, T_n^S et T_n^N sera discrétisée par l'intermédiaire d'une partition de [0,1] beaucoup plus fine que celle des points $\{t_i^{(n)}\}$. En fait, on va substituer à X l'espace X_N où $N >> n$.

Plus précisément : soient n et N deux entiers soumis aux conditions suivantes :

i) $N >> n \geq 2$

ii) $(N-1)/(n-1)$ est un entier.

Associés à ces deux entiers on a respectivement les espaces X_n et X_N dont on précise les bases des fonctions chapeaux et ses adjointes.

$$\mathcal{H}_n = \{e_1^{(n)},\ldots,e_n^{(n)}\} \quad , \qquad \mathcal{H}_n^* = \{e_1^{(n)*},\ldots,e_n^{(n)*}\}$$

$$\mathcal{H}_N = \{e_1^{(N)},\ldots,e_N^{(N)}\} \quad , \qquad \mathcal{H}_N^* = \{e_1^{(N)*},\ldots,e_N^{(N)*}\}$$

associées aux points

$$t_j^{(n)} = \frac{j-1}{n-1} \quad j = 1,\ldots,n \; ; \qquad t_i^{(N)} = \frac{i-1}{N-1} \quad i = 1,\ldots,N$$

D'après la condition ii) ci-dessus on a l'inclusion

$$\{t_j^{(n)}\} \subseteq \{t_i^{(N)}\}$$

Cette inclusion entraîne le fait suivant qui est capital :

X_n *est un sous-espace de* X_N.

On peut donc exprimer $e_j^{(n)}$ comme une combinaison linéaire (unique) des éléments de la base $\mathcal{H}_N$:

$$e_j^{(n)} = \sum_{i=1}^{N} p_{ij}\, e_i^{(N)} \qquad j = 1,\ldots,n$$

les coefficients p_{ij} étant donnés par

$$p_{ij} = \langle e_j^{(n)}, e_i^{(N)*} \rangle \quad i = 1,\ldots,N \; ; \; j = 1,\ldots,n$$

La matrice $p = (p_{ij})$ de taille $N \times n$ est identifiée à une application linéaire de C^n dans C^N que l'on appelle *prolongement* de X_n dans X_N. Si $u \in C^n$ est la colonne des coordonnées d'un vecteur de X_n relatives à la base $\mathcal{H}_n$ alors $pu \in C^N$ est la colonne des coordonnées relatives à la base $\mathcal{H}_N$ de ce même vecteur considéré cette fois-ci comme élément de l'espace X_N. On peut penser à p comme étant la représentation matricielle, relative aux bases $\mathcal{H}_n$ et $\mathcal{H}_N$, de *l'injection canonique* de X_n dans X_N.

D'autre part, soit $r = (r_{ji})$ la matrice de taille $n \times N$ définie par

$$r_{ji} = \langle e_i^{(N)}, e_j^{(n)*} \rangle \quad j = 1,\ldots,n \; ; \quad i = 1,\ldots,N$$

où de façon équivalente :

$$r_{ji} = \begin{cases} 1 & \text{si } t_j^{(n)} = t_i^{(N)} \\ 0 & \text{si non} \end{cases} \qquad j = 1,\ldots,n \; ; \; i = 1,\ldots,N.$$

On identifie r à une application linéaire de C^N dans C^n dite *restriction* de X_N à X_n. Si $x \in C^N$ est la colonne des coordonnées d'un vecteur de X_N par rapport à la base $\mathcal{H}_N$ alors $rx \in C^n$ est la colonne des coordonnées, relatives à la base $\mathcal{H}_n$, de l'interpolant de ce vecteur dans l'espace X_n aux noeuds $\{t_j^{(n)}\}$.

On définit la matrice

$$\pi = pr$$

qui est une matrice carrée de taille N.

Lemme 4.1.1

i) $p : C^n \to C^N$ *est injectif*

ii) $r : C^N \to C^n$ *est surjectif*

iii) $rp = I_n$

iv) $\pi^2 = \pi$

Preuve : Triviale. □

Remarquons que $\pi : C^N \to C^N$ représente en termes de coordonnées relatives à la base $\mathcal{H}_N$ la *projection d'interpolation* de X_N sur X_n aux noeuds $\{t_j^{(n)}\}$.

Ceci implique que dans la base $\mathcal{H}_N^*$, $\pi^H = r^H p^H$ représente une projection de X_N^* sur X_n^*. Les fonctionnelles $e_j^{(n)*}$ peuvent s'exprimer en termes de la base $\mathcal{H}_N^*$ en faisant intervenir la matrice $r^H = (r_{ij})$. En effet :

$$e_j^{(n)*} = \sum_{i=1}^{N} r_{ji}\, e_i^{(N)*} \qquad j = 1,\ldots,N$$

et l'on en déduit que r^H correspond à la représentation, en termes de coordonnées relatives aux bases $\mathcal{H}_n^*$ et $\mathcal{H}_N^*$, du prolongement de X_n^* dans X_N^*.

4.2 DISCRETISATION PAR PROJECTION

L'approximation T_n^G admet une représentation à partir de la matrice carrée $\mathcal{G}_n^G$ de taille n dont le coefficient en position (ligne k, colonne j) est

$$(\mathcal{G}_n^G)_{kj} = \int_0^1 \kappa(t_k^{(n)},s) e_j^{(n)}(s)ds \qquad k,j = 1,\ldots,n.$$

Or, dûment plongé dans X_N, l'opérateur T_n^G sera représenté, par rapport à la base $\mathcal{H}_N$, par la matrice carrée de taille N :

$$[T_n^G] = p\, \mathcal{G}_n^G\, r$$

De même, pour la représentation matricielle carrée de taille N des opérateurs T_n^P et T_n^S, plongés dans X_N, on considère les matrices $\mathcal{C}_n^P$ de taille $n \times N$ et $\mathcal{C}_n^S$ de taille $N \times n$ définies par

$$(\mathcal{C}_n^P)_{ki} = \int_0^1 \kappa(t_k^{(n)},s)e_i^{(N)}(s)ds \quad k = 1,\ldots,n;\ i = 1,\ldots,N.$$

$$(\mathcal{C}_n^S)_{ik} = \int_0^1 \kappa(t_i^{(N)},s)e_k^{(n)}(s)ds \quad i = 1,\ldots,N\ ;\ k = 1,\ldots,n.$$

qui permettent de les représenter dans la base $\mathcal{H}_N$ par les matrices :

$$[T_n^P] = p\,\mathcal{C}_n^P$$

$$[T_n^S] = \mathcal{C}_n^S\,r$$

Si l'on prend comme représentation de T la matrice carrée, de taille N, $\mathcal{C}_N^G$ définie par

$$(\mathcal{C}_N^G)_{i\ell} = \int_0^1 \kappa(t_i^{(N)},s)e_\ell^{(N)}(s)ds \quad i,\ell = 1,\ldots,N.$$

Alors on a le :

Lemme 4.2.1

i) $[T_n^G] = \pi\mathcal{C}_N^G\,\pi$

ii) $[T_n^P] = \pi\mathcal{C}_N^G$

iii) $[T_n^S] = \mathcal{C}_N^G\pi$

Preuve : Triviale □

Lemme 4.2.2

i) $[T_n^G]^H = [T_n^{G*}]$

ii) $[T_n^P]^H = [T_n^{P*}]$

iii) $[T_n^S]^H = [T_n^{S*}]$

Preuve : Cela vient du fait que l'on a pris pour ces représentations matricielles des couples de bases qui sont *adjointes* : $\mathcal{H}_n$ et $\mathcal{H}_n^*$ en dimension n et $\mathcal{H}_N$ et $\mathcal{H}_N^*$ en dimension N.

□

On introduit les notations suivantes :

$$\mathcal{R}_n^G(z) := (\mathcal{E}_n^G - z\, I_n)^{-1} \qquad \text{pour } z \in \rho(\mathcal{E}_n^G)$$

λ_n : valeur propre simple non nulle de $\mathcal{E}_n^G$

u_n : vecteur propre de $\mathcal{E}_n^G$ associé à λ_n

v_n : vecteur propre de $(\mathcal{E}_n^G)^H$ associé à $\bar{\lambda}_n$ et tel que $v_n^H u_n = 1$

$$\mathcal{P}_n^G := u_n v_n^H$$

$$\mathcal{S}_n^G := \lim_{z \to \lambda_n} \mathcal{R}_n^G(z)\,(I_n - \mathcal{P}_n^G)$$

On remarque que $\mathcal{P}_n^G$ est la projection spectrale de $\mathcal{E}_n^G$ associée à λ_n et que $\mathcal{S}_n^G$ est donc la résolvante réduite correspondante.

Théorème 4.2.3

Si $z \in \rho(\mathcal{E}_n^G)$ *et* $z \neq 0$ *alors*

$z \in \rho([T_n^G]) \cap \rho([T_n^P]) \cap \rho([T_n^S])$ *et l'on a*

$$[R_n^G(z)] = p(\mathcal{R}_n^G(z) + \frac{1}{z} I_n)r - \frac{1}{z} I_N$$

$$[R_n^P(z)] = \frac{1}{z}\,(p\mathcal{R}_n^G(z)\mathcal{E}_n^P - I_N)$$

$$[R_n^S(z)] = \frac{1}{z}\,(\mathcal{E}_n^S \mathcal{R}_n^G(z) r - I_N)$$

Preuve : On considère l'équation $([T_n^G]-z\ I_N)x = f$.
Elle s'écrit aussi $(p\ \mathcal{C}_n^G\ r - z\ I_N)x = f$.
On multiplie à gauche par r et l'on obtient, d'après le lemme 4.1.1., successivement :

$$(rp\ \mathcal{C}_n^G\ r - zr)x = rf$$

$$(\mathcal{C}_n^G - z\ I_n)rx = rf$$

$$rx = \mathcal{R}_n^G(z)\ rf$$

et en multipliant à gauche par p :

$$\pi x = p\mathcal{R}_n^G(z)rf.$$

Or, si l'on pose $x = \pi x + (I_N-\pi)x$ dans l'équation de départ on a les relations qui suivent :

$$(p\ \mathcal{C}_n^G r - zI_N)(p\mathcal{R}_n^G(z)rf + (I_N-\pi)x) = f$$

$$p\ \mathcal{C}_n^G\mathcal{R}_n^G(z)rf + p\ \mathcal{C}_n^G r(I_N-\pi)x - zp\mathcal{R}_n^G(z)rf - z(I_N-\pi)x = f$$

Mais $\mathcal{C}_n^G\mathcal{R}_n^G(z) = I_n + z\mathcal{R}_n^G(z)$

$$r(I_N-\pi) = r-rpr = 0$$

donc finalement $(I_N-\pi)x = -\frac{1}{z}\ (I_N-\pi)f$.

Ceci montre que

$$x = (p(\mathcal{R}_n^G(z) + \frac{1}{z}\ I_n)r - \frac{1}{z}\ I_N)f$$

d'où le résultat concernant $[R_n^G(z)]$.

Maintenant on considère l'équation $([T_n^P]-zI_N)x = f$ que l'on multiplie à gauche par $I_N-\pi$ pour obtenir

$$(I_N-\pi)x = -\frac{1}{z}\ (I_N-\pi)f.$$

Ensuite on substitue $\pi x+(I_N-\pi)x$ à x et l'on multiplie l'équation de départ par π à gauche. Ceci donne

$$([T_n^G]-zI_N)\pi x = \pi f + \frac{1}{z}\pi[T_n^P](I_N-\pi)f$$

En utilisant le résultat obtenu pour $[R_n^G(z)]$ on aboutit à l'expression

$$x = \frac{1}{z}(p\mathcal{R}_n^G(z)\mathcal{C}_n^P - I_N)f$$

si l'on tient compte de l'identité

$$[R_n^G(z)][T_n^P]\pi = I_N + z[R_n^G(z)].$$

On laisse au lecteur la démonstration du résultat concernant $[R_n^S(z)]$, (Cf. Le cas de $[R_n^N(z)]$ dans le théorème 4.3.1.).

□

Théorème 4.2.4

Si $\lambda_n \neq 0$ *est une valeur propre simple de* $\mathcal{C}_n^G$ *alors elle l'est aussi des matrices* $[T_n^G]$, $[T_n^P]$ *et* $[T_n^S]$. *En outre, étant donnés* u_n *et* v_n *les vecteurs déjà définis on a que*

$\phi_n^G = p\,u_n$ *est un vecteur propre de* $[T_n^G]$

$\phi_n^P = p\,u_n$ *l'est de* $[T_n^P]$

$\phi_n^S = \frac{1}{\lambda_n}\mathcal{C}_n^S u_n$ *l'est de* $[T_n^S]$

tous associés à λ_n ;

$\phi_n^{G*} = r^H v_n$ *est un vecteur propre de* $[T_n^{G*}]$

$\phi_n^{P*} = \frac{1}{\bar{\lambda}_n}(\mathcal{C}_n^P)^H v_n$ *l'est de* $[T_n^{P*}]$

$\phi_n^{S*} = r^H v_n$ *l'est de* $[T_n^{S*}]$

tous associés à $\bar{\lambda}_n$.

D'autre part, les projections spectrales et les résolvantes réduites des matrices $[T_n^G], [T_n^P]$ *et* $[T_n^S]$ *, en* λ_n*, sont respectivement :*

$$[P_n^G] = p\,\mathcal{P}_n^G\,r$$

$$[P_n^P] = \frac{1}{\lambda_n}\,p\,\mathcal{P}_n^G\,\mathcal{C}_n^P$$

$$[P_n^S] = \frac{1}{\lambda_n}\,\mathcal{C}_n^S\,\mathcal{P}_n^G\,r$$

et

$$[S_n^G] = p\,\mathcal{S}_n^G\,r - \frac{1}{\lambda_n}\,(I_N - \pi)$$

$$[S_n^P] = \frac{1}{\lambda_n}\,p\,\mathcal{S}_n^G\,\mathcal{C}_n^P + \frac{1}{\lambda_n^2}\,p\,\mathcal{P}_n^G\,\mathcal{C}_n^P - \frac{1}{\lambda_n}\,I_N$$

$$[S_n^S] = \frac{1}{\lambda_n}\,\mathcal{C}_n^S\,\mathcal{S}_n^G\,r + \frac{1}{\lambda_n^2}\,\mathcal{C}_n^S\,\mathcal{P}_n^G\,r - \frac{1}{\lambda_n}\,I_N$$

Preuve :

i) Les vecteurs ϕ_n^G, ϕ_n^P et ϕ_n^S :

D'après le lemme 4.1.1. on a $\phi_n^G \neq 0$. Or $[T_n^G]\phi_n^G = p\,\mathcal{C}_n^G\,rpu_n = \lambda_n\,pu_n = \lambda_n\phi_n^G$. Mais si x est un vecteur propre de $[T_n^G]$ associé à λ_n alors $\mathcal{C}_n^G\,rx = \lambda_n\,rx$ donc il existe $\alpha \in C$ tel que $rx = \alpha u_n$ d'où $\alpha\lambda_n\phi_n^G = \lambda_n x$ et $x = \alpha\phi_n^G$ car $\lambda_n \neq 0$. Ceci montre que λ_n est une valeur propre simple de $[T_n^G]$ et que ϕ_n^G est un vecteur propre associé. De même, $\phi_n^P \neq 0$ et $[T_n^P]\phi_n^P = \pi\,\mathcal{C}_N^G\phi_n^P = pr\,\mathcal{C}_N^G\,pu_n = p\,\mathcal{C}_n^G\,u_n = \lambda_n\,\phi_n^P$. Si x est un vecteur propre de $[T_n^P]$ associé à λ_n alors $\pi\,\mathcal{C}_N^G\,x = \lambda_n x$ d'où $x \in X_n$ et $x = \pi x$. Or $\mathcal{C}_n^G\,rx = \lambda_n rx$ implique $x = \alpha\phi_n^P$ pour un certain $\alpha \in C$. ϕ_n^S vérifie $r\phi_n^S = u_n \neq 0$ donc $\phi_n^S \neq 0$. D'autre part $[T_n^S]\phi_n^S = \mathcal{C}_n^S\,r\phi_n^S = \mathcal{C}_n^S\,u_n = \lambda_n\phi_n^S$. Si x est un vecteur propre de $[T_n^S]$ associé à λ_n alors $\mathcal{C}_n^S\,rx = \lambda_n x$ et $\mathcal{C}_n^G\,rx = \lambda_n\,rx$ donc $rx = \alpha u_n$ pour un $\alpha \in C$. Ceci implique $\alpha\,\mathcal{C}_n^S\,u_n = \lambda_n\,x$ d'où $x = \alpha\,\phi_n^S$.

ii) Pour ce qui est de ϕ_n^{G*}, ϕ_n^{P*}, ϕ_n^{S*} les démonstrations sont analogues aux précédentes.

iii) Les formules des projections spectrales découlent du fait que

$$(\phi_n^{G*})^H \phi_n^G = (\phi_n^{P*})^H \phi_n^P = (\phi_n^{S*})^H \phi_n^S = 1$$

donc

$$[P_n^G] = \phi_n^G (\phi_n^{G*})^H \quad , \quad [P_n^P] = \phi_n^P (\phi_n^{P*})^H$$

et

$$[P_n^S] = \phi_n^S (\phi_n^{S*})^H .$$

iv) Les formules des résolvantes réduites sont une conséquence des formules des résolvantes (Cf. Théorème 4.2.3), des formules des projections spectrales et des définitions

$$[S_n^G] := \lim_{z \to \lambda_n} [R_n^G(z)](I_N - [P_n^G])$$

$$[S_n^P] := \lim_{z \to \lambda_n} [R_n^P(z)](I_N - [P_n^P])$$

$$[S_n^S] := \lim_{z \to \lambda_n} [R_n^S(z)](I_N - [P_n^S]).$$

□

4.3 DISCRETISATION PAR QUADRATURE APPROCHEE

On considère maintenant les approximations T_n^F et T_n^N définies dans l'exemple 1.2.2.

Soit $\mathcal{C}_n^F$ la matrice carrée de taille n dont le coefficient en position (ligne k, colonne j) est

$$(\mathcal{C}_n^F)_{kj} = \omega_j^{(n)} \kappa(t_k^{(n)}, t_j^{(n)}) \qquad k,j = 1,\ldots,n.$$

La discrétisation à N points de la variable qui reste continue dans T_n^N nous conduit à définir la matrice $\mathcal{C}_n^N$ de taille $N \times n$:

$$(\mathcal{C}_n^N)_{ij} = \omega_j^{(n)} \kappa(t_i^{(N)}, t_j^{(n)}) \qquad i = 1,\ldots,N \; ; \; j = 1,\ldots,n.$$

Alors, plongés dans X_N, les opérateurs T_n^F et T_n^N admettent les représentations matricielles carrées de taille N suivantes par rapport à la base $\mathcal{H}_N$:

$$[T_n^F] = p\,\mathcal{T}_n^F\,r$$

$$[T_n^N] = \mathcal{T}_n^N\,r$$

et il en est de même pour les adjoints :

$$[T_n^{F*}] = [T_n^F]^H$$

$$[T_n^{N*}] = [T_n^N]^H$$

On introduit les notations suivantes :

$$\mathcal{R}_n^F(z) := (\mathcal{T}_n^F - z\,I_n)^{-1} \quad \text{pour } z \in \rho(\mathcal{T}_n^F)$$

λ_n : valeur propre simple non nulle de $\mathcal{T}_n^F$

u_n : vecteur propre de $\mathcal{T}_n^F$ associé à λ_n

v_n : vecteur propre de $(\mathcal{T}_n^F)^H$ associé à $\bar{\lambda}_n$ et tel que $v_n^H u_n = 1$

$\mathcal{P}_n^F := u_n v_n^H$, projection spectrale de $\mathcal{T}_n^F$ en λ_n

$\mathcal{S}_n^F := \lim_{z\to\lambda_n} \mathcal{R}_n^F(z)(I_n - \mathcal{P}_n^F)$, résolvante réduite associée.

Théorème 4.3.1

Si $z \in \rho(\mathcal{T}_n^F)$ *et* $z \neq 0$ *alors*

$z \in \rho([T_n^F]) \cap \rho([T_n^N])$ *et l'on a*

$$[R_n^F(z)] = p(\mathcal{R}_n^F(z) + \frac{1}{z}\,I_n)r - \frac{1}{z}\,I_N$$

$$[R_n^N(z)] = \frac{1}{z}\,(\mathcal{T}_n^N \mathcal{R}_n^F(z) r - I_N).$$

Preuve : La formule de $[R_n^F(z)]$ se démontre de la même façon que la formule de $[R_n^G(z)]$ dans le théorème 4.2.3. : il n'y a qu'à changer G par F.

Nous démontrons celle de $[R_n^N(z)]$. Les relations suivantes peuvent être facilement vérifiées :

$([T_n^N] - zI_N)x = f$ s'écrit $(\mathcal{C}_n^N r - zI_N)x = f$ d'où $(r\mathcal{C}_n^N - zI_n)rx = rf$
qui équivaut à $(\mathcal{C}_n^F - z\, I_n)rx = rf$, donc $rx = \mathcal{R}_n^F(z)\, rf$ et
$\pi x = p\mathcal{R}_n^F(z)rf$.
Or $([T_n^N] - zI_N)(\pi x+(I_N-\pi)x) = f$ implique
$(\mathcal{C}_n^N r - zI_N)\pi x + \mathcal{C}_n^N r(I_N-pr)x - z(I_N-\pi)x = f$
d'où $x = \frac{1}{z}[\mathcal{C}_n^N\mathcal{R}_n^F(z)\, r - I_N]f$ ce qui démontre le théorème.

□

Théorème 4.3.2

Si $\lambda_n \neq 0$ *est une valeur propre simple de* $\mathcal{C}_n^F$ *alors elle l'est aussi des matrices* $[T_n^F]$ *et* $[T_n^N]$. *En outre, si* u_n *et* v_n *sont les vecteurs définis dans cette section alors*

$\phi_n^F = p\, u_n$ *est un vecteur propre de* $[T_n^F]$

$\phi_n^N = \frac{1}{\lambda_n}\mathcal{C}_n^N u_n$ *l'est de* $[T_n^N]$

tous deux associés à λ_n;

$\phi_n^{F*} = r^H v_n$ *est un vecteur propre de* $[T_n^{F*}]$

$\phi_n^{N*} = r^H v_n$ *l'est de* $[T_n^{N*}]$

associés ceux-ci à $\bar{\lambda}_n$.

D'autre part on a les formules suivantes pour les projections spectrales et les résolvantes réduites respectives :

$$[P_n^F] = p\,\mathcal{P}_n^F\,r$$

$$[P_n^N] = \frac{1}{\lambda_n}\,\mathcal{T}_n^N\,\mathcal{P}_n^F\,r$$

et

$$[S_n^F] = p\,\mathcal{S}_n^F\,r - \frac{1}{\lambda_n}\,(I_N - \pi)$$

$$[S_n^N] = \frac{1}{\lambda_n}\,\mathcal{T}_n^N\,\mathcal{S}_n^F\,r + \frac{1}{\lambda_n^2}\,\mathcal{T}_n^N\,\mathcal{P}_n^F\,r - \frac{1}{\lambda_n}\,I_N$$

Preuve : Le résultat concernant ϕ_n^F se démontre de la même manière que celui de ϕ_n^G dans le Théorème 4.2.4. où le lemme 4.2.1. n'a pas été utilisé. Pour ϕ_n^N on applique formellement la preuve donnée pour ϕ_n^S où le Lemme 4.2.1 n'a pas été utilisé non plus. Les autres sont alors triviaux.

□

Normalement l'opérateur T est approché dans ce contexte par une discrétisation complète par quadrature à N points: $\mathcal{T}_N^F$.

$$(\mathcal{T}_N^F)_{i\ell} = \omega_\ell^{(N)}\,\kappa(t_i^{(N)}, t_\ell^{(N)}) \qquad i,\ell = 1,\ldots,N.$$

Il faut pourtant signaler que l'on n'a pas un lemme analogue au Lemme 4.2.1 : on ne peut pas établir une formule qui donne $[T_n^F]$ ou $[T_n^N]$ à partir de $\mathcal{T}_N^F$ en faisant intervenir la seule projection π. Remarquons à ce propos que les poids de la quadrature ne sont pas les mêmes dans $\mathcal{T}_n^F$ et dans $\mathcal{T}_N^F$.

4.4 DISCRETISATION PAR PROJECTION AVEC QUADRATURE FINE

Si l'on part de la discrétisation complète par quadrature à N points, $\mathcal{T}_N^F$, définie à la fin de la section précédente, on peut construire trois approximations en prennant, dans un certain sens, le Lemme 4.2.1. comme définition de celles-ci. Il s'agit donc d'approximations par projection d'ordre n (soit sur X_n) couplée avec une quadrature d'ordre N. On les notera comme suit :

$$\mathcal{T}_n^{qG} := r\,\mathcal{T}_N^F\,p, \qquad [T_n^{qG}] := \pi\mathcal{T}_N^F\,\pi = p\,\mathcal{T}_n^{qG}\,r$$

$$\mathcal{T}_n^{qP} := r\,\mathcal{T}_N^F, \qquad [T_n^{qP}] := \pi\mathcal{T}_N^F = p\,\mathcal{T}_n^{qP}$$

$$\mathcal{T}_n^{qS} := \mathcal{T}_N^F\,p, \qquad [T_n^{qS}] := \mathcal{T}_N^F\,\pi = \mathcal{T}_n^{qS}\,r$$

Bien sûr, le résultat suivant va de soi.

Théorème 4.4.1

Le lemme 4.2.2. et les théorèmes 4.2.3 et 4.2.4. restent valables si l'on substitue qG *à* G , qP *à* P *et* qS *à* S .

Preuve : Triviale. □

Sur la convergence vers T de ces approximations on a le lemme suivant :

Lemme 4.4.2

Lorsque $n \to \infty$ *et* $N \to \infty$ *avec* $N \geq n$ *on a*

$$T_n^{qG} \xrightarrow{cc} T,\ T_n^{qP} \xrightarrow{cc} T \quad \text{et} \quad T_n^{qS} \xrightarrow{cc} T$$

Preuve : Cf. Chatelin 1983, pp. 198-199. □

4.5 LES LIMITES DES SUITES $\{(\lambda^k, \phi^k)\}$ DANS $C \times X_N$.

Lorsque les itérations proposées au chapitre 3 sont traduites dans leurs représentations matricielles il se pose la question de la limite vers laquelle elles vont converger. Or, dans le chapitre 4, seules les deux discrétisations de grande taille $\mathcal{T}_N^G$ et $\mathcal{T}_N^F$ ont été considérées pour la représentation de T. Elles correspondent aux représentations matricielles dans la base $\mathcal{H}_N$ des approximations T_N^G et T_N^F respectivement, lesquelles constituent des discrétisations complètes de T dont le domaine est X et l'espace image X_N. $\mathcal{T}_N^G$ (resp. $\mathcal{T}_N^F$) représente en fait l'opérateur restreint $T_N^G|_{X_N}$ (resp. $T_N^F|_{X_N}$).

Cela veut dire que si les évaluations de T, lors des itérations du chapitre 3, sont faites en substituant à T la matrice $\mathcal{T}_N^G$ (resp. $\mathcal{T}_N^F$) alors la limite de la suite d'itérés $\{(\lambda^k,\phi^k)\}$ correspondante -si la convergence a lieu- sera à une distance de $(\lambda,\phi^{(n)})$ de l'ordre de $\|(T_N^G-T)P\|$ (resp. $\|(T_N^F-T)P\|$) d'après les propriétés 1.3.2 et 1.3.5.

Autrement dit, *la limite de* $\{(\lambda^k,\phi^k)\}$, *lorsque la convergence a lieu, est le couple* (λ_N,ϕ_N) *d'éléments propres de la discrétisation de grande taille qui approchent* $(\lambda,\phi^{(n)})$. (Cf. Chatelin 1983, pp. 255-298).

Ceci nous permet de reformuler le problème entièrement en termes discrets.

Etant donné un opérateur intégral T sur X = C[0,1], de type Fredholm à noyau continu, on lui associe deux approximations représentées par les matrices carrées $\mathcal{T}_n$, de taille n, et $\mathcal{T}_N$, de taille N, où N >> n. $\mathcal{T}_N$ est la discrétisation dite *fine* et $\mathcal{T}_n$ est la discrétisation dite *grossière*. Le problème à résoudre est l'approximation numérique de la solution du problème de valeurs propres.

$$\mathcal{T}_N\phi_N = \lambda_N\phi_N \qquad \phi_N \in C^N, \quad \phi_N \neq 0$$

Les méthodes proposées au chapitre 3 lors de sa formulation matricielle, constituent des techniques itératives qui raffinent la solution du problème

$$\mathcal{T}_n u_n = \lambda_n u_n \qquad u_n \in C^n \qquad u_n \neq 0$$

comme approximation initiale de (λ_N,ϕ_N).

Remarquons, pourtant, que c'est parce que $\mathcal{T}_n$ et $\mathcal{T}_N$ sont *deux discrétisations d'un même opérateur* T que nous pouvons considérer $\mathcal{T}_n$ comme étant une approximation de $\mathcal{T}_N$.

4.6 COUT DES CALCULS MATRICIELS

On introduit les notations suivantes :

$\mathcal{C}_n$: matrice carrée de taille n dont λ_n est une valeur propre simple et non nulle.

$$\mathcal{C}_n \in \{\mathcal{C}_n^G, \mathcal{C}_n^F, \mathcal{C}_n^{qG}\}.$$

u_n : vecteur propre de $\mathcal{C}_n$ associé à λ_n.

v_n : vecteur propre de $\mathcal{C}_n^H$ associé à $\bar{\lambda}_n$ et tel que $v_n^H u_n = 1$.

$\mathcal{P}_n$, $\mathcal{S}_n$: projection spectrale et résolvante réduite de $\mathcal{C}_n$ en λ_n.

$\mathcal{C}_N$: matrice carrée de taille N qui sert à effectuer les produits qui représentent les évaluations de T.

E_D : ensemble des discrétisations présentées dans cette thèse et associées à un couple (n,N) donné

$$E_D = \{G,P,S,F,N,qG,qP,qS\}.$$

Le coût des différents calculs est estimé par le *nombre de produits* effectués. Par produit on entend soit une *multiplication*, soit une *division* (en général dans C). Les additions, soustractions, permutations de lignes d'une matrice ainsi que les comparaisons entre deux réels ne sont pas considérées lors du calcul des coûts.

Etant précisée une suite de calculs σ son coût sera noté $C(\sigma)$ tout en acceptant parfois quelques abus de notation dans ce contexte.

a) Calcul du produit $y_n = \mathcal{P}_n x_n$

Etant donné que u_n et v_n sont normalisés par $v_n^H u_n = 1$ on en déduit que

$$\mathcal{P}_n = u_n v_n^H$$

donc pour le produit $\mathcal{P}_n x_n$ avec $x_n \in C^n$ on a :

$$C(\mathcal{P}_n x_n) = 2n.$$

b) Factorisation matricielle du système linéaire associé au calcul de $\mathcal{S}_n$.

Soit $x_n \in C^n$. Le calcul de $y_n = \mathcal{S}_n x_n$ nécessite de la résolution du système linéaire de rang n

$$(\mathcal{C}_n - \lambda_n I_n) y_n = (I_n - \mathcal{P}_n) x_n$$

$$v_n^H y_n = 0$$

dont les inconnues sont les n coordonnées de y_n.

Supposons que l'on applique la méthode de triangularisation de Gauss (à pivot partiel) sur la matrice $A^{(1)}$ définie par :

$$A^{(1)} = \left[\begin{array}{c:c} \mathcal{C}_n - \lambda_n I_n & I_n \\ \hdashline v_n^H & 0 \ \ldots\ldots \ 0 \end{array}\right]$$

qui est de taille $(n+1) \times (2n)$.

Si à l'étape k du procédé de Gauss on a la structure

$$A^{(k)} = \left[\begin{array}{ccccc} 1 & & 0 & & 0 \\ 0 & 1 & & & \\ & 0 & a_{k,k}^{(k)} & 1 & \\ & & & & 1 \\ 0 & 0 & & & 0 \end{array}\right] = (a_{ij}^{(k)})$$

alors le passage à $A^{(k+1)}$ entraîne les calculs suivants :

i) Diviser par $a_{k,k}^{(k)}$ (qui est non nul quitte à faire une permutation de lignes) les coefficients $a_{k,j}^{(k)}$, $j = k+1,\ldots,n+k$ ce qui donne respectivement $a_{k,j}^{(k+1)}$ $j = k+1,\ldots,n+k$.
Ceci implique la réalisation de n produits.

ii) Multiplier par $a_{k+i,k}^{(k)}$ $i = 1,\ldots,n+1-k$ les coefficients $a_{k,j}^{(k+1)}$ $j = k+1,\ldots,n+k$.
Ceci a un coût de $n(n+1-k)$ produits.

Or le passage de $A^{(k)}$ à $A^{(k+1)}$ doit se faire pour $k = 1,\ldots,n$. Lorsque $k = n$ on aura la matrice

$$A^{(n+1)} = \left[\begin{array}{ccccc|cc} 1 & & & & & 0 & 0 \\ 0 & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & & & & & & \\ 0 & & & & 0 & 0 & 0 \end{array}\right]$$

avec $\mathcal{U}$ dans le bloc supérieur gauche et $\mathcal{L}^{-1}$ dans le bloc inférieur droit,

dont on se sert pour définir les matrices $\mathcal{U}$ et $\mathcal{L}^{-1}$.
Remarquons que $\mathcal{S}_n$, $\mathcal{U}$, $\mathcal{L}^{-1}$ et $\mathcal{P}_n$ vérifient

$$\mathcal{U}\mathcal{S}_n = \mathcal{L}^{-1}(I_n - \mathcal{P}_n)$$

$\mathcal{U}$ est une matrice régulière ; $\mathcal{L}^{-1}$ l'est aussi et la matrice

$$\mathcal{B} = \mathcal{L}^{-1}(I_n - \mathcal{P}_n)$$

est de rang n-1. Or, lors des calculs sur l'ordinateur il se peut que ce ne soit pas ainsi à cause des erreurs d'arrondi. Les zéros de la ligne n+1 de $A^{(n+1)}$ ne le sont qu'à la précision des calculs près.

Le coût de la construction des matrices $\mathcal{U}$ et $\mathcal{L}^{-1}$ est, d'après les paragraphes i) et ii) ci-dessus :

$$C(\mathcal{U}, \mathcal{L}^{-1}) = \frac{n^3+3n^2}{2}$$

L'algorithme précédent calcule $\mathcal{L}^{-1}$ au lieu de $\mathcal{L}$. Or, la factorisation $\mathcal{L},\mathcal{U}$ conventionnelle (selon le schéma dit de Crout) a un coût du même ordre que $C(\mathcal{U}, \mathcal{L}^{-1})$ et il faut remarquer que la résolution du système triangulaire $\mathcal{L}x = y$ a le même coût que le produit $y = \mathcal{L}^{-1}x$.

D'autre part il est utile de signaler que la factorisation $\mathcal{U}$, $\mathcal{L}^{-1}$ (ou bien $\mathcal{L},\mathcal{U}$) ne convient pas si l'on a à résoudre moins de n fois le système concernant le calcul de la résolvante réduite $\mathcal{S}_n$. Néanmoins, pour fixer les idées on supposera dans la suite que l'on fait la décomposition en facteurs triangulaires $\mathcal{U}$, $\mathcal{L}^{-1}$ décrite ci-dessus.

c) Calcul du produit $y_n = \mathcal{S}_n x_n$ ($\mathcal{U}$ et $\mathcal{L}^{-1}$ donnés).

Etant donnés les matrices $\mathcal{U}$ et $\mathcal{L}^{-1}$ du paragraphe b) précédent, le calcul de $y_n = \mathcal{S}_n x_n$ nécessite :

i) Du produit $b_n = (I_n - \mathcal{P}_n)x_n$ qui coûte $2n$.

ii) Du produit $c_n = \mathcal{L}^{-1} b_n$ qui coûte $\frac{(n+1)n}{2}$ et

iii) De la résolution du système linéaire triangulaire de taille n : $\mathcal{U} y_n = c_n$ qui coûte $\frac{(n-1)n}{2}$.

Donc le coût total est

$$C(\mathcal{U}^{-1}\mathcal{L}^{-1}(I_n - \mathcal{P}_n)x_n) = n^2+2n.$$

d) Calcul de la restriction $x_n = r\, x_N$.

Ce calcul est gratuit. Etant donné $x_N \in C^N$ le vecteur $x_n = r\, x_N$ n'est constitué que de certaines coordonnées de x_N dûment choisies. On le voit, par exemple, sur la Figure 4.6.1 où $N = 9$ et $n = 3$.

e) Calcul du prolongement $x_N = p\, x_n$.

Etant donné $x_n \in C^n$ le calcul de $x_N = p\, x_n$ entraîne le calcul de N-n valeurs dont chacune est une combinaison linéaire convexe de deux coordonnées de x_n. (Voir Figure 4.6.2).

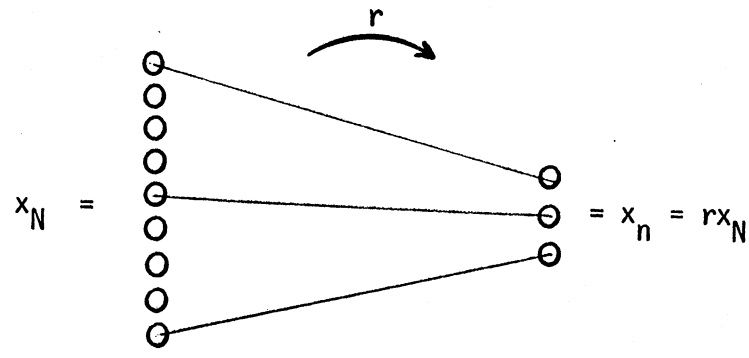
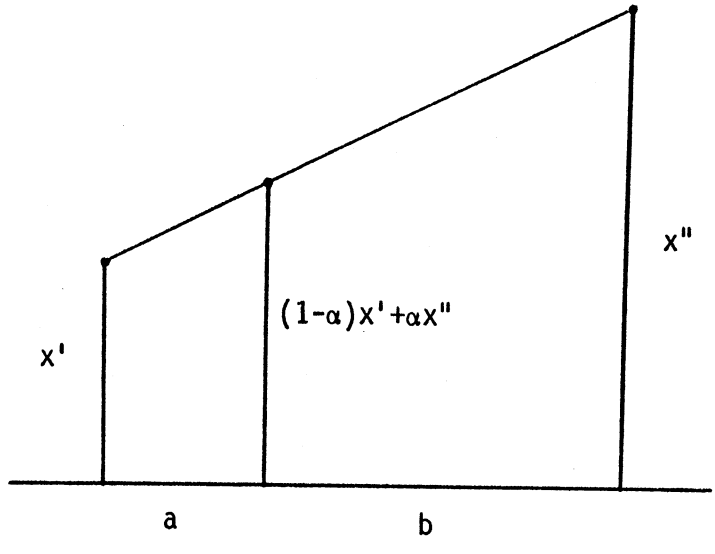


FIGURE 4.6.1.



$$\alpha = \frac{a}{a + b}$$

FIGURE 4.6.2.

Cela veut dire que l'on a à calculer N-n expressions du type $(1-\alpha)x'+\alpha x''$ y compris le calcul du paramètre α qui est un quotient différent pour chacune d'elles. Donc :

$$C(p\ x_n) = 3(N-n).$$

f) Calcul du produit $y_N = \mathcal{C}_N x_N$.

$\mathcal{C}_N$ étant, en général, une matrice pleine, le calcul de $y_N = \mathcal{C}_N\ x_N$ a un coût

$$C(\mathcal{C}_N\ x_N) = N^2$$

g) Calcul des vecteurs ϕ_n^{D*} $\quad D \in E_D$

$$C(\phi_n^{D*}) = \begin{cases} 0 & \text{si } D \in \{G,F,qG,S,N,qS\} \\ Nn+n & \text{si } D \in \{P,qP\} \end{cases}$$

h) Calcul des résidus.

$$C(r^k) = C(\tilde{r}^k) = \begin{cases} N^2+N+n & \text{si } D \in \{G,F,qG,S,N,qS\} \\ N^2+2N & \text{si } D \in \{P,qP\} \end{cases}$$

Ceci du fait que $(\phi_n^{D*})^H x_N$ coûte N si $D \in \{P,qP\}$ et n si non.

i) Calcul des vecteurs ϕ_n^D $\quad D \in E_D$.

$$C(\phi_n^D) = \begin{cases} 3(N-n) & \text{si } D \in \{G,F,qG,S,N,qS\} \\ Nn+n & \text{si } D \in \{P,qP\} \end{cases}$$

j) Calcul de $y_N = [S_n^D]x_N \quad D \in E_D$.

Sans tenir compte du coût de calcul de $\mathcal{L}^{-1}, \mathcal{U}$, on a

$$C([S_n^D]x_N) = \begin{cases} 4N + n^2-n & \text{si } D \in \{G,F,qG\} \\ Nn+4N+n^2 & \text{si } D \in \{P,qP\} \\ Nn+N+n^2+3n & \text{si } D \in \{S,N,qS\} \end{cases}$$

Or, pour le calcul de $y_N = [S_n^D][T_n^D]x_N$ *il ne convient pas de procéder séquentiellement :* $w_N = [T_n^D]x_N$, $y_N = [S_n^D]w_N$. En fait on peut construire pour $[S_n^D][T_n^D]$ une formule dûment simplifiée en termes de $\mathcal{E}_n^D$, $\mathcal{S}_n$, $\mathcal{P}_n$, r et p. Ceci donne les coûts suivants :

$$C([S_n^D T_n^D]x_N) = \begin{cases} 3N+2n^2-n & \text{si } D \in \{G,F,qG\} \\ 2Nn+6N+n^2-2n & \text{si } D \in \{P,qP\} \\ 2Nn+n^2+4n & \text{si } D \in \{S,N,qS\} \end{cases}$$

4.7 COUT DES ITERATIONS

Les coûts des différents calculs matriciels qui interviennent lors des itérations décrites au chapitre 3 ayant été établis dans la section 4.6 on peut maintenant faire les comptes pour ce qui est des itérations elles-mêmes.

a) Coût de base

On entend par cela le coût des calculs qui ne se font qu'une seule fois avant de commencer les itérations. Il comprend alors :

i) La construction des matrices $\mathcal{U}$ et $\mathcal{L}^{-1}$.

ii) Le calcul des vecteurs ϕ_n^D et ϕ_n^{D*} $D \in E_D$.

On ne considère comme faisant partie de ce coût ni le calcul des matrices $\mathcal{E}_n$, $\mathcal{E}_n^D$, $\mathcal{E}_N$, $D \in E_D$, ni le calcul des éléments propres λ_n, u_n, v_n du fait que cette information peut nous parvenir par les moyens les plus divers.

Le coût de base est donc

$$C_B := C(\mathcal{U}, \mathcal{L}^{-1}, \phi_n^D, \phi_n^{D*}) = \begin{cases} \frac{n^3}{2} + 3N + \frac{3n^2}{2} - 3n & \text{si } D \in \{G,F,qG\} \\ \frac{n^3}{2} + Nn + 3N + \frac{3n^2}{2} - 2n & \text{si } D \in \{P,qP\} \\ \frac{n^3}{2} + Nn + \frac{3n^2}{2} + n & \text{si } D \in \{S,N,qS\} \end{cases}$$

b) Coût des itérations les plus simples.

D'après la section 3.5 on remarque que les itérations (A1,0) et (A2,0) admettent deux formulations selon si l'opérateur T_n intervient explicitement ou non. Ces deux formulations, bien qu'équivalentes en théorie, elles peuvent ne pas l'être en pratique.

On présente ci-dessous un résumé de ces itérations dites les plus simples (car $i(k) = 0 \ \forall k$) tout en donnant les différentes versions :

$$\phi^{k+1} = \phi^k - S_n r^k \qquad \text{(A1,0)}$$

$$\phi^{k+1} = \phi^k + \frac{1}{\lambda_n}(1 - S_n T_n) r^k \qquad \text{(A1,0) bis}$$

$$\phi^{k+1} = \phi^k + \lambda_n S_n \tilde{r}^k \qquad \text{(A2,0)}$$

$$\phi^{k+1} = \frac{1}{\lambda^{k+1}} T\phi^k + S_n T_n \tilde{r}^k \qquad \text{(A2,0) bis}$$

$$\phi^{k+1} = \phi^k + \frac{1}{\lambda_n}(1 - S_n T) r^k \qquad \text{(B1,0)}$$

$$\phi^{k+1} = \frac{1}{\lambda^{k+1}} T\phi^k + S_n \tilde{Tr}^k \qquad \text{(B2,0)}$$

avec

$$\phi^0 = \phi_n, \quad \lambda^{k+1} = \langle T\phi^k, \phi_n^* \rangle$$

$$r^k = T\phi^k - \lambda^{k+1}\phi^k, \quad \tilde{r}^k = \phi^k - \frac{1}{\lambda^{k+1}} T\phi^k$$

Bien entendu, tous les objets mathématiques ci-dessus seront remplacés lors du calcul par leur représentant matriciel correspondant dans l'espace C^N.

Le Tableau 4.7.1 montre les coûts correspondants.

COUT DE BASE ET COUT PAR ITERATION : LE CAS i(k) = 0				
METHODES		DISCRETISATIONS		
		G, F, qG	P, qP	S, N, qS
COUT DE BASE		$\frac{n^3}{2} + 3N + \frac{3n^2}{2} - 3n$	$\frac{n^3}{2} + Nn + 3N + \frac{3n^2}{2} - 2n$	$\frac{n^3}{2} + Nn + \frac{3n^2}{2} + n$
COUT PAR ITERATION	(A1,0)	N^2+5N+n^2	$N^2+Nn+6N+n^2$	$N^2+Nn+2N+n^2+4n$
	(A1,0)bis	$N^2+5N+2n^2$	$N^2+2Nn+9N+n^2-2n$	$N^2+2Nn+2N+n^2+5n$
	(A2,0)	N^2+6N+n^2	$N^2+Nn+7N+n^2$	$N^2+Nn+3N+n^2+4n$
	(A2,0)bis	$N^2+4N+2n^2$	$N^2+2Nn+8N+n^2-2n$	$N^2+2Nn+N+n^2+5n$
	(B1,0)	$2N^2+6N+n^2$	$2N^2+Nn+7N+n^2$	$2N^2+Nn+3N+n^2+4n$
	(B2,0)	$2N^2+5N+n^2$	$2N^2+Nn+6N+n^2$	$2N^2+Nn+2N+n^2+4n$

TABLEAU 4.7.1.

c) Coût des méthodes générales.

Le calcul de ϕ^{k+1} à partir de ϕ^k avec la méthode (A1, i(k)) implique le calcul des vecteurs suivants

Le résidu : r^k (donc $T\phi^k$ et λ^{k+1})

Le vecteur $u := S_n r^k$

et i(k) fois

$v := Tu$

$y := \langle v, \phi_n^* \rangle \phi^k$

$w := \lambda^{k+1} u$ et

$x := u - S_n(v-w+y)$ qui prend la place de u pour recommencer ce cycle.

Si l'on utilise la méthode (A2, i(k)) on a à calculer

Le résidu $\tilde{r}^k$ (donc λ^{k+1} et $\frac{1}{\lambda^{k+1}} T\phi^k$)

Le vecteur $\tilde{u} := \tilde{S}_n \tilde{r}^k$

et i(k) fois

$\tilde{v} := T\tilde{u}$

$\tilde{y} := \langle \tilde{v}, \phi_n^* \rangle . \frac{1}{\lambda^{k+1}} T\phi^k$

$\tilde{w} := \frac{1}{\lambda^{k+1}} (\tilde{v}-\tilde{y})$ et

$\tilde{x} := \tilde{u} - \tilde{S}_n(\tilde{u}-\tilde{w})$ qui va substituer à $\tilde{u}$ pour recommencer ce cycle.

La méthode (B1,i(k)) nécessite pour sa part les calculs suivants :

Le résidu r^k

Les vecteurs $u' := T\, r^k$, $u'' := S_n u'$ et

$$u := \frac{1}{\lambda_n}(r^k - u'')$$

et i(k) fois :

$$v := Tu$$

$$y := \langle v, \phi_n^* \rangle \phi^k$$

$$w := \lambda^{k+1} u$$

$$x' := T(v-y-w), \quad x'' := S_n\, x' \quad \text{et}$$

$$x := \frac{1}{\lambda_n}(v-y-w-x'')$$

qui devient le nouvel u du cycle.

Finalement, pour ce qui est de la méthode (B2,i(k)) on a à calculer

Le résidu $\tilde{r}^k$

Les vecteurs $\tilde{u}' := T\, \tilde{r}^k$

$$\tilde{u} := \tilde{r}^k - S_n \tilde{u}'$$

et i(k) fois

$$\tilde{v} := T\tilde{u}$$

$$\tilde{y} := \langle \tilde{v}, \phi_n^* \rangle . \frac{1}{\lambda^{k+1}} T\phi^k$$

$$\tilde{w} := \frac{1}{\lambda^{k+1}}(\tilde{v}-\tilde{y})$$

$$\tilde{x}' := T(\tilde{u}-\tilde{w}) \quad \text{et} \quad \tilde{x} := \tilde{u} - S_n \tilde{x}'$$

qui prend la place de $\tilde{u}$ pour recommencer le cycle.
Voir Tableau 4.7.2.

COUT DES METHODES GENERALES				
METHODES		DISCRETISATIONS		
		G, F, qG	P, qP	S, N, qS
COUT DE BASE		$\frac{n^3}{2} + 3N + \frac{3n^2}{2} - 3n$	$\frac{n^3}{2} + Nn + 3N + \frac{3n^2}{2} - 2n$	$\frac{n^3}{2} + Nn + \frac{3n^2}{2} + n$
(A1,i(k))	1 fois	$N^2 + 5N + n^2$	$N^2 + Nn + 6N + n^2$	$N^2 + Nn + 2N + n^2 + 4n$
	+ i(k) fois	$N^2 + 6N + n^2$	$N^2 + Nn + 7N + n^2$	$N^2 + Nn + 3N + n^2 + 4n$
(A2,i(k))	1 fois	$N^2 + 6N + n^2$	$N^2 + Nn + 7N + n^2$	$N^2 + Nn + 3N + n^2 + 4n$
	+ i(k) fois	$N^2 + 7N + n^2$	$N^2 + Nn + 8N + n^2$	$N^2 + Nn + 4N + n^2 + 4n$
(B1,i(k))	1 fois	$2N^2 + 6N + n^2$	$2N^2 + Nn + 7N + n^2$	$2N^2 + Nn + 3N + n^2 + 4n$
	+ i(k) fois	$2N^2 + 7N + n^2$	$2N^2 + Nn + 8N + n^2$	$2N^2 + Nn + 4N + n^2 + 4n$
(B2,i(k))	1 fois	$2N^2 + 5N + n^2$	$2N^2 + Nn + 6N + n^2$	$2N^2 + Nn + 2N + n^2 + 4n$
	+ i(k) fois	$2N^2 + 6N + n^2$	$2N^2 + Nn + 7N + n^2$	$2N^2 + Nn + 3N + n^2 + 4n$

TABLEAU 4.7.2.

4.8 REMARQUES ET COMMENTAIRES BIBLIOGRAPHIQUES

Signalons premièrement que le problème de valeurs propres d'un opérateur intégral T peut être posé dans des espaces autres que C[0,1] :

Etant donné $p \in [1,+\infty[$ on définit *l'espace de Lebesgue* $L^p[0,1]$ comme l'espace des fonctions x définies sur [0,1] presque partout (p.p. dans la suite) telles que

$$\|x\|_p := [\int_0^1 |x(t)|^p \, dt]^{1/p} \text{ est fini.}$$

Deux fonctions x,y de $L^p[0,1]$ sont dites égales ($x = y$) si $\|x-y\|_p = 0$ (c'est-à-dire, $x(t) = y(t)$ p.p.).

L'espace $L^\infty[0,1]$ est défini comme l'espace des fonctions x définies p.p. sur [0,1] telles que

$$\|x\|_\infty := \text{Inf}\{c : |x(t)| \leq c \quad \text{p.p. sur } [0,1]\}$$

existe en tant que nombre réel. On remarque que C[0,1] est un sous-espace vectoriel de $L^\infty[0,1]$ et que la norme $\|\ \|$ que nous avons considérée sur C[0,1] est la norme induite par $\|\ \|_\infty$. Evidemment C[0,1] est fermé dans $L^\infty[0,1]$.

Quant à la compacité des opérateurs intégraux on rappelle le résultat suivant dû à Graham et Sloan 1979 :

Si le noyau κ défini p.p. sur $[0,1] \times [0,1]$ vérifie les conditions

i) $\quad \text{Sup}\{\|\kappa_t\|_p : t \in [0,1]\}$ est fini.

ii) $\quad \lim_{t' \to t} \|\kappa_{t'} - \kappa_t\|_p = 0 \qquad \forall t \in [0,1]$

où $\quad \kappa_t : [0,1] \to C$ est défini par

$$\kappa_t(s) = \kappa(t,s) \qquad \forall s \in [0,1]$$

et

$$1 \leq p \leq +\infty$$

alors l'opérateur intégral de Fredholm T dont le noyau est κ est un opérateur compact de $L^r[0,1]$ dans $C[0,1]$ pour tout r appartenant à l'intervalle $[q,+\infty]$ où q est défini par

$$\frac{1}{q} + \frac{1}{p} = 1$$

A fortiori, si i) et ii) sont vérifiés alors T est compact en tant qu'opérateur de C[0,1] dans lui-même.

Ceci permet de considérer des opérateurs à noyau non continu, comme par exemple

$$\kappa(t,s) = \frac{1}{|t-s|^{\alpha}} \qquad 0 \leq t \neq s \leq 1 \qquad 0 < \alpha < 1$$

Ce noyau appartient à une classe de noyaux dits *faiblement singuliers*. Pour les opérateurs à noyau faiblement singulier on peut construire les approximations de Galerkin, Projection et Sloan mais, par contre, les approximations de Fredholm et Nyström *ne sont pas toujours définies* dans ce cas. Pour un traitement numérique plus raffiné de ces noyaux d'autres techniques de quadrature approchée ont été développées ; notamment *l'intégration produit* (Cf. Atkinson 1976, Schneider 1979, Sloan 1980).

Si l'on pose le problème de valeurs propres de T dans l'espace de Hilbert $H = L^2[0,1]$ muni du produit scalaire habituel

$$\langle x,y \rangle := \int_0^1 x(t)\, \overline{y(t)}\, dt$$

et de la norme induite $\| \|_2$ alors les approximations T_n^G, T_n^P et T_n^S peuvent être définies à l'aide d'une *base hilbertienne othonormale* de H, comme par exemple, les *polynômes de Legendre :*

$$\forall t \in [0,1] : \qquad L_1(t) = 1$$

$$L_2(t) = \sqrt{3}\,(2t-1)$$

et pour $m \geq 3$

$$L_m(t) = \frac{(2m-1)^{1/2}}{m-1}\,(2m-3)^{1/2}\,(2t-1)L_{m-1}(t) - \frac{m-2}{(2m-5)^{1/2}}\,L_{m-2}(t)$$

(Cf. Vogel 1953). La suite de projections $\{\pi_n\}$ est alors définie par

$$\pi_n x = \sum_{i=1}^{n} \langle x, L_i \rangle L_i \qquad \forall x \in H$$

On a :

$$\pi_n \xrightarrow{p} 1$$

$$\langle x, y \rangle = 0 \qquad \forall x \in \text{Ker } \pi_n \qquad \forall y \in \pi_n H$$

donc

$$\pi_n^* = \pi_n$$

et

$$T_n^G \xrightarrow{\| \|_2} T, \quad T_n^P \xrightarrow{\| \|_2} T, \quad T_n^S \xrightarrow{\| \|_2} T.$$

La représentation matricielle de ces approximations est très simple car la base de $\pi_N H$ est obtenue en rajoutant $L_{n+1}, \ldots, L_N$ à la base $\{L_1, \ldots, L_n\}$ de $\pi_n H$ (Cf. Ahués et al 1982). Cependant, le calcul de matrice $\mathcal{G}_N^G$, de grande taille, est *très cher* à cause des évaluations des polynômes de Legendre de degré élevé.

De retour dans l'espace de Banach $X = C[0,1]$ disons qu'une quadrature approchée peut être considérée d'une manière générale comme une fonctionnelle

$$I^{(n)} : X \to C$$

définie par

$$I^{(n)} x = \sum_{i=1}^{n} \omega_i^{(n)} x(t_i^{(n)}) \qquad \forall x \in X$$

On veut, naturellement, que la suite $\{I^{(n)}\}$ soit ponctuellement convergente vers la fonctionnelle

$$I : X \to C$$

définie par

$$Ix = \int_0^1 x(t)dt \qquad \forall x \in X$$

Or, d'après le théorème de Banach-Steinhaus -ou de la borne uniforme- (Cf. Lang 1977, p. 199), pour cela il faut et il suffit que

i) $$\text{Sup} \{ \sum_{i=1}^{n} |\omega_i^{(n)}| : n \in \mathbb{N} \} \quad \text{soit fini}$$

ii) Il existe un sous-ensemble W de X, dense dans X, tel que

$$\lim_{n\to\infty} I^{(n)}x = Ix \quad \forall x \in W.$$

Très souvent, W est pris égal à l'espace des polynômes définis sur [0,1], ou égal à l'espace des fonctions linéaires par morceaux, ainsi qu'on l'a fait dans ce travail.

En ce qui concerne le calcul des opérateurs résolvante, projection spectrale et résolvante réduite d'une matrice carrée de taille N correspondants à une approximation T_n, à partir de ceux d'une matrice carrée de taille $n \leq N$, une formulation beaucoup plus générale est donnée dans Chatelin 1983 (pp. 183-189) où l'on ne tient pas compte du caractère matriciel de ces opérateurs. Cependant nous pensons que nos démonstrations, bien que valables sous des hypothèses plus restrictives, s'adaptent mieux à notre besoin : la programmation des calculs numériques sur l'ordinateur.

A propos du calcul de la résolvante réduite, reprenons les notations de la section 4.6. Le calcul de $y_n = \mathcal{S}_n x_n$ nécessite de la résolution du système linéaire de rang n, à n+1 équations :

$$(\mathcal{C}_n - \lambda_n I_n) y_n = (I_n - \mathcal{P}_n) x_n$$

$$v_n^H y_n = 0$$

dont les inconnues sont les n coordonnées de y_n.

Ce système peut être numériquement mal conditionné lorsqu'il existe des valeurs propres de $\mathcal{C}_n$ autres que λ_n qui sont, pourtant, proches de celle-ci. Le coût de la factorisation $\mathcal{U}, \mathcal{L}^{-1}$ peut être réduit à $\frac{n^3}{3} + \frac{n^2}{2} - \frac{n}{6}$ produits si l'on applique le schéma de Crout. Ceci est fait dans d'Almeida 1983, où l'on trouve aussi une étude de la stabilité de ce système linéaire.

Considérons maintenant le problème du coût des itérations. Pour les méthodes les plus simples ($i(k) = 0$) il est intéressant de comparer d'une part (A1,0) et (A1,0) bis à (B1,0) et d'autre part (A2,0) et (A2,0) bis à (B2,0). D'une manière générale on constate (Cf. Tableau 4.7.1) que les méthodes (B1,0) et (B2,0) ont un coût par itération égal à $2N^2 + \ldots$ et que (A1,0) et (A2,0) ont un coût égal à $N^2 + \ldots$ Ainsi, les premières seront moins chères si elles nécessitent un nombre d'itérations moitié moindre (à précision désirée égale).

Notons ℓ le nombre d'itérations à faire pour obtenir une précision donnée, mesurée par exemple par le critère suivant :

ℓ est le plus petit entier tel que

$$\text{Max}\{\|r^{\ell}\|, \|\tilde{r}^{\ell}\|\} < \varepsilon_0$$

où ε_0 est un seuil fixé auparavant.

Prenons le cas suivant :

$$N = 100 \qquad n = 10 \qquad i(k) = 0 \qquad (4.8.1.)$$

Alors, du Tableau 4.7.1. on déduit le Tableau 4.8.1.

CONDITIONS POUR QUE (B1,0) ou (B2,0) SOIENT MOINS CHERES QUE (A1,0), (A1,0) bis, (A2,0), ou (A2,0) bis			
CONDITION	DISCRETISATION : N = 100, n = 10		
	G, F, qG	P, qP	S, N, qS
$\frac{\ell(B1,0)}{\ell(A1,0)} <$	0.51	0.54	0.53
$\frac{\ell(B1,0)}{\ell(A1,0)bis} <$	0.52	0.60	0.53
$\frac{\ell(B2,0)}{\ell(A2,0)} <$	0.52	0.54	0.54
$\frac{\ell(B2,0)}{\ell(A2,0)bis} <$	0.51	0.60	0.57

TABLEAU 4.8.1.

Il est évident que pour ce qui est des familles générales à paramètre i(k) la situation n'est guère différente.

Approximativement, si l'on prend $i(k) = i_0$, constant et égal pour les quatre familles, alors on reproduit le Tableau 4.8.1. Cependant la situation de B1 et B2 par rapport à A1 et A2 varie quelque peu si i(k) n'est pas constant.

A titre d'exemple, prenons le cas

$$n = 100 \qquad n = 10 \qquad i(k) = k \qquad (4.8.2.)$$

Si $D \in \{G,F,qG\}$ alors on trouve que la méthode (B1,k) est moins chère que la méthode (A1,k) si les nombres d'itérations respectifs $\ell(B1,k)$ et $\ell(A1,k)$ satisfont, approximativement, à l'inégalité

$$20800.\ell(B1,k) < 10^4(2.2\ell(A1,k)^2+6.6\ell(A1,k)+9.7)^{1/2}-31100$$

et il en est de même pour (B2,k) par rapport à (A2,k).
Dans le cas des autres discrétisations ces comparaisons donnent des seuils un peu plus grands pour $\ell(B1,k)$ (et $\ell(B2,k)$) pour que ces méthodes soient moins chères que les autres. Les résultats sont présentés sur le Tableau 4.8.2.

CONDITIONS POUR QUE (B1,k) ou (B2,k) SOIENT MOINS CHERES QUE (A1,k) ou (A2,k)			
CONDITION	DISCRETISATION N = 100 n = 10		
si $\ell(A1,k) = \ell(A2,k) =$	G, F, qG	P, qP	S, N, qS
	alors $\ell(B1,k) = \ell(B2,k) <$		
10	6	7	6
50	35	36	36
100	71	73	72
500	358	368	364
1000	717	738	728

TABLEAU 4.8.2.

Finalement on considère le cas

$$N = 100 \qquad n = 10 \qquad i(k) = 2^{k+1} \quad (4.8.3.)$$

On constate que dans ce cas il suffit que $(B1, 2^{k+1})$ ou $(B2,2^{k+1})$ fassent moins d'itérations que $(A1,2^{k+1})$ ou $(A2,2^{k+1})$ pour qu'elles soient moins chères que ces dernières.

Jusque là on a considéré que i(k) est la même fonction de k pour les quatre familles. Or, si les itérations δ_j^k, $j \in \mathbb{N}$ sont arrêtées par une condition de précision atteinte alors il se peut, évidemment que i(k) varie d'une famille à l'autre et que ceci, comptes faits du coût total des calculs, soit favorable à l'utilisation des méthodes $(B1, i(k))$ ou $(B2,i(k))$.

C'est donc maintenant que l'expérimentation numérique est indispensable.

Disons finalement que les méthodes proposées dans cette thèse sont beaucoup moins chères que les algorithmes généraux qu'on utilise pour la résolution numérique du problème de valeurs propres de grandes matrices. Prenons par exemple la méthode d'Arnoldi en version itérative (Cf. Saad 1979, 1980, 1983) :

Soit A une matrice complexe carrée de taille N. On fixe $n << N$ et on se donne x dans C^N de façon à ce que les vecteurs x, Ax, ..., $A^{n-1}x$ soient linéairement indépendants. On note X_n le sous-espace de C^N engendré par ces vecteurs. Soit

$$\pi_n : C^N \to C^N$$

la projection orthogonale sur X_n et soit

$$\hat{A}_n := \pi_n A|_{X_n}$$

$\hat{A}_n$ est une application linéaire de X_n dans lui-même et admet une représentation matricielle carrée de taille n par rapport à une base donnée dans X_n.

La *méthode d'Arnoldi* construit une base orthonormale de X_n par rapport à laquelle $\hat{A}_n$ est représenté par une matrice de Hessenberg $H_n = (h_{ij})$ de taille n. Cette base, notée $\{v_1,\ldots,v_n\}$, et cette matrice sont définies ainsi

$$v_1 := \frac{x}{\|x\|_2} \qquad h_{11} := v_1^H A\, v_1 \qquad h_{21} := \|x\|_2$$

Pour $j = 1,\ldots,n-1$

$$x_{j+1} := A\, v_j - \sum_{i=1}^{j} h_{ij}\, v_i$$

$$h_{j+1,j} := \|x_{j+1}\|_2 \quad , \text{(que l'on suppose non nul)}$$

$$v_{j+1} := \frac{x_{j+1}}{\|x_{j+1}\|_2}$$

$$h_{i,j+1} := v_i^H A\, v_{j+1} \quad \text{pour } i = 1,2,\ldots,j+1$$

Soit V_n la matrice de taille $N \times n$ dont les colonnes sont $v_1,\ldots,v_n$. Si μ est une valeur propre de H_n et ζ un vecteur propre associé à μ alors $\phi = V_n\zeta$ est un vecteur propre de $\pi_n A\, \pi_n$ associé à μ qui est aussi une valeur propre de cette application. On dit que μ,ϕ sont des *éléments propres approchés* de A (obtenus par la méthode d'Arnoldi).

La version itérative de cette méthode consiste à *recommencer le cycle de calculs en prenant* $x = \phi$ comme point de départ.

Le calcul de μ,ζ se fait couramment par la méthode QR (Cf. Watkins 1982). Notons $C(\mu,\zeta)$ le coût de ce calcul. Alors chaque itération de la méthode d'Arnoldi a un coût approximativement égal à

$$C(1 \text{ itération Arnoldi}) \approx (n-1)N^2+n^2N-nN+5N+C(\mu,\zeta)$$

(où on a négligé quelques calculs de racines carrées...).

Il est très important de remarquer, encore une fois, que ce qui rend nos méthodes plus performantes que les méthodes générales (de type itérations simultanées, QR, Arnoldi, Lanczos, etc.) est la *relation intime qui existe, via l'opérateur* T, *entre la grande matrice* $\mathcal{Z}_N$

et la petite matrice $\mathcal{C}_n$ sur laquelle on fait les calculs.

Des techniques similaires ont été développées dans le but d'accélérer la convergence de la méthode de la puissance itérée (et de la méthode des itérations simultanées) aussi bien dans le cas des opérateurs intégraux (Cf. Chu et Spence 1981) que dans le cas général où la grande et la petite matrice ne proviennent pas de la discrétisation d'un opérateur (Cf. Chatelin et Miranker 1982, 1983).

CHAPITRE 5

SUR LES EXPERIENCES NUMERIQUES

Exemples et Conclusions

INTRODUCTION

Dans ce chapitre on montre les résultats de quelques expériences numériques. Nous avons fait un grand nombre d'essais avec différents noyaux et différentes discrétisations. La plupart de ces essais ont été consacrés à l'étude des méthodes les plus simples.

Dès les premières expériences on s'est aperçu d'un fait qui fut finalement confirmé par l'ensemble des situations considérées : la méthode (B2,0) *est plus rapide* que les autres. Malgré son coût élevé, elle est, en fin de comptes, plus avantageuse que les autres.

Les méthodes générales ont été testées dans le but de montrer la possibilité d'avoir, en pratique, une convergence superlinéaire (ou même quadratique) dès les premières itérations. Il en est ainsi ; pourtant, leur coût très élevé ne les rend pas plus performantes que les méthodes les plus simples (qui sont à convergence linéaire). On revient de nouveau sur l'appréciation antérieure : la méthode (B2,0) s'avère le meilleur choix "presque partout".

La longueur de cette thèse étant forcément limitée, nous avons choisi un échantillon représentatif d'exemples numériques.

Je remercie Mauricio Télias pour son inestimable collaboration en ce qui concerne la programmation, la mise en oeuvre et la validation des méthodes testées sur l'ordinateur.

5.1 GENERALITES

Dans tous les cas les paramètres de discrétisation n et N sont fixés à

$$n = 10 \quad , \quad N = 100$$

Les éléments propres de la matrice $\mathcal{C}_n^D$, où $D \in \{F,G,qG\}$, sont calculés par la méthode QR en utilisant un sous-programme de la bibliothèque NAG.

L'indice ℓ du dernier itéré est défini par

$$\ell := \text{Min } \{k \in \mathbb{N} : ||\text{Res}(k)|| \leq 5.0\text{E-}10\}$$

où

$$\text{Res}(k) = \begin{cases} r^k & \text{pour A1 et B1} \\ \tilde{r}^k & \text{pour A2 et B2} \end{cases}$$

$$a.b\text{E-}q := (a + b\times 10^{-1}) \times 10^{-q}$$

et $\|\ \|$ est la norme induite sur X_N par la norme du max de l'espace $X = C[0,1]$, c'est-à-dire

$$\|x\| = \text{Max}\{|x(t_i^{(N)})| \ : 1 \leq i \leq N\} \quad \forall x \in X_N.$$

A titre d'illustration, pour apprécier l'effet du raffinement de la valeur propre λ_n, on donne dans chaque cas les différences $\lambda_n - \lambda$ et $\lambda^{\ell+1} - \lambda$ où λ est la valeur propre exacte. Font exception à ceci deux exemples où l'on ne connait pas la valeur exacte.

Toutes les expériences ont été faites sur l'ordinateur CII-Honeywell-Bull 68 du Centre Interuniversitaire de Calcul de Grenoble (C.I.C.G.).

5.2 LE NOYAU DE CONVECTION - DIFFUSION

Le problème de convection-diffusion unidimensionnel à l'état stationnaire et à vitesse constante s'écrit

$$-\varepsilon Z'' + uZ' = f \qquad \text{dans }]0,1[$$

$$Z(0) = Z(1) = 0$$

(Cf. Roache 1972, Telias 1983).

ε est une constante positive qui mesure la diffusivité et u représente la vitesse que l'on suppose constante.

La solution Z admet la représentation intégrale

$$Z(t) = \int_0^1 \kappa(t,s)\, f(s)ds \qquad t \in [0,1] \qquad (5.2.1.)$$

(Cf. Courant et Hilbert 1953, pp. 351-375, Chatelin 1983 pp. 90-92) où κ est le noyau de Green suivant

$$\kappa(t,s) = \begin{cases} \alpha(e^{u(1-s)/\varepsilon}-1)\,(1-e^{ut/\varepsilon}) & \text{si } 0 \le t \le s \le 1 \\ \alpha(e^{u(1-s)/\varepsilon}-e^{u/\varepsilon})\,(1-e^{u(1-t)/\varepsilon}) & \text{si } 0 \le s \le t \le 1 \end{cases}$$

La constante α est définie par :

$$\alpha = \frac{1}{u(1-e^{u/\varepsilon})}$$

Il est bien connu que $\frac{\partial \kappa(t,s)}{\partial t}$ n'est pas défini sur la diagonale $0 \le s = t \le 1$, et que

$$\left| \lim_{t \to s^+} \frac{\partial \kappa(t,s)}{\partial t} - \lim_{t \to s^-} \frac{\partial \kappa(t,s)}{\partial t} \right| = \frac{1}{\varepsilon}$$

Les éléments propres de l'opérateur intégral $T : f \to Z$ défini dans $C[0,1]$ par (5.2.1) sont les suivants :

Pour chaque $j \in \mathbf{N} \setminus \{0\}$,

$$\overset{(j)}{\phi}(t) = e^{ut/(2\varepsilon)} \sin j\pi t, \qquad t \in [0,1]$$

est une fonction propre associée à la valeur propre simple

$$\overset{(j)}{\lambda} = \frac{4\varepsilon}{(2\pi j\varepsilon)^2+u^2}$$

Du point de vue numérique il est intéressant de tester le cas où $\varepsilon << u$ (Cf. Telias 1983).

5.3 LE NOYAU DE DIRICHLET DANS UNE ELLIPSE

Le problème de Dirichlet

$$\frac{\partial^2 Z}{\partial x^2} + \frac{\partial^2 Z}{\partial y^2} = 0 \text{ dans } \Omega$$

$$Z = f \text{ sur } \partial\Omega \qquad (5.3.1.)$$

où Ω est le domaine elliptique défini par

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} < 1 \qquad (a > b > 0)$$

et $\partial\Omega$ sa frontière, nous amène à considérer l'équation intégrale

$$u(t) + \frac{1}{\pi} \int_0^1 \frac{\partial w(x(t),y(t),s)}{\partial s} u(s)ds = \frac{1}{\pi} g(t) \qquad t \in [0,1] \quad (5.3.2).$$

Ici, $x(t) = a \cos 2\pi t$, $y(t) = b \sin 2\pi t$, $t \in [0,1]$ est une paramétrisation de $\partial\Omega$ et

$$w(x,y,s) := \text{Arctg} \frac{y(s)-y}{x(s)-x} \qquad s \in [0,1]$$

$$g(t) := f(x(t),y(t)) \qquad t \in [0,1]$$

Si u est une solution de l'équation (5.3.2) alors

$$Z(x,y) = \int_0^1 \frac{\partial w(x,y,s)}{\partial s} u(s)ds$$

est une solution du problème (5.3.1) (Cf. Kantorovich et Krylov 1955, p. 119).

Un calcul assez long montre que

$$\frac{\partial w(x(t),\, y(t),s)}{\partial s} = \frac{\frac{2\pi ab}{a^2-b^2}}{\frac{a^2+b^2}{a^2-b^2} - \cos 2\pi(s+t)}$$

Si l'on pose

$$\gamma = \frac{a-b}{a+b}$$

alors (5.3.2.) s'écrit

$$u(t) + \int_0^1 \kappa(t,s)u(s)ds = \frac{1}{\pi} g(t) \qquad t \in [0,1]$$

où

$$\kappa(t,s) = \frac{1-\gamma^2}{1+\gamma^2 - 2\gamma\cos 2\pi(s+t)} \qquad 0 \le t,\ s \le 1$$

L'opérateur intégral T défini par ce noyau a les éléments propres suivants :

Pour chaque $j \in \mathbb{N}$

$$\phi_c^{(j)}(t) = \cos 2\pi jt \qquad t \in [0,1]$$

est une fonction propre associée à la valeur propre simple

$$\lambda_+^{(j)} = +\gamma^j$$

et pour chaque $j \in \mathbb{N}\setminus\{0\}$

$$\phi_s^{(j)}(t) = \sin 2\pi jt \qquad t \in [0,1]$$

est une fonction propre associée à la valeur propre simple

$$\lambda_-^{(j)} = -\gamma^j$$

Le cas où a >> b, c'est-à-dire que γ est proche de 1, s'avère intéressant car

$$\frac{\text{Max}\{|\kappa(t,s)| \ : \ 0 \le t,\ s \le 1\}}{\text{Min}\{|\kappa(t,s)| \ : \ 0 \le t,\ s \le 1\}} = \left(\frac{1+\gamma}{1-\gamma}\right)^2$$

5.4 LES METHODES LES PLUS SIMPLES : $i(k) = 0 \quad \forall k \in \mathbb{N}$

On présente douze exemples concernant le cas

$$i(k) = 0 \qquad \forall k \in \mathbb{N}.$$

Le Tableau 5.4.1. montre les tests effectués avec le noyau de convection-diffusion et le Tableau 5.4.2. montre les exemples traités avec le noyau de Dirichlet. On remarque dans la plupart de ces cas la supériorité de (B2,0). Elle converge si rapidement que le coût total la rend plus avantageuse que les autres méthodes, malgré son coût par itération élevé, presque dans tous les cas. Seuls certains cas concernant les discrétisations par projection P et qP font exception à ces remarques. Dans le cas de l'exemple 3 l'avantage de (B1,0) par rapport à (B2,0) n'est pas significatif en termes relatifs.

La convergence des méthodes les plus simples est *linéaire*. En pratique, on constate que les constantes des bornes d'erreur sont assez modérées. La figure 5.4.1. montre la convergence de la méthode (B2,0) dans le cas de l'exemple 5.4.10 du Tableau 5.4.2.

Noyau de convection-diffusion								ℓ = Nombre d'itérations			
Exemple	u	ε	$\lambda = \lambda^{(j)}$	(j)	D	$\lambda_n - \lambda$	$\lambda^{\ell+1} - \lambda$	A1	A2	B1	B2
5.4.1.	0.	1.00	0.025330	2	S	-1.0E-03	-8.3E-06	6	6	4	4
5.4.2.	0.	1.00	0.0063326	4	P	-9.2E-04	-8.5E-06	8	8	9	5
5.4.3.	-1.	1.00	0.098818	1	F	1.0E-03	8.0E-06	32	32	19	23
5.4.4.	1.	0.05	0.057680	5	N	2.3E-02	1.7E-04	22	14	15	10
5.4.5.	1.	0.05	0.105917	3	G	-1.5E-03	1.0E-06	3	4	4	2
5.4.6.	1.	0.05	0.105917	3	P	-1.5E-03	1.0E-06	16	17	15	16

TABLEAU 5.4.1.

Noyau de Dirichlet							ℓ = Nombre d'itérations			
Exemple	γ	$\lambda = -\gamma^j$	j	D	$\lambda_n - \lambda$	$\lambda^{\ell+1} - \lambda$	A1	A2	B1	B2
5.4.7.	0.25	-0.250	1	qP	4.8E-01	1.0E-12	6	2	6	2
5.4.8.	0.50	-0.250	2	qS	3.8E-02	1.0E-12	8	6	9	4
5.4.9.	0.70	-0.343	3	N	-1.4E-01	1.0E-08	32	20	20	10
5.4.10.	0.70	-0.343	3	qP	8.7E-02	1.0E-09	17	19	19	5
5.4.11.	0.85	-0.850	1	qG	2.7E-02	-2.0E-07	15	16	7	8
5.4.12	0.85	-0.850	1	qS	2.7E-02	-2.0E-07	14	15	7	7

TABLEAU 5.4.2.

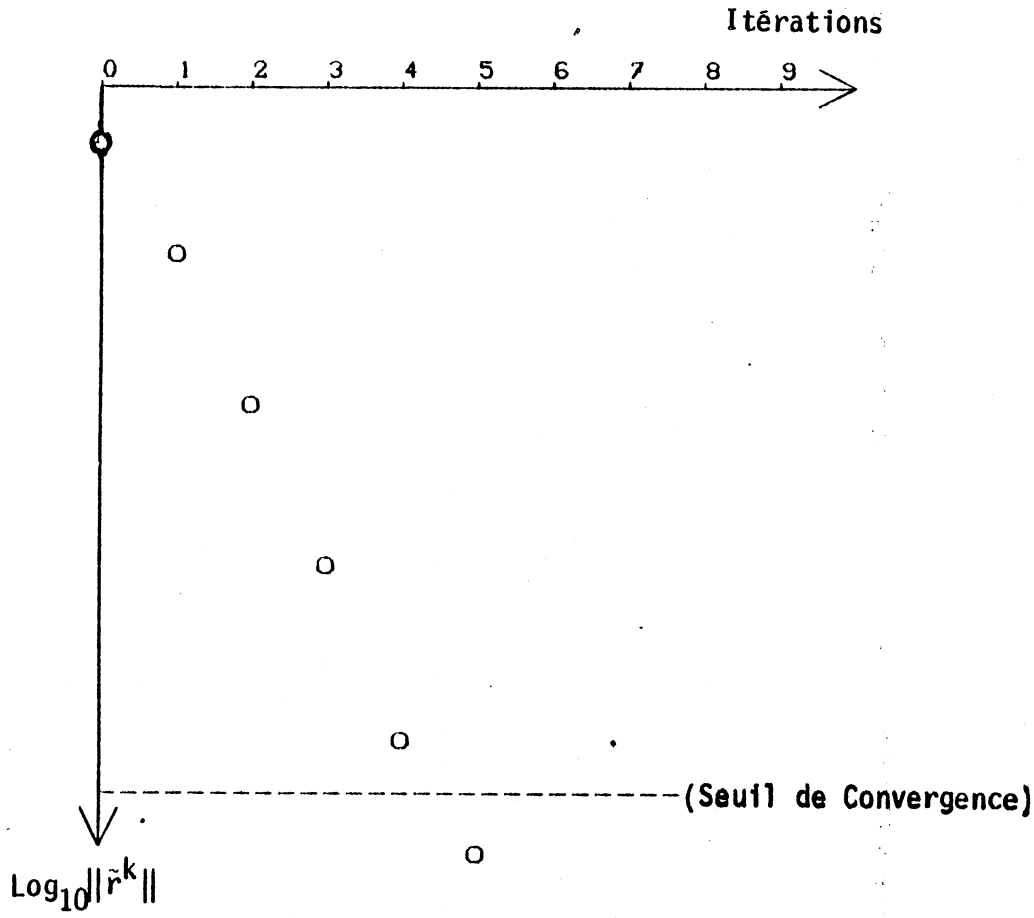

Itérations
0
1
2
3
4
5
6
7
8
9
(Seuil de Convergence)
$Log_{10}\|\tilde{r}^k\|$

FIGURE 5.4.1

5.5 LES METHODES GENERALES

On présente quatre exemples qui montrent la possibilité d'avoir une convergence *superlinéaire* (et parfois *quadratique)* avec un choix approprié de i(k).

On note

$$\nabla\delta_i^k := \delta_i^k - \delta_{i-1}^k$$

$$\eta_k := \text{borne supérieure de } \|Res(k)\|$$

Exemple 5.5.1

Noyau de convection-diffusion.
Les paramètres u, ε sont fixés à

$$u = 0 \quad \varepsilon = 1.00$$

La valeur propre à approcher est

$$\lambda = \overset{(4)}{\lambda} = 0.00633257 \text{ (valeur tronquée)}$$

On définit pour cet exemple

$$i(k) := \text{Min}\{i \in \mathbb{N} : \|\nabla\delta_i^k\| \leq 0.01\|Res(k)\| \}$$

Le Tableau 5.5.1 montre les résultats des quatre méthodes en utilisant la discrétisation $\mathcal{C}_n^p$. Cet exemple doit être comparé avec l'exemple 5.4.2 du Tableau 5.4.1.

□

Exemple 5.5.2

Noyau de convection-diffusion
Les paramètres u, ε sont fixés à

$$u = 1 \quad \varepsilon = 0.05$$

La valeur propre exacte est

$$\lambda = \overset{(5)}{\lambda} = 0.057680 \text{ (valeur tronquée)}$$

On définit dans ce cas

$$\ell_k := \text{Min}\{i \in \mathbb{N} : \|\nabla\delta_i^k\| \leq 0.02 \|\text{Res}(k)\| \}$$

$$i(k) := \text{Min}\{\ell_k, k+2\}.$$

Le Tableau 5.5.2 montre les résultats des quatre méthodes en utilisant la discrétisation $\mathcal{Z}_n^N$. On doit comparer avec l'exemple 5.4.4. du Tableau 5.4.1.

□

Exemple 5.5.3

Noyau de Dirichlet

Le paramètre γ est fixé à

$$\gamma = 0.25$$

La valeur propre à approcher est

$$\lambda = -0.25$$

On définit dans cet exemple

$$i(k) := \text{Min}\{i \in \mathbb{N} : \|\nabla\delta_i^k\| \leq 0.02 \|\text{Res}(k)\|\}$$

Le Tableau 5.5.4 montre les résultats des quatre méthodes en utilisant la discrétisation $\mathcal{Z}_n^{qP}$. Voir Exemple 5.4.7 du Tableau 5.4.2.

□

Exemple 5.5.4

Noyau de Dirichlet

On fixe le paramètre γ à

$$\gamma = 0.50$$

La valeur propre exacte que l'on veut approcher est

$$\lambda = -0.25$$

Noyau de convection-diffusion				
$u = 0.$ $\varepsilon = 1.00$ $D = P$ $\lambda = {}^{(4)}\lambda$				
Méthode	k	η_k	η_k/η_{k-1}	i(k)
A1	0	8.7E-04	-	5
	1	7.1E-06	8.2E-03	5
	2	5.2E-10	7.3E-05	5
	3	4.0E-15	7.7E-06	-
A2	0	1.4E-01	-	2
	1	1.7E-04	1.2E-03	2
	2	5.0E-08	2.9E-04	2
	3	2.5E-12	5.0E-05	-
B1	0	8.7E-04	-	6
	1	7.1E-06	8.2E-03	6
	2	4.2E-10	5.9E-05	-
B2	0	1.4E-01	-	2
	1	1.5E-04	1.1E-03	2
	2	1.7E-10	1.1E-06	-
$\lambda_n - {}^{(4)}\lambda$	-9.2E-04			
$\lambda^{\ell+1} - {}^{(4)}\lambda$	-8.5E-06			

TABLEAU 5.5.1

Noyau de convection-diffusion

$u = 1.$ $\varepsilon = 0.05$ $D = N$ $\lambda = \lambda^{(5)}$

Méthode	k	η_k	η_k/η_{k-1}	i(k)	(*)
A1	0	2.2E-02	-	2	N.
	1	8.3E-04	3.8E-02	3	N.
	2	2.3E-05	2.8E-02	4	N.
	3	3.8E-07	1.7E-02	5	N.
	4	2.4E-09	6.4E-03	6	N.
	5	6.3E-12	2.7E-03	-	-
A2	0	5.0E-01	-	2	N.
	1	4.0E-02	8.0E-02	3	0.
	2	1.6E-04	4.0E-03	3	0.
	3	2.5E-07	1.7E-03	4	0.
	4	3.8E-10	1.6E-02	-	-
B1	0	2.2E-02	-	2	N.
	1	5.8E-03	2.7E-01	3	N.
	2	2.8E-04	4.9E-02	4	N.
	3	1.0E-06	3.6E-03	5	N.
	4	5.9E-10	5.9E-04	6	0.
	5	2.4E-13	4.1E-05		-
B2	0	5.0E-01	-	2	N.
	1	5.0E-02	1.0E-01	2	0.
	2	1.5E-04	3.0E-03	2	0.
	3	4.8E-08	3.2E-04	3	0.
	4	1.2E-11	2.5E-04	-	-

$\lambda_n - \lambda^{(5)}$	2.3E-02
$\lambda^{\ell+1} - \lambda^{(5)}$	1.7E-04

(*) Est-ce que la condition $\|\nabla \delta^k_{i(k)}\| \leq 0.02 \|Res(k)\|$ est vérifiée ?
N = Non, 0 = oui.

TABLEAU 5.5.2

Noyau de Dirichlet				
$\gamma = 0.25$ $\quad D = qP \quad$ $\lambda = -\gamma$				
Méthode	k	η_k	η_k/η_{k-1}	i(k)
A1	0	9.9E-03	-	2
	1	7.1E-07	7.2E-05	2
	2	5.0E-11	7.0E-05	-
A2	0	4.0E-02	-	1
	1	5.8E-11	1.5E-09	-
B1	0	9.9E-03	-	2
	1	7.1E-07	7.2E-05	2
	2	5.0E-11	7.0E-05	-
B2	0	4.0E-02	-	1
	1	8.0E-13	2.0E-11	-
$\lambda_n - \lambda$		4.8E-01		
$\lambda^{\ell+1} - \lambda$		1.0E-12		

TABLEAU 5.5.3

Noyau de Dirichlet					
$\gamma = 0.50$ $\quad D = qS \quad \lambda = -\gamma^2$					
Méthode	k	η_k	η_k/η_{k-1}	i(k)	(*)
A1	0	1.3E-03	-	2	N.
	1	6.4E-06	4.9E-03	3	0.
	2	4.0E-09	6.3E-04	3	0.
	3	1.1E-12	2.8E-04	-	-
A2	0	5.1E-03	-	2	0.
	1	2.6E-07	5.1E-05	2	0.
	2	7.9E-12	3.0E-05	-	-
B1	0	1.3E-03	-	2	N.
	1	1.1E-05	8.5E-03	3	N.
	2	1.8E-08	1.2E-03	4	0.
	3	2.5E-12	1.9E-04	-	-
B2	0	5.1E-03	-	1	0.
	1	8.7E-07	1.7E-04	1	0.
	2	8.7E-12	9.9E-06	-	-
$\lambda_n - \lambda$	3.8E-02				
$\lambda^{\ell+1} - \lambda$	1.0E-12				
(*) Est-ce que la condition $\|\nabla\delta^k_{i(k)}\| \leq 0.02\|Res(k)\|$ est vérifiée ? N = Non, 0 = oui.					

TABLEAU 5.5.4

On définit i(k) comme dans l'exemple 5.5.2. Le Tableau 5.5.4 montre les résultats avec la discrétisation $\mathcal{Z}_n^{qS}$ (voir aussi Tableau 5.4.2).

□

On voit bien que les convergences superlinéaire et quadratique sont possibles en pratique avec une fonction i(k) "raisonnable".

Néanmoins, compte tenu du coût total, il semble plus avantageux d'utiliser la méthode simple (B2,0). D'autre part, il semble que la formulation (3.4.1.) est meilleure que (3.1.3.), c'est-à-dire que le choix est entre A2 ou B2.

5.6 L'EQUATION DE 2EME ESPECE : ATKINSON/BRAKHAGE

Rappelons que la méthode A2 consiste à résoudre l'équation dans $(1-P_n)X$:

$$\tilde{J}(\phi^k)\delta^k = \tilde{F}(\phi^k)$$

en utilisant un *inverse approché de type Atkinson :*

$$-\lambda_n S_n\big|_{(1-P_n)X} : (1-P_n)X \to (1-P_n)X$$

tandis que la méthode B2 le fait en utilisant un *inverse approché de type Brakhage :*

$$(1-S_nT)\big|_{(1-P_n)X} : (1-P_n)X \to (1-P_n)X$$

On se pose donc, tout naturellement, la question suivante :

Est-ce que pour les équations intégrales de 2ème espèce (Cf. Exemples 2.3.1 et 2.3.2) *la méthode de Brakhage est meilleure que celle d'Atkinson ?*

Il semble que la réponse est *oui*. Effectivement, dans les exemples qui suivent on constate la supériorité de la méthode de Brakhage face à la méthode d'Atkinson pour la résolution de

$$(T-z)x = f.$$

Dans cette section on note

$$Res(k) = (T-z)u^k - f$$

Le dernier itéré u^ℓ est caractérisé par l'indice

$$\ell = Min\{k \in \mathbb{N} : \|Res(k)\| \leq 5.0E\text{-}10\}$$

Exemple 5.6.1 Noyau de convection-diffusion

Les valeurs des paramètres u, ε et z sont

$$u = 1 \quad \varepsilon = 0.05 \qquad z = 2.0$$

Le second membre est défini par

$$f(t) = -t^2+5t-2.2 \qquad t \in [0,1]$$

et on considère la discrétisation $\mathcal{G}_n^P$ ($n = 10$). Les évaluations de T sont faites à l'aide de $\mathcal{G}_N^G$ ($N = 100$).

On obtient :

$$\ell(\text{Atkinson}) = 16 \text{ itérations}$$
$$\ell(\text{Brakhage}) = 7 \text{ itérations}$$

La solution exacte est

$$u(t) = 1.1-2t \qquad t \in [0,1]$$

et on a l'estimation a posteriori :

$$\text{Max}\{|u(t_i^{(N)})-u^\ell(t_i^{(N)})| : 1 \leq i \leq N\} \leq 1.0E\text{-}10$$

pour les deux méthodes.

□

Exemple 5.6.2 Noyau de Dirichlet

Les paramètres γ et z sont fixés à

$$\gamma = 0.50 \qquad z = 4.0$$

Le second membre est donné par

$$f(t) = 3.0 \qquad t \in [0,1]$$

On utilise la discrétisation $\mathcal{Z}_n^N$ (n = 10) et les évaluations de T sont faites avec $\mathcal{Z}_N^F$ (N = 100). Les résultats sont :

$$\ell(\text{Atkinson}) = 9 \text{ itérations}$$
$$\ell(\text{Brakhage}) = 6 \text{ itérations}$$

La solution exacte est

$$u(t) = -1.0 \qquad t \in [0,1]$$

et pour les deux méthodes on a l'estimation a posteriori

$$\textbf{Max}\ \{|u(t_i^{(N)}) - u^{\ell}(t_i^{(N)})| \ : 1 \le i \le N\} \le 1.5E\text{-}10 \ .$$

□

5.7 DES NOYAUX "PAS COMME LES AUTRES"

Jusqu'ici nous avons considéré des noyaux continus et dont les dérivées partielles premières existent et sont bornées presque partout sur $[0,1] \times [0,1]$.

Considérons maintenant le noyau

$$\kappa(t,s) = 2|\sin 10\pi s - \sin 10\pi t|^{1/2} \quad 0 \le t,s \le 1 \qquad (5.7.1)$$

dont *les dérivées partielles premières ne sont pas bornées* autour des ensembles :

$$U_\nu = \{(t,s) \in [0,1] \times [0,1] : |s-(-1)^\nu t| = \frac{\nu}{10}\} \quad \nu = 0,1,\ldots,9$$

et sur lesquels *elles ne sont pas définies.*

Exemple 5.7.1

On cherche à approcher la valeur propre dominante de la discrétisation $\mathcal{C}_N^F$ (N = 100) de l'opérateur T défini par le noyau (5.7.1.). On utilise pour cela la discrétisation $\mathcal{C}_n^F$ (n = 10) et on obtient avec les méthodes les plus simples les nombres d'itérations suivants :

ℓ(A1) = 11 itérations
ℓ(A2) = 10 itérations
ℓ(B1) = 6 itérations
ℓ(B2) = 5 itérations.

□

Certains noyaux *non continus* -dits faiblement singuliers- définissent, eux aussi, des opérateurs intégraux compacts de C[0,1] dans lui-même (Cf. Section 4.8.). C'est le cas, par exemple (Cf. Phillips 1972) du noyau

$$\kappa(t,s) = |t-s|^{-1/2} \quad 0 \le t,s \le 1 \quad t \neq s. \qquad (5.7.2.)$$

Exemple 5.7.2

On cherche à approcher la valeur propre sous-dominante de la discrétisation $\mathcal{C}_N^G$ (N = 100) de l'opérateur intégral T défini par le noyau (5.7.2.). On utilise les méthodes les plus simples avec la discrétisation $\mathcal{C}_n^S$ (n = 10) et on trouve que la méthode A1 converge très lentement car

$$0.995 \le \frac{\|Res(k)\|}{\|Res(k-1)\|} < 1.0 \qquad 1 \le k \le 100$$

Pour les autres méthodes on obtient

ℓ(A2) = 18 itérations
ℓ(B1) = 10 itérations
ℓ(B2) = 10 itérations.

□

Les appréciations faites dans les sections précédentes se confirment. Il en est de même pour l'équation intégrale de 2ème espèce.

Exemple 5.7.3

On considère l'équation

$$\int_0^1 \frac{u(s)}{|t-s|^{1/2}}\,ds - 5\,u(t) = \frac{3\pi}{8}(1+4t-4t^2)-5(4t-4t^2)^{3/4}$$

Les discrétisations utilisées sont $\mathcal{Z}_n^S$ ($n = 10$) et $\mathcal{Z}_N^G$ ($N = 100$) et les résultats sont

$$\ell(\text{Atkinson}) = 35 \text{ itérations}$$
$$\ell(\text{Brakhage}) = 18 \text{ itérations}$$

La solution est

$$u(t) = (4t-4t^2)^{3/4} \qquad t \in [0,1]$$

et le dernier itéré u^ℓ des deux méthodes vérifie

$$\text{Max}\{|u(t_i^{(N)})-u^\ell(t_i^{(N)})| : 1 \le i \le N\} \le 1.4\text{E-}02 \ .$$

□

5.8 CONCLUSIONS (provisoires...)

Dans le cadre de l'interprétation faite dans la section 3.5, on peut dire que dans la majorité des cas le lissage du résidu par l'opérateur T lui-même rend la convergence plus rapide que le lissage du résidu par l'approximation T_n. Ce fait est d'autant plus notable que la discrétisation T_n contient moins d'information sur l'opérateur T comme c'est le cas des discrétisations complètes de Fredholm et de Galerkin. On voit pourtant, que pour la discrétisation que nous avons appelée "de Projection" la supériorité de la méthode B2 face à la méthode A2 est discutable : des exemples 5.3.2, 5.4.6, 5.4.10, 5.5.1 et 5.5.3, seul l'exemple 5.4.10 montre nettement l'avantage de B2. Les cas de noyaux à dérivées non bornées

et des noyaux faiblement singuliers traités dans la section 5.7 confirment les observations que l'on vient de faire. Il est intéressant de constater à ce propos que *la méthode B2 montre sa supériorité même -et surtout- quand le noyau présente des caractéristiques "difficiles" ou "peu commodes" du point de vue de son approximation numérique* comme, par exemple, de fortes oscillations, de grandes variations, des discontinuités infinies, des dérivées infinies, etc... Pour ce qui est de l'équation intégrale de Fredholm de 2ème espèce, il semble aussi que la méthode de raffinement de Brakhage est supérieure à celle d'Atkinson. Ce fait avait été déjà remarqué dans la littérature sur ce sujet ; par exemple, dans Atkinson 1973, 1976 et Chatelin 1983.

D'autres exemples concernant les méthodes les plus simples se trouvent dans Ahués et al 1982, où l'on calcule la valeur propre dominante de trois opérateurs intégraux en utilisant différents rapports n/N pour chacune des discrétisations $\mathcal{G}_n^D$ présentées dans cette thèse. Dans Ahués et Telias 1983 on présente d'autres techniques de raffinement qui sont plus chères que B2 et lesquelles se sont démontrées moins performantes que celle-ci. Ces méthodes consistent à substituer à l'opérateur S_n, l'opérateur

$$A_{n,k} := (1-P_n)[(T_n-\lambda^{k+1})|_{(1-P_n)X}]^{-1}(1-P_n)$$

dans les formules des méthodes A2 et B2 et ainsi obtenir respectivement

$$\phi^0 := \phi_n$$

$$\phi^{k+1} := \frac{T\phi^k}{\lambda^{k+1}} + A_{n,k}\, T_n\, (\phi^k - \frac{T\phi^k}{\lambda^{k+1}}) \qquad k \geq 0$$

et

$$\phi^0 := \phi_n$$

$$\phi^{k+1} := \frac{T\phi^k}{\lambda^{k+1}} + A_{n,k}\, T(\phi^k - \frac{T\phi^k}{\lambda^{k+1}}) \qquad k \geq 0$$

où

$$\lambda^{k+1} := \langle T\phi^k, \phi_n^* \rangle .$$

Ces deux méthodes nécessitent la résolution d'un système linéaire *différent*, de taille n × (n+1), à chaque pas. Elles sont en théorie et en pratique *à convergence linéaire*. Les expériences faites dans Ahués et Telias 1983, montrent qu'elles ne sont pas plus performantes que A2 ou que B2.

Mais en tout cas, ces conclusions ne peuvent être que provisoires. Une expérimentation numérique complémentaire est nécessaire -elle continue à se faire- pour obtenir des critères plus définitifs quant à l'utilisation des diverses méthodes. L'étude de la programmation du cas de valeurs propres multiples est aussi en cours. En même temps que cette thèse subit le traitement du Service de Reprographie et se voit multipliée sans effort par dizaines d'exemplaires, une comparaison entre les méthodes générales et la méthode de Chatelin donnée par (3.5.2) est faite au Portugal (Cf. D'Almeida 1983).

On prépare aussi la documentation nécessaire pour que l'on puisse comparer, dans quelques semaines au Chili, les méthodes proposées ici avec l'algorithme d'Arnoldi (Cf. Saad 1979, 1983). D'autre part, on étudie des formulations nouvelles, équivalentes, qui ne calculent pas l'itéré en cours en fonction du résidu précédent. Ces reformulations devraient être plus stables dans le sens qu'elles n'auront pas à evaluer de fonctions d'argument arbitrairement petit. Néanmoins on remarque que nous n'avons pas eu de problèmes d'instabilité de ce genre.

Enfin, disons pour terminer que la méthode de Newton, le principe de correction du résidu, la notion de projection spectrale et la notion de convergence fortement stable nous ont fourni un cadre théorique et pratique très agréable pour l'étude et la mise en oeuvre de l'approximation numérique d'éléments propres d'un opérateur intégral compact.

Les opérateurs intégraux bidimensionnels et les opérateurs différentiels à résolvante compacte sont ad portas...

REFERENCES

AHUES, M. and CHATELIN, F. 1983
The Use of Defect Correction to Refine the Eigenelements of Compact Integral Operators.
SIAM J. Numer. Anal. (à paraître).

AHUES, M. and TELIAS, M. 1983
Refinement Methods of Newton Type for Approximate Eigenelements of Integral Operators.
SIAM J. Numer. Anal.(soumis).

AHUES, M. ; D'ALMEIDA, F. and TELIAS, M. 1982
On the Defect Correction Method with Applications to Iterative Refinement Techniques.
R.R. IMAG n° 324 .
Université de Grenoble .

AHUES, M.; CHATELIN, F.; D'ALMEIDA, F. and TELIAS, M. 1983 a
Iterative Refinement Techniques for the Eigenvalue Problem of Compact Integral Operators. In Treatment of Integral Equations by Numerical Methods.
C.T.H. Baker and G.F. Miller Eds. pp. 373-385.
Academic Press London .

AHUES, M.; D'ALMEIDA, F. and TELIAS, M. 1983 b
Iterative Refinement for Approximate Eigenelements of Compact Operators.
RAIRO Anal. Numér.
A paraître.

AHUES, M.; D'ALMEIDA, F. and TELIAS, M. 1983 c
Two Defect Correction Methods for the Eigenvalue Problem of Compact Operators in Banach Spaces.
J. Integral Equations (soumis).

ANSELONE, P.M. 1971
Collectively Compact Operator Approximation Theory .
Prentice-Hall, Engelwood Cliffs, New Yersey.

ANSELONE, P.M. and ANSORGE, R. 1979
Compactness Principle in Nonlinear Operator Approximation Theory.
Numer. Funct. Anal. Optim. 1, 589-618.

ATKINSON, K. 1973
Iterative Variants of the Nyström Method for the Numerical Solution of Integral Equations.
Numer. Math. 22, 17-31.

ATKINSON, K. 1976
A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind.
SIAM Philadelphia, Pennsylvania.

AXELSSON, O. 1982
On global convergence of iterative methods. In Lecture Notes in Mathematics 953 : Iterative Solution of Nonlinear Systems of Equations.
R. Ansorge, Th. Meis and W. Törnig Eds. pp. 1-19.
Springer-Verlag Berlin, Heidelberg, New-York.

BANK, R.E. and ROSE, D.J. 1981
Global Approximate Newton Methods.
Numer. Math. 37, 279-295.

BARTELS, R.H. and STEWART, G.W. 1972
A Solution of the Equation AX + XB = C.
Commun. ACM 15, 802-826.

BJÖRCK, Å. and GOLUB, G. 1973
Numerical Methods for Computing Angles Between Linear Subspaces.
Math. Comp. 27, 579-594.

BÖHMER, K. 1981
Discrete Newton Methods and Iterated Defect Corrections.
Numer. Math. 37, 167-192.

BRAKHAGE, H. 1960
Uber die numerische Behandlung von Integralgleichungen nach der Quadraturformelmethode.
Numer. Math. 2, 183-196.

CARTAN, H. 1972
Calculo Diferencial.
Editorial Omega, Barcelona.

CHATELIN, F. 1983
Spectral Approximation of Linear Operators.
Academic Press, New-York.

CHATELIN, F. and MIRANKER, W.L. 1982
Acceleration by Aggregation of Successive Approximation Methods.
Linear Algebra Appl. 43, 17-47.

CHATELIN, F. and MIRANKER, W.L. 1983
Aggregation/Disaggregation for Eigenvalue Problems.
SIAM J. Numer. Anal.
(soumis)

CHU, K.W. and SPENCE, A. 1981
Deferred Correction for the Integral Equation Eigenvalue Problem.
J. Austral. Math. Soc. 22, 474-487.

COURANT-HILBERT 1953
Methods of Mathematical Physics.
Vol. 1.
Interscience Publishers. New-York.

CUBILLOS, P.O. 1980
On the Numerical Solution of Fredholm Integral Equations of the Second Kind.
Ph.D. Thesis.
University of Iowa.

D'ALMEIDA, F. 1983
Problemas de valores propios de operadores integrais compactos. Métodos de refinamento de soluções aproximadas.
Tese de Doutoramento .
Univ. de Porto (à paraître).

DAVIS, CH. and KAHAN, W.M. 1970
The Rotation of Eigenvectors by a Perturbation III.
SIAM J. Numer. Anal. 7, 1-46.

DEMBO, F.S.; EISENSTAT, S.C. and STEIHAUG, T. 1982
Inexact Newton Methods.
SIAM J. Numer. Anal. 19, 400-408.

DENNIS, J.E. 1969
On the Kantorovic Hypothesis for Newton's Method.
SIAM J. Numer. Anal. 6, 493-507.

DENNIS, J.E. 1971
Toward a Unified Convergence Theory for Newton-Like Methods.
In Nonlinear Functional Analysis and Applications.
L.B. Rall Ed. pp. 425-472.
Academic Press New-York.

DENNIS, J.E. and MORE, J.J. 1974
A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods.
Math. Comp. 28, 549-560.

DUNFORD, N. and SCHWARTZ, J.T. 1958
Linear Operators, Part. I : General Theory.
Wiley (Interscience) New-York.

GOLUB, G.H., NASH, S. and VAN LOAN, C. 1979
A Hessenberg-Schur Method for the Problem AX + XB = C.
IEEE Trans. Autom. Control AC-24, 909-913.

GRAHAM, I.G. and SLOAN, I.H. 1979
On the Compactness of Certain Integral Operators.
J. Math. Anal. Appl. 68, 580-594.

HACKBUSCH, W. 1979
On the Computation of Approximate Eigenvalues and Eigenfunctions of Elliptic Operators by Means of a Multigrid Method.
SIAM J. Numer. Anal. 16, 201-215.

HACKBUSCH, W. 1981 a
Introduction to Multigrid Methods for the Numerical Solution of Boundary Value Problems.
Sém. Anal. Numér. n° 358.
Université de Grenoble.

HACKBUSCH, W. 1981 b
On the Convergence of Multigrid Iterations.
Beiträge Numer. Math. 9, 213-239.

HEMKER, P.W. 1982 a
The Defect Correction Principle.
BAIL II short course lecture notes.

HEMKER, P.W. 1982 b
Extensions of the Defect Correction Principle.
BAIL II short course lecture notes.

HEMKER, P.W. and SCHIPPERS, H. 1981
Multigrid Methods for the Solution of Fredholm Equations of the Second Kind.
Math. Comp. 36, 215-232.

JIANG, Z.Q. 1982
Experimentation des Méthodes Itératives de Newton et Gauss-Seidel en Variables Discrètes.
Thèse d'Université.
Université de Grenoble.

KANTOROVICH, L.V., and AKILOV, G.P. 1964
Functional Analysis in Normed Spaces.
Pergamon Press New-York.

KANTOROVICH, L.V. and KRYLOV, V.I. 1955
Approximate Methods for Higher Analysis.
Wiley (Interscience), New-York.

KASPAR, B. 1982
Overrelaxation in monotonically convergent iteration methods.
In Lecture Notes in Mathematics 953 : Iterative Solution of Nonlinear Systems of Equations.
R. Ansorge, Th. Meis and W. Törnig Eds. pp. 80-87.
Springer Verlag Berlin, Heidelberg, New-York.

KATO, T. 1966
Perturbation Theory for Linear Operators.
Springer Verlag Berlin, Heidelberg, New-York.

KRASNOSELSKII, M.A.; VAINIKKO, G.M.; ZABREIKO, P.P.; RUTITSKII, Ya.B. and STETSENKO, V.Ya. 1972
Approximate Solutions of Operator Equations.
Wolters-Noordhoff, Groningen, The Netherlands.

KULKARNI, R. 1982
Convergence and Computation of Approximate Eigenelements.
Ph. D. Thesis
Indian Institute of Technology, Bombay.

LANG, S. 1977
Analyse Réelle.
Inter Editions, Paris.

LEMORDANT, J. 1980
Localisation de Valeurs Propres et Calcul de Sous-espaces Invariants.
Thèse d'Etat.
Université de Grenoble.

LIN QUN 1982
Iterative Refinement of Finite Element Approximations for Elliptic Problems.
RAIRO Anal. Numér. 16, 39-47.

MORE, J.J. and SORENSEN, D.C. 1982
Newton's Method.
Technical Report ANL-82-8.
Argonne National Laboratory, Illinois.

NICOLAIDES, R.A. 1979
On Some Theoretical and Practical Aspects of Multigrid Methods.
Math. Comp. 33, 933-952.

ORTEGA, J.M. and RHEINBOLDT, W.C. 1970
Iterative Solution of Nonlinear Equations in Several Variables.
Academic Press New-York.

PHILLIPS, J. 1972
The Use of Collocation as a Projection Method for Solving Linear Operator Equations.
SIAM J. Numer. Anal. 9, 14-28.

POTRA, F. 1982
On the convergence of a class of Newton-like Methods.
In Lecture Notes in Mathematics 953 : Iterative Solution of Nonlinear Systems of Equations.
R. Ansorge, Th. Meis and W. Törnig Eds. pp. 125-137.
Springer-Verlag Berlin, Heidelberg, New-York.

POTRA, F. and PTAK, V. 1980
Sharp Error Bounds for Newton's Process.
Numer. Math. 34, 63-72.

POTRA, F. and PTAK, V. 1983
Nondiscrete Induction and an Inversion-free Modification of Newton's Method.
Unpublished manuscript.

RALL, L.
Computational Solution of Nonlinear Operator Equations.
Wiley (Interscience) New-York.

ROACHE, P.J. 1972
Computational Fluid Dynamics.
Hermosa Publishers, Albuquerque. New Mexico.

ROBERT, F. 1981
Itérations Discrètes .
Cours polycopié INPG-USMG, Grenoble.

SAAD, Y. 1979
Etude de la Convergence du Procédé d'Arnoldi pour le Calcul des Eléments Propres de Grandes Matrices Creuses Non Symétriques.
Sém. Anal. Numér. N° 321.
Université de Grenoble.

SAAD, Y. 1980
On the Rates of Convergence of the Lanczos and the Block-Lanczos Methods.
SIAM J. Numer. Anal. 17, 687-706.

SAAD, Y. 1983
Méthodes Numériques pour la Résolution de Problèmes Matriciels de Grandes Dimensions.
Thèse d'Etat, Université de Grenoble.

SATO, K. 1981
Global Convergence Features of Newton's Method Applied to Polynomial Equations.
In Lecture Notes in Numerical and Applied Analysis : The Newton Method and Related Topics.
H. Fujita and M. Yamaguti Eds. pp. 23-56. Ed. Kinokunyia, Kyoto.

SCHNEIDER, C. 1979
Regularity of the Solution of a Class of Weakly Singular Fredholm Integral Equations of the Second Kind.
Integral Equations Operator Theory 2, 62-68.

SLOAN, I.H. 1980
On Choosing Points in Product Integration.
J. Math. Phys. 21, 1032-1039.

SORENSEN, D.C. 1982
Newton's Method with a Model Trust Region Modification.
SIAM J. Numer. Anal. 19, 409-426.

STETTER, H. 1978
The Defect Correction Principle and Discretization Methods.
Numer. Math. 29, 425-443.

STEWART, G.W. 1973
Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems.
SIAM Review, 15, 727-764.

TANABE, K. 1981
Feasibility-improving gradient-acute-projection Methods : A Unified Approach to Nonlinear Programming.
In Lecture Notes in Numerical and Applied Analysis. The Newton Method and Related Topics.
H. Fujita and M. Yamaguti Eds pp. 57-76.
Ed. Kinokunyia, Kyoto.

TELIAS, M. 1983
Résolution de l'Equation de Convection-diffusion et d'un Modèle des Circulations Océaniques Générales par des Méthodes d'Eléments Finis.
Thèse de Docteur-Ingénieur.
Université de Grenoble - Institut National Polytechnique de Grenoble.

VARAH, J.M. 1970
Computing Invariant Subspaces of a General Matrix when the Eigensystem is Poorly Conditioned.
Math. Comp. 24, 137-149.

VESENTINI, E. 1968
On the Subharmonicity of the Spectral Radius.
Bollet. Unione Mat. Ital. 4, 427-429.

VOGEL, Th. 1953
Les Fonctions Orthogonales dans les Problèmes aux Limites de la Physique Mathématique.
Eds. C.N.R.S. Paris.

WATKINS, D. 1982
Understanding the QR Algorithm.
SIAM Review 24, 427-440.

WERNER, W. 1982
On the Simultaneous Determination of Polynomial Roots.
In Lecture Notes in Mathematics 953 : Iterative Solution of Nonlinear Systems of Equations.
R. Ansorge, Th. Meis and M. Törnig Eds. pp. 188-202.
Springer-Verlag Berlin, Heidelberg, New-York.

YOSIDA, K. 1971
Functional Analysis.
Springer Verlag Berlin, Heidelberg, New-York.

AUTORISATION DE SOUTENANCE

VU les dispositions de l'article 3 de l'arrêté du 16 avril 1974,

VU les rapports de présentation de

. Madame F. CHATELIN, Professeur
. Monsieur P.J LAURENT, Professeur

Monsieur AHUES BLANCHAIT Mario Paul

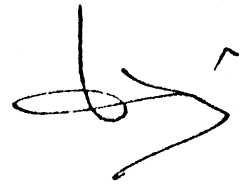
est autorisé à présenter une thèse en soutenance pour l'obtention du diplôme de DOCTEUR-INGENIEUR, spécialité "Mathématiques Appliquées".

Fait à Grenoble, le 18 mai 1983

Le Président de l'I.N.P.-G,

D. BLOCH

P.O. le Vice-Président,



RESUME :

Dans cette thèse on propose quatre familles de méthodes itératives pour le raffinement d'éléments propres approchés d'un opérateur compact dans un espace de Banach complexe. Ces méthodes sont de type Newton et le calcul de l'inverse de la dérivée de l'opérateur non linéaire dont on calcule un zéro est fait à l'aide de techniques fondées sur le Principe de Correction du Résidu. Selon la précision de ce calcul on peut atteindre une convergence quadratique, superlinéaire ou linéaire. On présente des applications aux opérateurs intégraux à noyau continu ou faiblement singulier. Les discrétisations considérées sont les approximations de Galerkin, Projection et Sloan – avec ou sans quadrature – et les approximations de Fredholm et Nyström. On conclut ce travail par de nombreux exemples numériques.

Mots clés : *Opérateur compact, Projection Spectrale, Valeurs et vecteurs propres, Convergence fortement stable, Convergence Collectivement compacte, Convergence en norme, Méthode de Newton, Principe de Correction du Résidu, Opérateur intégral.*