Problèmes liés à l'utilisation des méthodes d'éléments finis pour le calcul des valeurs propres

Christian Lajaunie

THESE

présentée à

l' Université Scientifique et Médicale de Grenoble

et à

l' Institut National Polytechnique de Grenoble

pour obtenir le grade de
DOCTEUR INGENIEUR
«Analyse numérique»

par

LAJAUNIE Christian

PROBLEMES LIES A L'UTILISATION DES METHODES D'ELEMENTS FINIS POUR LE CALCUL DES VALEURS PROPRES.

Thèse soutenue le 26 juin 1980 devant la commission d'examen.

Madame	F. CHATELIN	Présidente
Messieurs	N. GASTINEL	
	J.F. MAITRE	Examinateurs
	A. PONCET	

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

Monsieur Gabriel CAU : Président
Monsieur Joseph KLEIN : Vice-Président

MEMBRES DU CORPS ENSEIGNANT DE L'U.S.M.G.

PROFESSEURS TITULAIRES

MM.	AMBLARD Pierre	Clinique de dermatologie
	ARNAUD Paul	Chimie
	ARVIEU Robert	I.S.N.
	AUBERT Guy	Physique
	AYANT Yves	Physique approfondie
Mme	BARBIER Marie-Jeanne	Electrochimie
MM.	BARBIER Jean-Claude	Physique expérimentale
	BARBIER Reynold	Géologie appliquée
	BARJON Robert	Physique nucléaire
	BARNOUD Fernand	Biosynthèse de la cellulose
	BARRA Jean-René	Statistiques
	BARRIE Joseph	Clinique chirurgicale A
	BEAUDOING André	Clinique de pédiatrie et puériculture
	BELORIZKY Elie	Physique
	BARNARD Alain	Mathématiques pures
Mme	BERTRANDIAS Françoise	Mathématiques pures
MM.	BERTRANDIAS Jean-Paul	Mathématiques pures
	BEZES Henri	Clinique chirurgicale et traumatologie
	BLAMBERT Maurice	Mathématiques pures
	BOLLIET Louis	Informatique (I.U.T. B)
	BONNET Jean-Louis	Clinique ophtalmologie
	BONNET-EYMARD Joseph	Clinique hépato-gastro-entérologie
Mme	BONNIER Marie-Jeanne	Chimie générale
MM.	BOUCHERLE André	Chimie et toxicologie
	BOUCHEZ Robert	Physique nucléaire
	BOUSSARD Jean-Claude	Mathématiques appliquées
	BOUTET DE MONVEL Louis	Mathématiques pures
	BRAVARD Yves	Géographie
	CABANEL Guy	Clinique rhumatologique et hydrologique
	CALAS François	Anatomie
	CARLIER Georges	Biologie végétale
	CARRAZ Gilbert	Biologie animale et pharmacodynamie

.../...

MM.	CAU Gabriel	Médecine légale et toxicologie
	CAUQUIS Georges	Chimie organique
	CHABAUTY Claude	Mathématiques pures
	CHARACHON Robert	Clinique ot-rhino-laryngologique
	CHATEAU Robert	Clinique de neurologie
	CHIBON Pierre	Biologie animale
	COEUR André	Pharmacie chimique et chimie analytique
	COUDERC Pierre	Anatomie pathologique
	DEBELMAS Jacques	Géologie générale
	DEGRANGE Charles	Zoologie
	DELORMAS Pierre	Pneumophtisiologie
	DEPORTES Charles	Chimie minérale
	DESRE Pierre	Métallurgie
	DODU Jacques	Mécanique appliquée (I.U.T. I)
	DOLIQUE Jean-Michel	Physique des plasmas
	DREYFUS Bernard	Thermodynamique
	DUCROS Pierre	Cristallographie
	FONTAINE Jean-Marc	Mathématiques pures
	GAGNAIRE Didier	Chimie physique
	GALVANI Octave	Mathématiques pures
	GASTINEL Noël	Analyse numérique
	GAVEND Michel	Pharmacologie
	GEINDRE Michel	Electroradiologie
	GERBER Robert	Mathématiques pures
	GERMAIN Jean-Pierre	Mécanique
	GIRAUD Pierre	Géologie
	JANIN Bernard	Géographie
	KAHANE André	Physique générale
	KLEIN Joseph	Mathématiques pures
	KOSZUL Jean-Louis	Mathématiques pures
	KRAVTCHENKO Julien	Mécanique
	LACAZE Albert	Thermodynamique
	LACHARME Jean	Biologie végétale
Mme	LAJZEROWICZ Janine	Physique
MM.	LAJZEROWICZ Joseph	Physique
	LATREILLE René	Chirurgie générale
	LATURAZE Jean	Biochimie pharmaceutique
	LAURENT Pierre	Mathématiques appliquées
	LEDRU Jean	Clinique médicale B
	LE ROY Philippe	Mécanique (I.U.T. I)

.../...

MM.	LLIBOUTRY Louis	Géophysique
	LOISEAUX Jean-Marie	Sciences nucléaires
	LONGEQUEUE Jean-Pierre	Physique nucléaire
	LOUP Jean	Géographie
Mlle	LUTZ Elisabeth	Mathématiques pures
MM.	MALINAS Yves	Clinique obstétricale
	MARTIN-NOEL Pierre	Clinique cardiologique
	MAYNARD Roger	Physique du solide
	MAZARE Yves	Clinique Médicale A
	MICHEL Robert	Minéralogie et pétrographie
	MICOUD Max	Clinique maladies infectieuses
	MOURIQUAND Claude	Histologie
	MOUSSA André	Chimie nucléaire
	NEGRE Robert	Mécanique
	NOZIERES Philippe	Spectrométrie physique
	OZENDA Paul	Botanique
	PAYAN Jean-Jacques	Mathématiques pures
	PEBAY-PEYROULA Jean-Claude	Physique
	PERRET Jean	Séméiologie médicale (neurologie)
	RASSAT André	Chimie systématique
	RENARD Michel	Thermodynamique
	REVOL Michel	Urologie
	RINALDI Renaud	Physique
	DE ROUGEMONT Jacques	Neuro-Chirurgie
	SARRAZIN Roger	Clinique chirurgicale B
	SEIGNEURIN Raymond	Microbiologie et hygiène
	SENGEL Philippe	Zoologie
	SIBILLE Robert	Construction mécanique (I.U.T. I)
	SOUTIF Michel	Physique générale
	TANCHE Maurice	Physiologie
	VAILLANT François	Zoologie
	VALENTIN Jacques	Physique nucléaire
Mme	VERAIN Alice	Pharmacie galénique
MM.	VERAIN André	Physique biophysique
	VEYRET Paul	Géographie
	VIGNAIS Pierre	Biochimie médicale

.../...

PROFESSEURS ASSOCIES

MM.	CRABBE Pierre	CERMO
	SUNIER Jules	Physique

PROFESSEURS SANS CHAIRE

Mlle	AGNIUS-DELORS Claudine	Physique pharmaceutique
	ALARY Josette	Chimie analytique
MM.	AMBROISE-THOMAS Pierre	Parasitologie
	ARMAND Gilbert	Géographie
	BENZAKEN Claude	Mathématiques appliquées
	BIAREZ Jean-Pierre	Mécanique
	BILLET Jean	Géographie
	BOUCHET Yves	Anatomie
	BRUGEL Lucien	Energétique (I.U.T. I)
	BUISSON René	Physique (I.U.T. I)
	BUTEL Jean	Orthopédie
	COHEN-ADDAD Jean-Pierre	Spectrométrie physique
	COLOMB Maurice	Biochimie médicale
	CONTE René	Physique (I.U.T. I)
	DELOBEL Claude	M.I.A.G.
	DEPASSEL Roger	Mécanique des fluides
	GAUTRON René	Chimie
	GIDON Paul	Géologie et minéralogie
	GLENAT René	Chimie organique
	GROULADE Joseph	Biochimie médicale
	HACQUES Gérard	Calcul numérique
	HOLLARD Daniel	Hématologie
	HUGONOT Robert	Hygiène et médecine préventive
	IDELMAN Simon	Physiologie animale
	JOLY Jean-René	Mathématiques pures
	JULLIEN Pierre	Mathématiques appliquées
Mme	KAHANE Josette	Physique
MM.	KRAKOWIACK Sacha	Mathématiques appliquées
	KUHN Gérard	Physique (I.U.T. I)
	LUU DUC Cuong	Chimie organique - pharmacie
	MICHOULIER Jean	Physique (I.U.T. I)
Mme	MINIER Colette	Physique (I.U.T. I)

.../...

MM.	PELMONT Jean	Biochimie
	PERRIAUX Jean-Jacques	Géologie et minéralogie
	PFISTER Jean-Claude	Physique du solide
Mlle	PIERY Yvette	Physiologie animale
MM.	RAYNAUD Hervé	M.I.A.G.
	REBECQ Jacques	Biologie (CUS)
	REYMOND Jean-Charles	Chirurgie générale
	RICHARD Lucien	Biologie végétale
Mme	RINAUDO Marguerite	Chimie macromoléculaire
MM.	SARROT-REYNAULD Jean	Géologie
	SIROT Louis	Chirurgie générale
Mme	SOUTIF Jeanne	Physique générale
MM.	STIEGLITZ Paul	Anesthésiologie
	VIALON Pierre	Géologie
	VAN CUTSEM Bernard	Mathématiques appliquées

MAITRES DE CONFERENCES ET MAITRES DE CONFERENCES AGREGES

MM.	ARMAND Yves	Chimie (I.U.T. I)
	BACHELOT Yvan	Endocrinologie
	BARGE Michel	Neuro-chirurgie
	BEGUIN Claude	Chimie organique
Mme	BERIEL Hélène	Pharmacodynamie
MM.	BOST Michel	Pédiatrie
	BOUCHARLAT Jacques	Psychiatrie adultes
Mme	BOUCHE Liane	Mathématiques (CUS)
MM.	BRODEAU François	Mathématiques (I.U.T. B) (Personne étrangère habilitée à être directeur de thèse)
	BERNARD Pierre	Gynécologie
	CHAMBAZ Edmond	Biochimie médicale
	CHAMPETIER Jean	Anatomie et organogénèse
	CHARDON Michel	Géographie
	CHERADAME Hervé	Chimie papetière
	CHIAVERINA Jean	Biologie appliquée (EFP)
	COLIN DE VERDIERE Yves	Mathématiques pures
	CONTAMIN Charles	Chirurgie thoracique et cardio-vasculaire
	CORDONNER Daniel	Néphrologie
	COULOMB Max	Radiologie
	CROUZET Guy	Radiologie

MM.	CYROT Michel	Physique du solide
	DENIS Bernard	Cardiologie
	DOUCE Roland	Physiologie végétale
	DUSSAUD René	Mathématiques (CUS)
Mme	ETERRADOSSI Jacqueline	Physiologie
MM.	FAURE Jacques	Médecine légale
	FAURE Gilbert	Urologie
	GAUTIER Robert	Chirurgie générale
	GIDON Maurice	Géologie
	GROS Yves	Physique (I.U.T. I)
	GUIGNIER Michel	Thérapeutique
	GUITTON Jacques	Chimie
	HICTER Pierre	Chimie
	JALBERT Pierre	Histologie
	JUNIEN-LAVILLAVROY Claude	O.R.L.
	KOLODIE Lucien	Hématologie
	LE NOC Pierre	Bactériologie-virologie
	MACHE Régis	Physiologie végétale
	MAGNIN Robert	Hygiène et médecine préventive
	MALLION Jean-Michel	Médecine du travail
	MARECHAL Jean	Mécanique (I.U.T. I)
	MARTIN-BOUYER Michel	Chimie (CUS)
	MASSOT Christian	Médecine interne
	NEMOZ Alain	Thermodynamique
	NOUGARET Marcel	Automatique (I.U.T. I)
	PARAMELLE Bernard	Pneumologie.
	PECCOUD François	Analyse (I.U.T. B) (Personnalité étrangère habilitée à être directeur de thèse)
	PEFFEN René	Métallurgie (I.U.T. I)
	PERRIER Guy	Géophysique-glaciologie
	PHELIP Xavier	Rhumatologie
	RACHALL Michel	Médecine interne
	RACINET Claude	Gynécologie et obstétrique
	RAMBAUD Pierre	Pédiatrie
	RAPHAEL Bernard	Stomatologie
Mme	RENAUDET Jacqueline	Bactériologie (pharmacie)
MM.	ROBERT Jean-Bernard	Chimie-physique
	ROMIER Guy	Mathématiques (I.U.T. B) (Personnalité étrangère habilitée à être directeur de thèse)
	SAKAROVITCH Michel	Mathématiques appliquées

.../...

MM.	SCHAERER René	Cancérologie
Mme	SEIGLE-MURANDI Françoise	Crytogamie
MM.	STOEBNER Pierre	Anatomie pathologie
	STUTZ Pierre	Mécanique
	VROUSOS Constantin	Radiologie

MAITRES DE CONFERENCES ASSOCIES

MM.	DEVINE Roderick	Spectro Physique
	KANEKO Akira	Mathématiques pures
	JOHNSON Thomas	Mathématiques appliquées
	RAY Tuhina	Physique

MAITRE DE CONFERENCES DELEGUE

M.	ROCHAT Jacques	Hygiène et hydrologie (pharmacie)

Fait à Saint Martin d'Hères, novembre 1977

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Année universitaire 1979-1980

Président : M. Philippe TRAYNARD

Vice-Présidents : M. Georges LESPINARD
M. René PAUTHENET

PROFESSEURS DES UNIVERSITES

MM.	ANCEAU François	Informatique fondamentale et appliquée
	BENOIT Jean	Radioéléctricité
	BESSON Jean	Chimie Minérale
	BLIMAN Samuel	Electronique
	BLOCH Daniel	Physique du Solide - Cristallographie
	BOIS Philippe	Mécanique
	BONNETAIN Lucien	Génie Chimique
	BONNIER Etienne	Métallurgie
	BOUVARD Maurice	Génie Mécanique
	BRISSONNEAU Pierre	Physique des Matériaux
	BUYLE-BODIN Maurice	Electronique
	CHARTIER Germain	Electronique
	CHERADAME Hervé	Chimie Physique Macromoléculaires
Mme	CHERUY Arlette	Automatique
MM.	CHIAVERINA Jean	Biologie, Biochimie, Agronomie
	COHEN Joseph	Electronique
	COUMES André	Electronique
	DURAND Francis	Métallurgie
	DURAND Jean-Louis	Physique Nucléaire et Corpusculaire
	FELICI Noël	Electrotechnique
	FOULARD Claude	Automatique
	GUYOT Pierre	Métallurgie Physique
	IVANES Marcel	Electrotechnique
	JOUBERT Jean-Claude	Physique du Solide - Cristallographie
	LACOUME Jean-Louis	Géographie - Traitement du Signal
	LANCIA Roland	Electronique - Automatique
	LESIEUR Marcel	Mécanique
	LESPINARD Georges	Mécanique
	LONGEQUEUE Jean-Pierre	Physique Nucléaire Corpusculaire
	MOREAU René	Mécanique
	MORET Roger	Physique Nucléaire Corpusculaire
	PARIAUD Jean-Charles	Chimie - Physique
	PAUTHENET René	Physique du Solide - Cristallographie
	PERRET René	Automatique

.../...

MM.	PERRET Robert	Electrotechnique
	PIAU Jean-Michel	Mécanique
	PIERRARD Jean-Marie	Mécanique
	POLOUJADOFF Michel	Electrotechnique
	POUPOT Christian	Electronique - Automatique
	RAMEAU Jean-Jacques	Chimie
	ROBERT André	Chimie Appliquée et des matériaux
	ROBERT François	Analyse numérique
	SABONNADIERE Jean-Claude	Electrotechnique
Mme	SAUCIER Gabrielle	Informatique fondamentale et appliquée
M.	SOHM Jean-Claude	Chimie - Physique
Mme	SCHLENKER Claire	Physique du Solide - Cristallographie
MM.	TRAYNARD Philippe	Chimie - Physique
	VEILLON Gérard	Informatique fondamentale et appliquée
	ZADWORNY François	Electronique

CHERCHEURS DU C.N.R.S. (Directeur et Maître de Recherche)

M.	FRUCHART Robert	Directeur de Recherche
MM.	ANSARA Ibrahïm	Maître de Recherche
	BRONOEL Guy	Maître de Recherche
	CARRE René	Maître de Recherche
	DAVID René	Maître de Recherche
	DRIOLE Jean	Maître de Recherche
	KAMARINOS Georges	Maître de Recherche
	KLEITZ Michel	Maître de Recherche
	LANDAU Ioan-Doré	Maître de Recherche
	MERMET Jean	Maître de Recherche
	MUNIER Jacques	Maître de Recherche

Personnalités habilitées à diriger des travaux de recherche (décision du Conseil Scientifique)

E.N.S.E.E.G.

MM.	ALLIBERT Michel	
	BERNARD Claude	
	CAILLET Marcel	
Mme	CHATILLON Catherine	
MM.	COULON Michel	
	HAMMOU Abdelkader	
	JOUD Jean-Charles	
	RAVAINE Denis	
	SAINFORT	C.E.N.G.

.../...

MM. SARRAZIN Pierre
SOUQUET Jean-Louis
TOUZAIN Philippe
URBAIN Georges — Laboratoire des Ultra-Réfractaires ODEILLO

E.N.S.M.E.E.

MM. BISCONDI Michel
BOOS Jean-Yves
GUILHOT Bernard
KOBILANSKI André
LALAUZE René
LANCELOT Françis
LE COZE Jean
LESBATS Pierre
SOUSTELLE Michel
THEVENOT François
THOMAS Gérard
TRAN MINH Canh
DRIVER Julian
RIEU Jean

E.N.S.E.R.G.

MM. BOREL Joseph
CHEHIKIAN Alain
VIKTOROVITCH Pierre

E.N.S.I.E.G.

MM. BORNARD Guy
DESCHIZEAUX Pierre
GLANGEAUD François
JAUSSAUD Pierre
Mme JOURDAIN Geneviève
MM. LEJEUNE Gérard
PERARD Jacques

E.N.S.H.G.

M. DELHAYE Jean-Marc

E.N.S.I.M.A.G.

MM. COURTIN Jacques
LATOMBE Jean-Claude
LUCAS Michel
VERDILLON André

Je tiens a remercier tous ceux qui m'ont aidé ou encouragé dans la réalisation de ce travail.

Je pense tout particulièrement à Monsieur le Professeur N. Gastinel qui en a assuré la direction, et dont les suggestions ont été maintes fois utiles.

Je pense aussi à Madame F. Chatelin, qui a acceptée la présidence du Jury. Je tiens à la remercier pour son enseignement et ses précieuses indications concernant les problèmes de valeurs propres.

Mes remerciements vont également à Monsieur A. Poncet sans lequel la mise en oeuvre n'aurait pas été possible. Merci pour la patience et la gentillesse avec laquelle il m'a initié au programme d'éléments finis "DELTA", et pour son aide permanente.

Je suis profondément reconnaissant à Monsieur J.F. Maitre de s'être intéressé à ce travail, et d'avoir accepté de participer au Jury.

Enfin, je voudrais profiter de cette occasion pour exprimer le plaisir que j'ai eu a travailler en compagnie des membres de l'Equipe d'Analyse Numérique, et pour remercier Mesdames G. Assadas et Cl. Meyrieux pour le soin et la célérité qu'elles ont apporté à la frappe de cette thèse.

PREMIERE PARTIE

TABLE DES MATIERES DE LA PREMIERE PARTIE

INTRODUCTION

L'objet de cette première partie est la description et l'étude de quelques méthodes de calcul de valeurs propres "extrêmes", qui nous ont semblées intéressantes par la simplicité des calculs "de base" mis en jeu.

Cette caractéristique les rend particulièrement adaptées au cas de problèmes provenant de l'utilisation de méthodes d'éléments finis, pour le calcul des valeurs propres d'un opérateur autoadjoint.

On a alors un problème "matriciel" du type :

$$Ax = \lambda Bx$$

où les matrices A et B sont définies respectivement à partir d'une forme bilinéaire symétrique et continue, et d'un produit scalaire sur certains espaces :

$$a_{ij} = a\ (P_i\ ,\ P_j) \qquad ,\quad A = (a_{ij})$$

$$b_{ij} = \quad (P_i\ ,\ P_j) \qquad ,\quad B = (b_{ij})$$

Les P_i sont ici les fonctions de "base" d'un espace de dimension finie, en général inclus dans l'espace fonctionnel dans lequel le problème est posé.

La taille des matrices est égale à la dimension de l'espace d'approximation utilisé ; on a typiquement des valeurs de l'ordre du millier pour celle-ci. Cependant, les fonctions de base P_i étant à support étroit, il en résulte que les matrices sont très creuses. Si on utilise, par exemple des éléments de Lagrange de degré un (avec trois noeudspar triangle) les lignes de chaque matrice contiendront, en général moins d'une dizaine d'éléments non nuls.

Ce fait montre bien tout l'intérêt qu'il y a éviter les transformations de matrices telles que les décompositions de Choleski, car, même s'il est possible par une permutation d'indices, de donner une structure "bande" aux matrices, la longueur de la bande restera toujours considérablement plus grande que le nombre d'éléments non nuls contenus dans une ligne.

C'est là le principal atout des méthodes décrites ici par rapport aux méthodes telles que l'algorithme de Lanczos [9] ou la méthode des itérations simultanées [5] , qui exigent, l'une et l'autre, une "factorisation" de B sous la forme

$$B = L\ L^T$$

de manière à pouvoir résoudre à chaque étape le système linéaire associé.

Les matrices A et B sont, d'autre part, symétriques définies positives : B par le fait de l'usage d'un produit scalaire, et A dès que la forme bilinéaire vérifie l'hypothèse d'ellipticité forte :

$$\alpha > 0 \text{ tel que } \quad \forall u \qquad a(u,u) \geq \alpha ||u||^2$$

Ces hypothèses resteront toujours sous-jacentes tout au long de cette première partie, bien que, en fait, nous puissions supposer A seulement semi-définie positive.

De ce fait, les valeurs propres sont toutes réelles positives, et nous pouvons les ordonner :

$$0 \leq \lambda_1 \leq \lambda_2 \qquad \leq \lambda_n$$

On cherchera à calculer λ_1, plus petite des valeurs propres du problème, en tirant partie des propriétés extrémales du quotient de Rayleigh. C'est-à-dire que nous cherchons

$$\lambda_1 = \underset{x \neq 0}{\text{Min}} \ \frac{x^e Ax}{x^t Bx}$$

Le calcul des valeurs propres superieures : λ_i , $i > 1$ est possible par deux procédés que nous décrivons rapidement : Le premier est un calcul recursif basé sur des techniques de déflation ; le second utilise un calcul simultanné de plusieurs vecteurs , avecune methode de projection (comme dans la methode des itérations simultannées)

Nous renvoyons à (4) et (11) pour un tour d' horizon des principales méthodes utilisées pour le probleme Ax = Bx , ainsi q'une bibliographie s'y rapportant

Le Plan de cette première partie est le suivant :

Dans un premier temps, on montrera comment des exigences très naturelles de monotonie des suites de quotients de Rayleigh, permettent de définir une classe assez vaste d'algorithmes.

Le fait que les méthodes de puissance itérée et de Gradient puissent être retrouvées avec ce formalisme n'est cité ici qu'à titre anecdotique, et nous n'avons pas cherché à faire l'étude de ces deux méthodes. On trouvera dans [6] une méthode de Gradient pour le problème généralisé à matrices non symétriques, ainsi qu'une bibliographie sur l'application du Gradient au calcul de valeurs propres.

Le paragraphe A.23 décrit deux méthodes :
l'une (Successive Over Relaxation) a été abondamment étudiée par Axel Ruhe [7] et [8] . L'autre apparait comme une généralisation d'une méthode utilisée par Buffoni [1] .Les paragraphes suivants décrivent des versions par blocs (à notre connaissance originales) de ces algorithmes.

Le second chapitre est consacré à l'établissement de preuves de convergence des algorithmes introduits en A. On montrera que ces méthodes produisent des suites de vecteurs dont les quotients de Rayleigh convergent vers λ_i , valeur propre du problème. Une étude plus fine montrera que, d'une manière générale, $\lambda_i = \lambda_i$ ou λ_n selon un choix que l'utilisateur aura fixé. Cependant, les cas de convergences vers des valeurs propres intermédiaires ne pourront pas être en toute rigueur définitivement écartés.

Enfin un troisième chapitre est consacré à la description de détails pratiques d'implementation ; et au compte rendu d'essais numériques

(A) DEFINITION DE QUELQUES ALGORITHMES

Précisons d'abord les notations utilisées par la suite dans cette première partie :

Les algorithmes que nous allons décrire génèrent des suites de vecteurs de R^n, $\{x_N\}_{N\in\mathbb{N}}$. Les approximants de valeurs propres seront données par les quotients de Rayleigh de ces vecteurs :

$$\mu_N = \frac{x_N^t A x_N}{x_N^t B x_N}$$

La matrice B définit une norme sur R^n notée de la manière suivante :

$$||x||_B = [\, x^t B x \,]^{1/2}$$

On introduira les vecteurs :

$p_N = x_{N+1} - x_N$ accroissement de x_N à l'étape (N+1)

$r_N = (A-\mu_N B)x_N$ "residuel" du problème

$\hat{x}_N = x_N / ||x_N||_B$ (si $x_N \neq 0$)

$\tilde{r}_N = (A-\mu_N B)\hat{x}_N$.

A.1 Expression de $\mu_{N+1} - \mu_N$

Nous allons donner une expression générale qui nous permettra de trouver des déterminations des directions d'accroissement, p_N, telles que les suites μ_N générées soient monotones.

$$\mu_{N+1} - \mu_N = \frac{x_{N+1}^t A\, x_{N+1}}{x_{N+1}^t B\, x_{N+1}} - \mu_N$$

$$= \frac{1}{||x_{N+1}||_B^2} \left[x_{N+1}^t (A-\mu_N B)\, x_{N+1}\right]$$

$$= \frac{1}{||x_{N+1}||_B^2} \left[x_N^t (A-\mu_N B) x_N + p_N^t (A-\mu_N B) p_N + 2\, p_N^t\, r_N\right]$$

mais $$x_N^t (A-\mu_N B) x_N = ||x_N||_B^2 \left(\frac{x_N^t A x_N}{x_N^t B x_N} - \mu_N \right) = 0$$

On a donc la relation :

(R) $$\mu_{N+1} - \mu_N = \frac{1}{||x_{N+1}||_B^2} \left[p_N^t (A - \mu_N B) p_N + 2 p_N^t r_N \right]$$

A.2 On cherchera à donner à $\mu_{N+1} - \mu_N$ un signe constant, par des déterminations appropriées de p_N.

A.2.1 Dans une méthode de Gradient , l'accroissement sera pris dans la direction de r_N, puisque nous savons que :

$$\text{grad } [\mu(x)]_{x=x_0} = \frac{2}{||x_0||_B} (A-\mu(x_0)B)x_0 = \alpha_0 r_0$$

on a donc : $$p_N = k_N r_N$$

x_{N+1} appartient donc au plan déterminé par x_N et r_N :

$$x_{N+1} = \rho_N x_N + \theta_N r_N$$

Ces paramètres peuvent être choisis de manière optimale, c'est-à-dire tels que μ_{N+1} soit minimum :

$$\mu_{N+1} = \frac{\rho_N^2 a_1^N + 2\rho_N\theta_N a_2^N + \theta_N^2 a_3^N}{\rho_N^2 b_1^N + 2\rho_N\theta_N b_2^N + \theta_N^2 b_3^N}$$

avec : $a_1^N = x_N^t A x_N$ $\quad a_2^N = x_N^t A r_N$ $\quad a_3^N = r_N^t A r_N$

$b_1^N = x_N^t B x_N$ $\quad b_2^N = x_N^t B r_N$ $\quad b_3^N = r_N^t B r_N$

μ_{N+1} est alors la plus petite valeur propre du problème suivant :

$$\begin{bmatrix} a_1^N & a_2^N \\ a_2^N & a_3^N \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{bmatrix} b_1^N & b_2^N \\ b_2^N & b_3^N \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

tandis que ρ_N et Θ_N seront les composantes d'un vecteur propre associé, choisies pour que

$$||x_{N+1}||_B = 1$$

Ceci constitue une première possibilité.

A.2.2 Plus généralement, on imposera une relation du type

$$r_N = -M_N P_N$$

M_N étant une matrice, à choisir, facile à inverser

alors :

$$\mu_{N+1} - \mu_N = \frac{1}{||x_{N+1}||_B^2} \; p_N{}^t H_N p_N$$

où $H_N = A - \mu_N B - (M_N + M_N{}^t)$

La monotonie peut être assurée par un choix de M_N tel que $\pm H_N$ soit définie positive.

On envisagera plusieurs choix possibles :

a) une première possibilité consiste à prendre :

$$r_N = - k_N A p_N \qquad (k_N \text{ scalaire})$$

alors $H_N = (1 - 2k_N)A - \mu_N B$

On est donc assuré d'avoir $- H_N$ définie positive dès que

$$k_N \geq \frac{1}{2}$$

du point de vue du calcul, on a :

$$x_{N+1} = x_N + p_N = \left(1 - \frac{1}{k_N}\right) x_N + \left(\frac{\mu_N}{k_N}\right) A^{-1} B x_N$$

On remarquera que, pour $k_N = 1$, on retrouve (à une normalisation de x_{N+1} près) la méthode de la puissance iterée inverse.

$$x_{N+1} = \rho_N A^{-1} B x_N$$

où ρ_N est un facteur de normalisation.

Ici, k_N peut être choisi optimal :

$$\mu_{N+1} = \frac{a_1^N - 2k_N a_2^N - k_N^2 a_3^N}{b_1^N - 2k_N b_2^N - k_N^2 b_3^N}$$

avec :

$$\begin{cases} a_1^N = r_N^t A^{-1} r_N \\ b_1^N = r_N^t A^{-1} {}^t B A^{-1} r_N \end{cases} \quad \begin{cases} a_2^N = x_N^t r_N \\ b_2^N = x_N^t {}^t B A^{-1} r_N \end{cases} \quad \begin{cases} a_3^N = x_N^t A x_N \\ b_3^N = x_N^t B x_N \end{cases}$$

μ_{N+1} sera donc la plus petite des deux valeurs λ_1 et λ_2 telles que :

$$p(\lambda) = \begin{vmatrix} a_1^N - \lambda b_1^N & a_2^N - \lambda b_2^N \\ a_2^N - \lambda b_2^N & b_3^N - \lambda b_3^N \end{vmatrix} = 0$$

(sous réserve, bien sur, que l'on puisse lui associer un vecteur propre de R^2 de la forme : $U^t = (1, k_N)$ avec $k_N \geq \frac{1}{2}$)

Chaque pas demande le calcul de :

$$\mu_N \; ; \; r_N \; ; \; Z_N = A^{-1} r_N \quad ; \quad b_1^N = Z_N^t B Z_N \quad \text{et} \quad B_2^N = x_N^t B Z_N$$

La nécessité d'inverser A fait que nous n'avons pas retenu cette possibilité.

b) un deuxième choix consiste à prendre

$$r_N = k_N B p_N$$

alors $M_N = A - (\mu_N - 2k_N)B$

Ce cas est très voisin du précédent, et, si $k_N \geq \mu_N/2$

alors H_N est définie positive, et la suite μ_N est croissante.

Ici encore, à chaque étape, on est conduit à résoudre un système linéaire, c'est à dire à calculer $B^{-1} r_N$

A.2.3 Dans le but d'éviter la difficulté de résolution de systèmes linéaires, (très encombrants, en pratique), nous avons porté notre attention sur des matrices obtenues à partir des parties triangulaires inférieures associées à A et B, c'est-à-dire :

$$T_{(A)} = D_{(A)} - E_{(A)}$$

$$T_{(B)} = D_{(B)} - E_{(B)}$$

où l'on a utilisé la décomposition additive standard : A = D(A) - E(A) - F(A),

soit

$$A = \begin{pmatrix} & & -F \\ & D & \\ -E & & \end{pmatrix}$$

On pose :

$$M_N = \alpha_N T_{(A)} + \beta_N T_{(B)} \quad ,$$

les paramètres α_N et β_N restant à déterminer. On a alors :

$$H_N = (1-\alpha_N)A - (\mu_N + \beta_N)B - \alpha_N D_{(A)} - \beta_N D_{(B)}$$

Nous avons envisagé trois cas :

(I) $\begin{cases} \alpha_N \geq 1 \\ \beta_N \geq 0 \end{cases}$; - H_N est alors définie positive, et $\{\mu_N\}$ est une suite décroissante

(II) $\begin{cases} \alpha_N \leq 0 \\ \beta_N \leq -\mu_N \end{cases}$ H_N est définie positive et $\{\mu_N\}$ croissante

(III) $\begin{cases} \alpha_N = 1 \\ \beta_N = -\mu_N \end{cases}$ ce cas correspond à la méthode S.O.R. étudiée par A. Ruhe [7].

Dans ce cas : $M_N = - D_{(A)} + \mu_N D_{(B)}$

$$p_N^t H_N p_N = \sum_{i=1}^{N} (b_{ii} \mu_N - a_{ii}) p_{iN}^2$$

Donc, si le vecteur initial est tel que :

$$\mu_0 < \min_i \left[a_{ii}/b_{ii} \right] = \min_i \mu(e_i)$$

(e_i est le i-ème vecteur de la base canonique)

alors $\forall N, \quad p_N^t H_N p_N < 0$

et la suite est décroissante (on a évidemment des résultats analogues si $\mu_0 > \max_i \mu(ei)$).

Les études faites par Ruhe montrent que le comportement assymptotique de la méthode SOR est "comparable" à celui de l'algorithme suivant, étudié par H.R. Schwartz [10] .

x_N donné ; $x_N^1 = x_N$

pour i = 1,...,n

- choisir $\xi_i \in R$ tel que $\mu [x_N^i + \xi_i e_i]$ soit minimum
- $x_N^{i+1} = x_N^i + \xi_i e_i$

$$x_{N+1} = \hat{x}_N^{n+1}$$

Guidés par cette remarque, nous avons cherché un algorithme assymptotiquement équivalent à des minimisations consécutives dans des sous-espaces.

A.2.4 Minimisation de $\mu(x)$ dans un sous-espace.

Soit E_p le sous-espace de R^n engendré par p vecteurs indépendants ($p << n$) et U la matrice dont les colonnes sont constituées par les composantes de ces vecteurs.

$$U = (u_1, \ldots, u_p)$$

x étant donné, nous cherchons $\bar{x}$, somme de x et d'un vecteur de Ep, "optimum", c'est-à-dire tel que le quotient de Rayleigh soit minimum.

$\bar{x} = x + U.\mu \qquad$ où $\mu \in R^p$

$$\text{alors} \quad \mu(\bar{x}) = \frac{x^tAx + \eta^tU^tAx + x^tAU\eta + \eta^tU^tAU\eta}{x^tBx + \eta^tU^tBx + x^tBU\eta + \eta^tU^tBU\eta}$$

Le minimum, Λ est la plus petite valeur propre du problème de taille p+1 suiva

$$\begin{pmatrix} x^tAx & x^tAu_1 & \cdots & x^tAu_p \\ u_1^tAx & u_1^tAu_1 & \cdots & u_1^tAu_p \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ u_p^tAx & u_p^tAu_1 & \cdots & u_p^tAu_p \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_p \end{pmatrix} =$$

$$\Lambda \begin{pmatrix} x^tBx & x^tBu_1 & \cdots & x^tBu_p \\ u_1^tBx & u_1^tBu_1 & \cdots & u_1^tBu_p \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ u_p^tBx & u_p^tBu_1 & \cdots & u_p^tBu_p \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_p \end{pmatrix}$$

qui est simplement le problème projetté dans l'espace engendré par x et E_p. On aurait alors $\eta^t = (y_1, y_2, \ldots, y_p)$, en choisissant (quand cela est possible) un vecteur propre associé tel que $y_0 = 1$.

Ceci est assez proche de ce qui est fait dans la méthode des itérations simultanées, où l'espace approximant est généré par l'utilisation de la méthode de puissance itérée sur (p+1) vecteurs indépendants : $z_0, \ldots, z_p$:

$$\bar{x} = \left[(B^{-1}A)^k Z_0 \ , \ (B^{-1}A)^k Z_1 \ , \ldots , (B^{-1}A)^k Z_p \right] \begin{pmatrix} \eta_0 \\ \eta_1 \\ \cdot \\ \cdot \\ \eta_p \end{pmatrix}$$

Ici, nous cherchons à la place une forme simplifiée, qui soit asymptotiquement (et sous réserve de convergence) "équivalente" à cette minimisation. Signalons, en anticipant sur la rédaction de cette première partie, le chapitre suivant, donnant un sens plus précis à cette affirmation, que si $\hat{x}$ est "proche" d'un vecteur propre v_1 de norme $||v_1||_B = 1$; associé à la valeur propre λ_1, c'est-à-dire :

$$\exists\, v_1 \text{ tq} : (A-\lambda_1 B)v_1 = 0$$

$$||v_1||_B = 1$$

$$||v_1 - \hat{x}||_B = \varepsilon \ ,$$

alors l'erreur commise en prenant $\mu(x)$ comme approximation de λ_1 est d'ordre ε^2.

Ce fait nous donne une approximation du premier ordre de notre problème :

$$[U^T(A-\mu B)U]\eta = -U^T(A-\mu B)x$$

En supposant la matrice $U^t(A-\mu B)U$ régulière, une approximation de $\bar{x}$ est :

$$\bar{x} = [I - U \left(U^t(A-\mu B)U\right)^{-1} U^T (A-\mu B)\ ]\ x$$

Nous allons simplifier cette expression en prenant un cas particulier : On supposera que les U_i sont p vecteurs de la base canonique. Alors, si $\tilde{A}$ et $\tilde{B}$ sont les sous matrices extraites de A et B en ne conservant que les indices relatifs à E_p , avec de même $\tilde{r}_N$ et $\tilde{x}_N$ extraits de r_N et x_N :

$$\tilde{x}^1 - \tilde{x} = -(\tilde{A}-\mu\tilde{B})^{-1}\ \tilde{r}_N$$

A.2.5 Méthodes par blocs

Notons $I\,[k,p[$ pour k et p entiers, tels que $k < p$ un "segment" de $\mathbb{N}$

$$I\,[k,p[\;=\; \{\, m \in \mathbb{N} \;/\; k \le m < p \,\}$$

On partitionnera le segment $I\,[1,n+1[$ en K segments consécutifs et disjoints :

$$I\,[1,n+1[\;=\; \bigcup_{i=1,k} I\,[m_i,m_{i+1}[$$

avec $m_1 = 1$ et $m_{k+1} = n+1$.

On pourra associer un partitionement de R^n en k sous espaces disjoints :

E_s = espace engendré par les vecteurs e_j, pour $j \in I\,[m_s,m_{s+1}[$

$$R^n = E_1 \oplus E_2 \oplus \ldots \oplus E_k$$

Nous allons maintenant donner une itération simple, qui, sous réserve de convergence est "asymptotiquement" équivalente à l'algorithme suivant :

$$x_N^{\,1} = x_N$$

pour $s = 1 \ldots k$:
$$\begin{cases} \mu' = \underset{\phi \in E_s}{\text{Min}}\; [\mu(x_N^{\,s}+\phi)], \text{ réalisé pour } \phi_s \\ x_N^{\,s+1} = x_N^{\,s} + \phi_s \end{cases}$$

$$x_{N+1} = \hat{x}_N^{\,k+1}$$

On associe au partitionnement des indices un découpage en blocs de la matrice $A - \mu B$:

$$A - \mu B = \begin{array}{|c|c|c|c|} \hline D_{11} & -F_{12} & & -F_{1K} \\ \hline -E_{21} & D_{22} & & -F_{2K} \\ \hline & & & \\ \hline -E_{K1} & -E_{K2} & & D_{KK} \\ \hline \end{array} \qquad = \bar{D}_{(A)} - \bar{E}_{(A)} - \bar{F}_{(A)}$$

et des vecteurs :

$$x = \begin{array}{|c|}\hline X_1 \\ \hline X_2 \\ \hline \\ \hline X_k \\ \hline \end{array}$$

D'après ce qui a été vu précédemment, une itération possible est :

* bloc 1 $\qquad \phi \in E_1$

$$\tilde{r} = D_{11} X_1 - \sum_{i>1} F_{1i} X_i$$

$$\tilde{A} - \mu \tilde{B} = D_{11}$$

d'où $\quad D_{11}(X'_1 - X_1) = - D_{11} X_1 + \sum_{i>1} F_{1i} X_i$

à l' issue de cette étape, x est remplacé par le vecteur.

$$\begin{array}{|c|}\hline X'_1 \\ \hline X_2 \\ \hline \\ \hline X_k \\ \hline \end{array}$$

* bloc s : à cette étape, le vecteur en entrée est :

$$x^t = \begin{array}{|c|c|c|c|c|c|c|}\hline X'_1 & X'_2 & & X'_{s-1} & X_s & X_{s+1} & & X_k \\ \hline \end{array}$$

$$\tilde{r} = - \sum_{j<s} E_{sj} X'_j + D_{ss} X_s - \sum_{j>s} F_{sj} X_j$$

$$\tilde{A} - \mu \tilde{B} = D_{ss}$$

d'où $D_{ss} X'_s - \sum_{j<s} E_{sj} X'_j = \sum_{j>s} F_{sj} X_j$

On identifie aisément la forme par blocs de l'algorithme de Gauss-Seidel linéaire :

$$(\bar{D}_{(A)} - \bar{E}_{(A)})X^{0} = \bar{F}_{(A)}\ X$$

A.2.6 Cet algorithme rentre dans le cadre donné précédemment, en prenant M_N combinaison linéaire des parties triangulaires inférieures par blocs issues de A et B :

$$M_N = \alpha_N \bar{T}_{(A)} + \beta_N \bar{T}_{(B)} \qquad \text{où} \quad \bar{T} = -\bar{E} + \bar{D}$$

On utilise la décomposition additive par blocs :

$$A = \bar{D}_{(A)} - \bar{E}_{(A)} - \bar{F}_{(A)}$$

Par la suite, les résultats du paragraphe sur les preuves de convergence étant valables pour les décompositions classiques et par blocs, on omettra de la préciser dans les notations.

La matrice H_N reste définie positive dans les deux cas suivants :

$$\underline{\text{I-B}} \quad \begin{cases} \alpha_N \geq 1 \\ \beta_N \geq 0 \end{cases} \qquad\qquad \underline{\text{II-B}} \quad \begin{cases} \alpha_N \leq 0 \\ \beta_N \leq -\mu_N \end{cases}$$

Le cas : $\underline{\text{III-B}}$:

$$\begin{cases} \alpha_N = 1 \\ \beta_N = -\mu_N \end{cases}$$

correspond à la méthode SOR par blocs.

Montrons que cette méthode génère des suites monotones :

$$\mu_{N+1} - \mu_N = -\frac{1}{||x_{N+1}||_B^2} \sum_{s=1}^{k} p_s^t (A_{ss} - \mu_N B_{ss}) p_s$$

Donc, si le vecteur initial a un quotient de Rayleigh inférieur à la plus petite valeur propre des problèmes blocs diagonaux, c'est-à-dire si :

$$\mu_0 < \min_{s=1\ldots k} \left[\min_{\phi \in E_s} \frac{\phi^t A \phi}{\phi^t B \phi} \right]$$

Alors, la suite est bien monotone décroissante, puisque tous les μ_N vérifieront, à fortiori, la condition énoncée.
Cette condition a pour conséquence également la régularité de tous les blocs diagonaux dans la suite de l'itération.

On se limitera pour la suite de l'étude théorique de cette première partie aux algorithmes I, II, SOR, ainsi qu'aux versions par blocs de ces algorithmes.

A.2.7 Versions "relaxées" de ces algorithmes :

Elles sont données par la condition :

$$(\alpha_N D_{(A)} + \beta_N D_B)(x_{N+1} - x_N) = w\,[(\alpha_N E_A + \beta_N E_B)(x_{N+1} - x_N) - (A - \mu_N B)x_N]$$

nous avons alors $T_A(w) = \frac{1}{w} D_A - E_A$; $T_B(w) = \frac{1}{w} D_B - E_B$

$$H_N = (1-\alpha_N)A - (\mu_N + \beta_N)B - \alpha_N(\frac{2}{w} - 1)D_A - \beta_N(\frac{2}{w} - 1)D_B$$

Les résultats précédents restent donc valables, pourvu que

$$\frac{2}{w} - 1 \geq 0 \qquad \text{soit} \quad w \in]0,2]$$

A.2.8 Calcul d'un vecteur initial

Comme nous venons de le voir, seules les méthodes SOR et SOR par blocs posent des conditions sur le vecteur initial. Nous proposons ici une technique pour initialiser ces algorithmes.

On commence par calculer $\bar{x}$ tel que :

$$\mu(\bar{x}) = \min_{s=1\ldots k} \{ \min_{x \in Es} \mu(x) \}$$

Il faut donc résoudre les k sous-systèmes extraits :

$$A_{ss} \tilde{X}_s = \lambda B_{ss} \tilde{X}_s$$

et garder le vecteur donnant le plus petit quotient de Rayleigh. Le calcul reste très raisonnable si on se limite à des blocs de petite taille (deux ou trois).

A ce stade, on a pas encore la condition d'inégalité stricte demandée, et un des blocs diagonaux de la matrice $M(\mu_0)$ est singulier.

Soit $D_{kk} = A_{kk} - \mu_0 B_{kk}$ ce bloc .

La détermination de $\tilde{p}_k$ est alors impossible. Illustrons ce fait dans le cas où les blocs sont de taille deux :

$$\begin{pmatrix} \sum_{\ell<i} (a_{i\ell} - \mu_0 b_{i\ell}) p_\ell \\ \\ \sum_{\ell<i} (a_{i+1,\ell} - \mu_0 b_{i+1,\ell}) p_\ell \end{pmatrix} + D_k \begin{pmatrix} p_i \\ \\ p_{i+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}$$

L'idée que nous avons utiliséeconsiste à remplacer D_k par $D_k + \rho I$ où ρ est un paramètre positif tel que $D_k + \rho I$ soit inversible.

L'accroissement initial est alors donné par :

$$[\ (D_A - E_A) - \mu_0 (D_B - E_B) + \rho [e_{ii} + e_{i+1,i+1}]\ ]\ p_0 = - r_0$$

ce qui conduit à :

$$\mu_1 - \mu_0 = \frac{-1}{||x_1||_B^2} [\ p_0^t\ (D_A - \mu_0 D_B)\ p_0 + 2\ \rho (p_{0i}^2 + p_{0i+1}^2)\ ]$$

x_1 est alors un bon vecteur de départ, puisque $\mu_1 < \mu_0$

(B) PREUVES DE CONVERGENCE

L'objet de ce chapitre est d'essayer de montrer que les algorithmes (I),(II) et SOR, ainsi que les versions relaxées et versions par blocs leur sont associées produisant des suites convergentes vers la plus petite ou la plus grande des valeurs propres du problème généralisé, ceci se détaillant comme suit :

convergence vers λ_1 :

* algorithmes (I)
* " (I.B)
* SOR si $\mu_0 < \min a_{ii}/b_{ii}$
* SOR-blocs si $\mu_0 < \underset{s}{\text{Min}} \left[\underset{\phi \epsilon Es}{\text{Inf}} \frac{\phi^t A \phi}{\phi^t B \phi} \right]$

convergence vers λ_N :

* algorithmes II
* " II.B
* SOR si $\mu_0 > \max a_{ii}/b_{ii}$
* SOR-blocs si $\mu_0 > \underset{s}{\text{Max}} \left[\underset{\phi \epsilon Es}{\text{Sup}} \left(\frac{\phi^t A \phi}{\phi^t B \phi} \right) \right]$

Les résultats étant analogues dans les deux cas, on ne parlera plus que des algorithmes générant des suites décroissantes.

Les suites étant bornées inférieurement, par λ_1 , elles convergent vers une limite, que nous noterons $\underline{\mu}$

Montrons d'abord le lemme suivant, déjà démontré par Ruhe dans le cas de la méthode SOR.

Lemme B.1

Si les suites α_N et β_N restent bornées supérieurement, alors $\| \tilde{r}_N \|_2 \to 0$

Démonstration $\quad \mu_{N+1} - \mu_N \to 0 \quad$ donc $\dfrac{|p_N{}^t H_N p_N|}{\|x_{N+1}\|_B^2} \to 0$

* montrons d'abord que : $\|p_N\|_2^2 \Big/ \|x_{N+1}\|_B^2 \to 0$

pour cela, on distinguera deux cas :

a) Méthode (I) et (I.B)

$$|p_N^t H_N p_N| = (\alpha_N - 1)||p_N||_A^2 + (\mu_N + \beta_N)||p_N||_B^2 + (\frac{2}{\omega} - 1)\beta_N||p_N||_{D_B}^2$$

$$+ (\frac{2}{\omega} - 1)\alpha_N ||p_N||_{D_A}^2$$

tous les coefficients sont, en effet, positifs.

alors, $|p_N^t H_N p_N| \geq \lambda_1 ||p_N||_B^2 \geq \phi||p_N||_2^2$

b) Méthodes SOR par blocs :

$$|p_N^t H_N p_N| = (\frac{2}{\omega} - 1)\left|\sum_{s=1}^{k} \tilde{p}_{Ns}^t (\tilde{A}_{ss} - \mu_N \tilde{B}_{ss})\tilde{p}_{Ns}\right|$$

$$\geq (\frac{2}{\omega} - 1)\left|\sum_{s=1}^{k} \tilde{p}_{Ns}^t (\tilde{A}_{ss} - \mu_0 \tilde{B}_{ss})\tilde{p}_{Ns}\right|$$

avec l'hypothèse sur μ_0 , $\exists\ \phi > 0$ telle que

$$\forall\ s = 1,\ldots,k \quad , \quad \tilde{p}_{Ns}^t (\tilde{A}_{ss} - \mu_0 \tilde{B}_{ss})\tilde{p}_{Ns} \geq \phi||\tilde{p}_{Ns}||_2^2$$

finalement $|p_N^t H_N p_N| \geq \rho(\frac{2}{\omega} - 1)||p_N||_2^2$

* montrons que $\dfrac{||p_N||_2^2}{||x_N||_B^2} \to 0$

en effet, $||x_N||_B \geq ||x_{N+1}||_B - ||p_N||_B$

soit $\dfrac{||x_N||_B}{||x_{N+1}||_B} \geq 1 - \dfrac{||p_N||_B}{||x_{N+1}||_B}$

nous avons vu que $\dfrac{||p_N||_B}{||x_{N+1}||_B} \to 0$

donc, certainement, à partir d'un certain rang :

$$\frac{||x_N||_B}{||x_{N+1}||_B} \geq \frac{1}{2}$$

d'où $\frac{1}{2}\frac{||p_N||_2}{||x_N||_B} \leq \frac{||p_N||_2}{||x_{N+1}||_B}$ donc $\lim_{N\to\infty} \frac{||p_N||_2}{||x_N||_B} = 0$

* passons maintenant à la convergence de $||\tilde{r}_N||$:

$r_N = - M_N p_N$ $\qquad ||r_N||_2 \leq ||M_N||_{22} ||p_N||_2$

avec $||M_N||_{22} = \sup_{||x||_2=1} ||M_N x||_2$

$||M_N|| \leq |\alpha_N| \; ||T_A||_{22} + |\beta_N| \; ||T_B||_{22}$

donc, avec l'hypothèse que α_N et β_N restent bornés, on a :

$||(A - \mu_N B)\hat{x}_N||_2 \to 0$ $\qquad$ soit $||\tilde{r}_N||_2 \to 0$ $\qquad$ C.Q.F.D.

La suite $\hat{x}_N$ est sur la frontière de la boule unité de R^n, (muni de la norme $||\;||_B$), qui est compacte. Nous pouvons donc en extraire une sous suite x_{N_k} convergente :

$\hat{x}_{N_k} \to v$ $\qquad$ et $\qquad ||v||_B = 1$

alors $(A - \mu_{N_k} B)\,\hat{x}_{N_k} \to (A - \underline{\mu} B)\, v$

Lemme B.2

$$(A - \underline{\mu} B) v = 0$$

$(\forall \varepsilon) > 0$ $\quad$ on montre que $||(A - \underline{\mu} B) v|| < \varepsilon$:

$(A - \underline{\mu} B) v = (A - \mu_{N_k} B)\hat{x}_{N_k} + (\mu_{N_k} - \underline{\mu}) B \hat{x}_{N_k} + (A - \underline{\mu} B)(v - x_{N_k})$

Nous savons que :

* K_1 tel que $k \geq K_1 \rightarrow ||(A - \mu_{N_k} B)\hat{x}_{N_k}|| = ||\tilde{r}_{N_k}|| < \frac{\varepsilon}{3}$

* K_2 tel que $k \geq K_2 \rightarrow |(\mu_{N_k} - \underline{\mu})| \cdot ||B\hat{x}_{N_k}|| < \frac{\varepsilon}{3}$

* K_3 tel que $k \geq K_3 \rightarrow ||(A - \underline{\mu}B|| \cdot ||v - \hat{x}_{N_k}|| < \frac{\varepsilon}{3}$

donc, pour $k \geq \sup [K_1, K_2, K_3]$ on a la majoration cherchée.

Ceci montre que $\underline{\mu}$ est une valeur propre du problème soit λ_i. Une étude assymptotique permettra de montrer que, en fait $\lambda_i = \lambda_1$

On aura par la suite besoin du théorème suivant, conséquence de la théorie des perturbations d'opérateurs symétriques : (voir [2]pour le cas $Ax = \lambda x$)

Théorème B.3

Soient δ un réel quelconque et x un vecteur de R^n

Soient λ_i une valeur propre du problème généralisé (à matrices symétriques)*

et $|V|$ le sous espace propre associé :

$$|V| = \text{Ker } [A - \lambda_i B]$$

On notera $\sigma(A,B)$ l'ensemble des valeurs propres du problème, et d la quantité :

$$d = \text{dist } [\delta, \left(\sigma(A,B) \setminus \lambda_i \right)] = \min_{\substack{\lambda_j \neq \lambda_i \\ \lambda_j \in \sigma(A,B)}} |\delta - \lambda_j|$$

et $\theta\left(x, |V|\right)$ l'angle entre x et $|V|$, défini par :

$$\theta(x, |V|) = \text{Arc sin} \left(\min_{y \in |V|} \frac{||x-y||}{||x||} \right)$$

On a alors la majoration suivante :

$$\sin \theta(x, |V|) \leq C \frac{||(A - \delta B)x||_2}{||x||_2} \cdot \frac{1}{d}$$

où C est une constante qui ne dépend que de B.

---

* B étant définie positive.

Une conséquence immédiate de ce théorème est que, si λ_i est simple alors

$$\hat{x}_N \to V_i$$

La démonstration se fera en plusieurs parties : on rappelera d'abord deux lemmes, classiques (Kato [2]).

Lemme B.3.1

Soit T un opérateur linéaire continu, autoadjoint de M, espace de Hilbert dans lui-même.

alors $$||T|| = \sup_{z \in \sigma(T)} |z|$$

et, si $z \in \rho(T)$, (ensemble résolvant), alors

$$||R_{(z)}|| = ||(T-z)^{-1}|| = \frac{1}{dist(z,\sigma(T))}$$

(nous notons 1 l'opérateur identité dans H).

Soit alors λ_0 une valeur propre isolée de T, et soit Γ un contour fermé dans l'ensemble résolvant de T, tel que Γ entoure λ_0 et l'isole du reste du spectre. Alors

$$P_{\lambda_0} = \frac{-1}{2i\pi} \int_\Gamma R(z)\, dz$$

est la projection sur le sous espace invariant associé à λ_0 . P_{λ_0} est autoadjoint et c'est une projection orthogonale.

On a alors le :

Lemme B.3.2

Pour tout réel δ , et tout $x \in M$

$$dist\left(\delta,\ (\sigma(T) \setminus \lambda_0)\right) .\ ||(1- P_{\lambda_0})x|| \leq ||(T-\delta)x||$$

où 1 est l'opérateur identité de H.

Nous pouvons ensuite passer à la démonstration :

on appliquera le dernier résultat à $T = L^{-1}A(L^{-1})^t$ avec $B = LL^t$

d'abord $(A-\lambda_i B)y = 0 \iff (T-\lambda_i)L^t y = 0$

On cherchera à borner la quantité :

$$\sin\ \Theta(y,|V|) = \underset{z\in|V|}{\text{Min}} \frac{||y-z||}{||y||}$$

Si P désigne la projection spectrale associée à la valeur propre λ_i , de la matrice $T = L^{-1}\ A(L^{-1})^t$

alors, il est facile de vérifier que :

$\forall y,\quad ((L^{-1})^t P\ L^t)y \in |V|$

donc $\sin\Theta(y,|V|) \leq \dfrac{||(L^{-1})^t\ (1-P)L^t\ y||}{||y||}$

$$\leq \frac{||(L^{-1})^t||}{d} \cdot \frac{||(T-\delta)L^t\ y||}{||y||}$$

$$\leq \frac{||(L^{-1})^t||}{d} \cdot \frac{||L^{-1}(A(L^{-1})^t-\delta L)L^t\ y||}{||y||}$$

$$\leq \frac{||(L^{-1})^t||\,||L^{-1}||}{d} \cdot \frac{||(A-\delta B)y||}{||y||}$$

Q.E.D.

Considérons une base B orthonormale de $|V|$, et soit V la matrice formée par les vecteurs de cette base :

$$V^t\ B\ V = I_k$$

alors $\overset{o}{P} = V\ V^t\ B$ est une projection

B - orthogonale sur $|V|$:

1) $V\ V^t\ B\ x \in |V|$ comme combinaison linéaire de vecteurs de $|V|$

2) $\overset{o2}{P} = V\ \underbrace{V^t\ B\ V}\ V^t\ B = V\ V^t\ B = \overset{o}{P}$

$= I_k$

3) $(\overset{o}{P}x\ ,\ (I_k - \overset{o}{P})y)_B = (x^t B\ V\ V^t)B(I_k - V\ V^t\ B)y = 0$

Considérons la quantité suivante :

$$\varepsilon(x) = \frac{||(I_k - \overset{o}{P})x||}{||x||}$$

Le théorème précédent nous apprend que :

$$\varepsilon(x_N) \to 0$$

Avant d'aller plus loin, nous allons formuler une hypothèse sur le comportement des suites α_N et β_N : (il s'agit en fait de fonctions de x_N) :

$$\alpha_N = \alpha(x_N) \quad ; \quad \beta_N = \beta(x_N))$$

Hypothèse (H) :

$$\lim_{N\to\infty} \alpha_N = \alpha$$

$$\lim_{N\to\infty} \beta_w = \beta$$

(limites qui ne dépendent que de λ_i).

le convergence ayant lieu à une vitesse au moins égale à celle de $\varepsilon(x_N)$.

Plus précisément, il existe C constante positive et un rang M tels que :

$$N > M \to \begin{cases} |\alpha_N - \alpha| < C \ \varepsilon(x_N) \\ |\beta_N - \beta| < C \ \varepsilon(x_N) \end{cases}$$

Cette hypothèse est très peu restrictive, puisque dans la pratique, on se limitera à prendre α_N et β_N constants pour les algorithmes I et II, et qu'elle se trouve naturellement vérifiée pour les méthodes SOR et SOR par blocs.

Le théorème suivant s'inspire d'un résultat donné par A. Ruhe pour la méthode SOR.

On posera : $$C(\lambda_i) = I - (\alpha T_A + \beta T_B)^{-1} \ (A - \lambda_i B)$$

Théorème B.4

Il existe une constante $\eta > 0$ et une fonction F , $R^n \to R^n$ dont la norme est homogène et telle que

$$\varepsilon(x) < \eta \Rightarrow ||F_{(x)}|| < C(\varepsilon_{(x)})^2 \ . \ ||x||$$

pour une certaine constante $C > 0$ (indépendante de x) ;

cette fonction étant telle que :

$$x \in |V| \rightarrow F(x) = 0$$

avec : pour tout x_0 tel que $\varepsilon(x_0) < \mu$

alors , $x_1 = \left[I - M_0^{-1} (A - \mu_0 B) \right] x_0$

vérifie :

$$x_1 = C(\lambda_i)\, x_0 + F(x_0)$$

---

démonstration : $x_1 - C(\lambda_i)\, x_0 = (\mu_0 - \lambda_i) M^{-1} B x_0 + (M^{-1} - M_0^{-1})(A - \mu_0 B) x_0$

$$= F_1(x_0) + F_2(x_0)$$

* le premier terme est facile à borner :

$$||F_1(x_0)|| = ||(\mu_0 - \lambda_i) M^{-1} B x_0|| \leq |(\mu_0 - \lambda_i)| \cdot ||M^{-1}|| \cdot ||B x_0||$$

$$|\mu_0 - \lambda_i| = \frac{x_o^t A x_o}{x_o^t B x_o} - \frac{x_o^t \overset{o}{P}^t A \overset{o}{P} x_o}{x_o^t \overset{o}{P}^t B \overset{o}{P} x_o} \leq \frac{1}{x_o^t B x_o} \left[x_o^t A x_o - x_o^t \overset{o}{P}^t A \overset{o}{P} x_o \right]$$

mais $x^t (I - \overset{o}{P})^t A \overset{o}{P} x = x^t (I - \overset{o}{P})^t B \overset{o}{P} x \cdot \lambda_i = 0$

il reste $$|\mu_0 - \lambda_i| \leq \frac{x_0^t (I - \overset{o}{P})^t A (I - \overset{o}{P}) x_0}{x_0^t B x_0}$$

$$\leq C \, \varepsilon(x_0)^2$$

et $$||F_1(x_0)|| \leq C \cdot ||M^{-1}|| \cdot ||B|| \quad \varepsilon(x_0)^2 \cdot ||x_0||$$

* $||F_2(x_0)|| \leq ||M^{-1} - M_0^{-1}|| \cdot ||(A - \mu_0 B) x_0||$

$$||r_0|| = ||(A - \mu_0 B) x_0|| \leq ||(A - \lambda_i B) x_0|| + |(\lambda_i - \mu_0)| \; ||B x_0||$$

$$||(A-\lambda_i B)x_0|| = ||(A-\lambda_i B)(I-\overset{\circ}{P})x_0||$$

$$\leq ||A-\lambda_i B|| \cdot \varepsilon(x_0) \cdot ||x_0||$$

d'où $||r_0|| \leq C_1 \quad \varepsilon(x_0) \cdot ||x_0||$

On va borner $M^{-1} - M_0^{-1}$: $M^{-1} - M_0^{-1} = M_0^{-1}(M_0 - M) M^{-1}$

$$||M^{-1} - M_0^{-1}|| \leq ||M_0^{-1}|| \cdot ||M_0 - M|| \; ||M^{-1}||$$

$$||M_0^{-1}|| \leq \frac{||M^{-1}||}{1 - ||M^{-1}|| \cdot ||M-M_0||}$$

donc $||M^{-1} - M_0^{-1}|| \leq \dfrac{||M^{-1}||^2 \cdot ||M-M_0||}{1-||M^{-1}|| \cdot ||M-M_0||}$

Or $M-M_0 = (\alpha - \alpha_0)T_A + (\beta - \beta_0)T_B$

$$||M-M_0|| \leq C\, \varepsilon(x_0)\, [||T_A|| + ||T_B||]$$

à partir d'un certain rang, on a certainement :

$$1 > 1 - ||M^{-1}|| \cdot ||M-M_0|| > \frac{1}{2}$$

d'où finalement :

$$||M^{-1} - M_0^{-1}|| \leq 2\, C\, \varepsilon(x_0) \cdot ||M^{-1}||^2 \cdot [||T_A|| + ||T_B||]$$

et $||(M^{-1} - M_0^{-1})\, r_0|| \leq C' \cdot \varepsilon(x_0)^2 \cdot ||x_0||$

ce qui achève la démonstration.

B. 5 Etude de l'itération linéaire

$$x_{N+1} = [\, I - M^{-1}(A-\lambda_i B)\,]\, x_N$$

$$C(\lambda_i)$$

D'après la régularité de la matrice M, les points fixes de l'itération linéaire sont les vecteurs propres du problème généralisé associés à la valeur propre $\lambda = \lambda_i$. En particulier 1 est valeur propre de $C(\lambda_i)$.

Nous allons donner deux propriétés de la matrice $C(\lambda_i)$, déjà énoncées par HR. Schwartz [10] au sujet de la matrice relative à la méthode SOR.

Lemme B.5.1

La valeur propre 1 de la matrice $C(\lambda_i)$ est semi-simple ; autrement dit, le bloc associé à cette valeur propre dans la forme de Jordan de C est diagonal.

En effet, s'il en était autrement, il existerait un vecteur h_0 tel que :

$$h_0 \in \text{Ker } (C-I)^2$$

$$h_0 \notin \text{Ker } (C-I)$$

Nous aurions donc : $C\, h_0 - h_0 = V_i$, vecteur propre du problème généralisé associé à la valeur propre λ_i .

En posant $h_1 = C\, h_0$

on a : $h_1 - h_0 = V_i$ donc $(A - \lambda_i B)(h_1 - h_0) = 0$

Il en résulte que :

$$h_1^t (A- \lambda_i B)h_1 = h_1^t (A- \lambda_i B)h_0 = h_0^t (A- \lambda_i B)h_1 = h_0^t (A- \lambda_i B)h_0$$

alors

$$0 = h_1^t (A- \lambda_i B)h_1 - h_0^t (A - \lambda_i B)h_0 = (h_1 - h_0)^t (A- \lambda_i B)(h_1 - h_0)$$

$$+ 2(h_1-h_0)^t (A- \lambda_i B)\, h_0 .$$

En utilisant : $[I - M^{-1} (A- \lambda_i B)]\, h_0 = h_1$

$$(A- \lambda_i B)\, h_0 = -M (h_1-h_0) ,$$

On a finalement :

$$0 = (h_1 - h_0)^t \; [A - \lambda_i B - (M+M^t)] \; (h_1 - h_0),$$

Ce qui implique que :

$$h_1 = h_0$$

$$(C - I)h_0 = 0$$

$$h_0 \in \text{Ker } (C-I) \; ,$$

ce qui est en contradiction avec les hypothèses.

Lemme B.5.2

pour $\lambda_i \neq \lambda_1$ le rayon spectral de $C(\lambda_i)$ est supérieur à 1.

Soit ϕ_0 le vecteur défini par :

$$\phi_0 = V_1 + V_i \qquad \text{où} \qquad A\,V_1 = \lambda_1 BV_1 \qquad ||V_1||_B = 1$$

$$A\,V_i = \lambda_i BV_i \qquad ||V_i||_B = 1$$

On a $(V_1 \; , \; V_i)_B = 0$

et $\mu(\phi_0) = \frac{1}{2}(\lambda_i + \lambda_1) < \lambda_i$

Alors la suite définie par la recurrence suivante :

$$\phi_0$$

$$\phi_{N+1} = [\, I - M^{-1} \; (A - \lambda_i B) \,] \; \phi_N$$

est telle que :

$$\mu(\phi_{N+1}) < \mu(\phi_N) < \ldots\ldots < \mu(\phi_0) < \lambda_i \qquad (2)$$

puisque $M = A - \lambda_i B - (M+M^t)$ est définie négative.

En supposant l'hypothèse non vérifiée, on aurait $\rho(C(\lambda_i)) = 1$

et, d'après le lemme B 5.1 :

$$\phi_N \to V_i \qquad \text{et} \qquad \mu_N \to \lambda_i$$

Ce qui est en contradiction avec (2).

Donc nécessairement $\rho\,[C(\lambda_i)] > 1$ si $\lambda_i \neq \lambda_1$

A cette étape, nous pouvons faire le point sur la situation.

Les méthodes considérées produisent des suites de vecteurs dont les quotients de Rayleigh convergent vers une valeur propre λ_i. Parallèlement, l'erreur sur les vecteurs, mesurée par la quantité

$$C(x) = \frac{||(I-\overset{o}{P}_{\lambda_i})x||}{||x||}$$

tend vers zero. Enfin, l'itération réelle diffère de l'itération linéaire :

$$x_{N+1} = C\,(\lambda_i)\,x_N$$

par une quantité d'ordre deux en $\varepsilon(x_N)$.

La matrice $C(\lambda_i)$ étant telle que :

$$\rho\,[C(\lambda_i)] > 1 \qquad \text{si} \quad \lambda_i \neq \lambda_1$$

l'itération linéaire ne peut (pratiquement) pas converger vers un vecteur du sous-espace propre $|V|$ associé à λ_i.

Pour pouvoir dire si la convergence de l'itération réelle vers $\lambda_i \neq \lambda_1$ est impossible ou non, il est nécessaire d'approfondir les relations entre les deux itérations.

C'est l'objet du paragraphe suivant :

B.6

Soit l'itération suivante :

$$x_{N+1} = C\,x_N + F(x_N)$$

où 1 est valeur propre,semi simple, de C avec un sous espace propre associé $|V|$
Ce sous espace est tel que :

$$x \in |V| \rightarrow F(x) = 0$$

P étant une projection sur $|V|$, l'erreur est mesurée par

$$\varepsilon(x) = \frac{||(I-P)x||}{||x||}$$

Nous savons qu'il existe des constantes positives η et K telles que :

$$\varepsilon(x) < \eta \qquad \frac{||F_{(x)}||}{||x||} \leq \varepsilon(x)^2 \cdot K$$

cette propriété restant vraie pour toutes les normes de R^n.

Nous allons essayer de répondre aux questions suivantes concernant l'espace $|V|$:

1) si $\rho(C) = 1$ — y a t'il convergence de l'itération réelle dans un secteur "proche" de $|V|$?

2) si $\rho(C) > 1$ — est-il possible d'exclure une convergence vers un élément de $|V|$? et que peut-on dire du caractère "répulsif"de $|V|$?

B.6.1 sous les hypothèses précédentes, et si $\rho(C) = 1$

alors l'espace $|V|$ est attractif :

$\tau > 0$ tel que si $\varepsilon(x) < \tau$ alors, l'itération :

$$x_0 = x$$

$$x_{N+1} = C\, x_N + F(x_N)$$

vérifie $\lim_{N \to \infty} \varepsilon(x_N) = 0$

Considérons la matrice $M = (1-P)\ C\ (1-P)$

son rayon spectral est :

$$\rho[M] = \max_{\mu_i \in [\rho[C] \setminus 1]} |\mu_i| < 1$$

donc il existe une norme ϕ sur R^n telle que :

$$\rho(M) < S_{\phi\phi}(M) < \frac{1}{2}\,[\rho(M) + 1] < 1$$

C'est la norme que nous allons utiliser :

$$\varepsilon_\phi(x) = \frac{||(I-P)x||_\phi}{||x||_\phi}$$

Nous allons montrer que $\exists\ \tau > 0$ et $C \in\ ]0,1[$ tels que :

$$\varepsilon(x_0) < \tau \rightarrow \varepsilon_\phi[C\mathcal{X}_o + F_{(x_0)}] \leqslant C\ \varepsilon_\phi(x_0)$$

posons : $x_1 = C\mathcal{X}_o + F_{(x_0)}$

$$\varepsilon_\phi(x_1) \leq \frac{||(1-P)\ C\mathcal{X}_o||_\phi + ||(1-P)F_{(x_0)}||_\phi}{\left|\,||C\mathcal{X}_o||_\phi - ||F_{x_0)}||_\phi\right|}$$

$$\varepsilon_\phi(x_1) \leq \frac{\dfrac{||(1-P)C\ (1-P)^2 x_0||_\phi}{||x_0||_\phi} + K\ ||1-P||_{\phi\phi}\ \varepsilon(x_0)^2}{\left|\dfrac{||C_{x_0}||_\phi}{||x_0||_\phi} - K\ \varepsilon_{(x_0)^2}\right|}$$

Nous allons borner inférieurement $\dfrac{||C_{x_0}||_\phi}{||x_0||_\phi}$

$$\frac{||C\ x_0\ ||_\phi}{||x_0||_\phi} \geq \left|\frac{||CP\ x_0||_\phi}{||x_0||_\phi} - \frac{||C(1-P)x_0||_\phi}{||x_0||_\phi}\right| = \left|\frac{||P_{x_0}||_\phi}{||x_0||_\phi} - S\phi\phi\ (C).\ \varepsilon_\phi(x_0)\right|$$

comme $\dfrac{||P_{x_0}||_\phi}{||x_0||_\phi} \geq 1 - \varepsilon_\phi(x_0)$

Il est licite de ne plus mettre de valeur absolue, puisque

$$\frac{||P_x\ ||_\phi}{||x_0||_\phi} \text{ et } \frac{||C_{x_0}||_\phi}{||x_0||_\phi} \rightarrow 1 \qquad \text{si} \quad \varepsilon_\phi(x_0) \rightarrow 0$$

Il reste finalement :

$$\varepsilon_\phi(x_1) \leq \frac{S_{\phi\phi}(M) \cdot \varepsilon_\phi(x_0) + K \cdot ||\cdot - P||_{\phi\phi} \quad \varepsilon_\phi(x_0)^2}{1 - [1+S_{\phi\phi}(C)]\ \varepsilon_\phi(x_0) - K\ \varepsilon_\phi(x_0)^2}$$

comme $S_{\phi\phi}(M) < 1$, il est facile de trouver $\chi_1 > 0$ tel que, par exemple :

$$\varepsilon_\phi(x_0) < \chi_1 \rightarrow \varepsilon_\phi(x_1) < \frac{1}{2}[1 + S_{\phi\phi}(M)] \cdot \varepsilon_\phi(x_0)$$

Il en résulte que : $\varepsilon(x_N) \rightarrow 0$

Le cas où $\rho(C) > 1$ est beaucoup plus délicat. On a d'abord étudié le problème dans lequel C est une matrice symétrique :

* dans R^2, il n'y a pas de difficulté : par un changement de base approprié, on se ramènera à l'exemple suivant :

$$U_{n+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1+k \end{pmatrix} U_n + F(U_n)$$

avec $U_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix}$

On a alors le résultat suivant :

$\exists\, \eta > 0$ tel que si $\varepsilon_{(U_0)} = \dfrac{y_0}{[x_0^2 + y_0^2]^{1/2}} \leq \eta$ [1]

alors $\exists$ K tel que $\varepsilon(u_K) > \eta$

En effet, si U_0 vérifie la condition [1] , alors :

$$\frac{y_1}{x_1} \geq \frac{|(1+k)y_0| - \varepsilon_0^2 K ||U_0||}{|x_0| + \varepsilon_0^2 K ||U_0||}$$

comme $\varepsilon_{(U_0)} \leq \left|\dfrac{y_0}{x_0}\right| \leq \dfrac{|y_0|}{|x_0^2+y_0^2|^{1/2}} \sqrt{2} \leq \sqrt{2}\ \varepsilon_{(U_0)}$

Nous pouvons borner : $\left| \frac{y_1}{x_1} \right| \geq \phi(\varepsilon_0) \left| \frac{y_0}{x_0} \right|$

avec $\phi(\varepsilon_0) = \frac{1+k-K\,\varepsilon_o}{1+\sqrt{2}\,K\,\varepsilon_o^2}$

comme $\lim_{\varepsilon \to 0} \phi(\varepsilon) = 1+k$ avec $k > 0$

Il est facile de trouver μ tel que si $\varepsilon_0 < \mu$ $\phi(\varepsilon_0) > C > 1$

(avec, par exemple, $C = 1 + \frac{k}{2}$)

Il suffit d'appliquer récursivement ce résultat tant que $\varepsilon(x_N) < \mu$

$$\left| \frac{y_N}{x_N} \right| > C^N \left| \frac{y_0}{x_0} \right|$$

d'où le résultat.

---

* passons maintenant au cas d'une dimension supérieure :

C est une matrice symétrique dans R^n avec $n > 2$.

On introduira les projections orthogonales suivantes :

$P : R^n \to E(1)$ dans le sous-espace propre associé à la valeur propre 1

P_μ projection sur le sous espace, associé à la valeur propre de plus grand module μ ($|\mu| > 1$).

Ici, la présence éventuelle de valeurs propres inférieures à 1 nous a contraint à imposer une condition supplémentaire, indiquant que la composante $P_\mu x$ n'est pas "trop petite". Le résultat peut s'énoncer de la manière suivante :

Lemme B.6.2

Pour tout réel $Q \in]0,1]$, on peut trouver $\eta > 0$ et $\alpha > 0$ tels que :

tout x_0 vérifiant : $\begin{cases} \varepsilon(x_0) < \eta \\ \text{et} \quad \frac{||P_\mu(x_0)||}{||P(x_0)||} \geq Q\,[\,\varepsilon(x_0) - \alpha\,\varepsilon^2(x_0)\,] \end{cases}$

est tel que l'itération : $x_{N+1} = C\,x_N + F_{(x_N)}$

vérifie : $\exists K$ tel que : $\varepsilon(x_K) > \eta$

Avant d'en donner une démonstration, essayons d'interprêter le résultat :

On définira une représentation plane de l'erreur de la manière suivante :

$$\hat{x}_0 = \frac{x_0}{||P_{(x_0)}||} \qquad (P_{(x_0)} \text{ supposé} \neq 0)$$

en abscisse, on portera : $||(1-P_\nu)(1-P)\hat{x}_0||$

en ordonnée : $||P_\nu\,\hat{x}_0|| = ||P_\nu\,(1-P)\hat{x}_0||$

Ce qui donne dans le plan x 0 y un point représentatif $M(x_0)$.

Alors $\varepsilon_{(x_0)} = \dfrac{||(1-P)\hat{x}_0||}{||\hat{x}_0||} = \dfrac{\overline{OM}_{(x_0)}}{||\hat{x}_0||}$

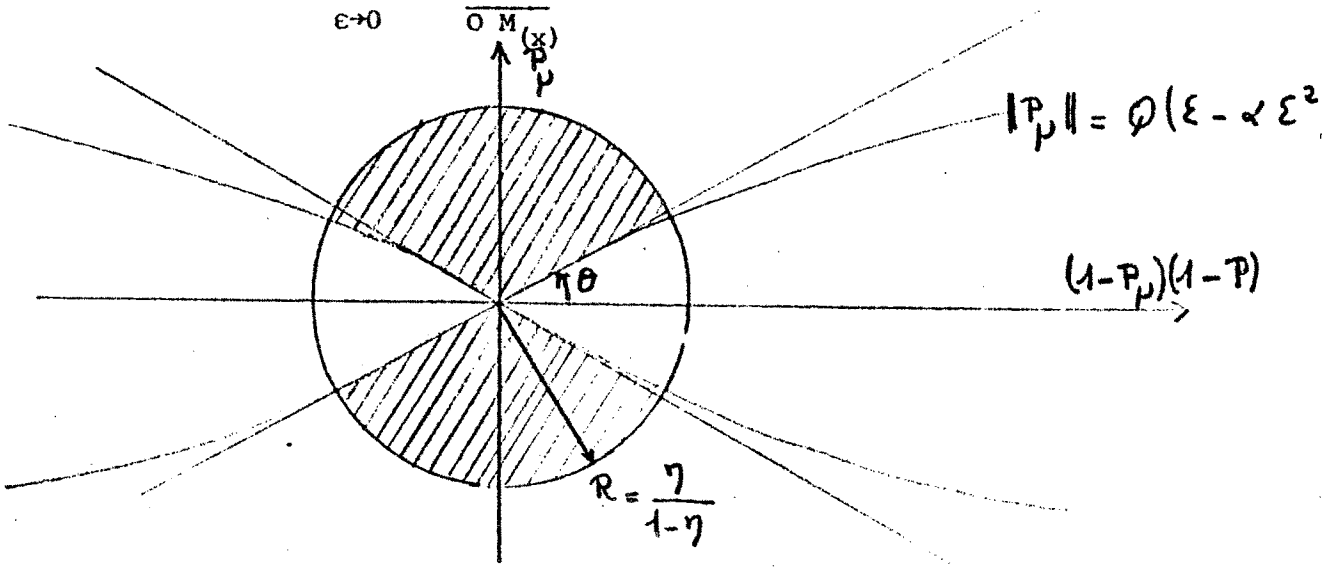
$$\overline{OM}_{(x_0)} = \frac{\varepsilon(x_0)}{1-\varepsilon(x_0)}$$

On représentera schématiquement la courbe :

$$||P_\mu\,(\hat{x})|| = \varrho\,(\varepsilon(x) - \alpha\,\varepsilon(x)^2)$$

qui présente à l'origine une tangente telle que :

$$\sin\Theta = \lim_{\varepsilon\to 0} \frac{||P_\mu(x)||}{\overline{O\,M}_{(x)}}$$



Donc, pour tout $\theta \in]0, \frac{\pi}{2}[$, il existe un rayon $R(\theta)$ tel que pour tout x_0 dont le point représentatif $M(x_0)$ appartienne au secteur hachuré

l'itération x_0

$$x_{N+1} = C x_N + F(x_N)$$

vérifie $\exists K$ tel que $\overline{OM}_{(x_K)} > R$

Ceci n'exclut pas la possibilité d'une convergence vers 0. Mais :

- la convergence doit être telle que la "trajectoire" de $M_{(x_N)}$ soit tangente en 0 à l'axe $x' 0x$
- toute perturbation $\bar{x} = x + \delta x$ de x telle que $P_\mu(\bar{x}) \neq 0$ entraîne une croissance de l'erreur.

Passons maintenant à la démonstration : On cherchera à montrer que $\exists$ A, constante > 1 telle que si x_0 satisfait aux hypothèses, alors :

$$\frac{||P_\mu x_1||}{||P x_1||} \geq A \frac{||P_\mu x_0||}{||P x_0||} \quad \text{et} \quad \frac{||P_\mu x_1||}{||P x_1||} \geq Q\left[\varepsilon(x_1) - \alpha \varepsilon(x_1)^2\right]$$

* en ce qui concerne le premier point :

$$\frac{||P_\mu x_1||}{||P x_1||} \geq \frac{||P_\mu C x_0|| - ||P_\mu F(x_0)||}{||P C x_0|| + ||P F(x_0)||}$$

en utilisant la symétrie de C :

$$||P_\mu C x_0|| = |\mu| ||P_\mu x_0||$$

$$||P C x_0|| = ||P x_0||$$

$$\frac{||P_\mu(x_1)||}{||P(x_1)||} \geq \frac{|\mu| \frac{||P_\mu(x_0)||}{||P(x_0)||} - K \varepsilon_{(x_0)}^2 \frac{||x_0||}{||P x_0||}}{1 + K \varepsilon_{(x_0)}^2 \frac{||x_0||}{||P x_0||}}$$

$$||P_{(x_0)}|| = ||x_0|| - ||(1-P)x_0|| \quad \text{donc} \frac{||x_0||}{||Px_0||} = \frac{1}{1-\varepsilon(x_0)}$$

en utilisant l'hypothèse (b)

$$\frac{||P_\mu (x_1)||}{||P (x_1)||} \geq \frac{|\mu| [1 - \varepsilon(x_0)] - \frac{K \varepsilon(x_0)}{Q[1 - \alpha\varepsilon(x_0)]}}{1 - \varepsilon(x_0) + K \varepsilon(x_0)^2} \frac{||P_\mu x_0||}{||P(x_0)||}$$

en posant $$f(\varepsilon) = \frac{|\mu|(1-\varepsilon)(1-\alpha\varepsilon) - \frac{K}{Q}\varepsilon}{(1-\varepsilon + K\varepsilon^2)(1-\alpha\varepsilon)}$$

nous avons : $\lim_{\varepsilon \to 0} f(\varepsilon) = |\mu| > 1$

Il est donc facile de trouver η_1 tel que si

$$\varepsilon < \eta_1 \qquad f(\varepsilon) > \frac{1}{2}(|\mu| + 1) > 1$$

alors $$\frac{||P_\mu(x_1)||}{||P (x_1)||} > \frac{1}{2}(|\mu|+1) \frac{||P_\mu x_0||}{||P (x_0)||}$$

il reste à vérifier le second point :

d'abord $$\varepsilon(x_1) \leq \frac{||(1-P)(x_o)|| + ||(I-P)F(x_0)||}{||Cx_0|| - ||F(x_0)||}$$

$$||C(x_0)|| = ||P C x_0|| + ||(I-P)C x_0|| \geq ||Px_0||$$

$$||(I-P)C x_0|| \leq |\mu| \; ||(I-P)x_0||$$

d'où $$\varepsilon(x_1) \leq \frac{|\mu|\varepsilon(x_0) + K \varepsilon(x_0)^2}{1 - \varepsilon(x_0) - K \varepsilon(x_0)^2}$$

On posera $$g(\varepsilon) = \varepsilon \frac{\mu + K\varepsilon}{1 - \varepsilon - K\varepsilon^2}$$

La fonction $\phi \to Q(\phi - \alpha\phi^2)$ étant croissante dans un voisinage de 0 ($Q > 0$), on a :

si $\varepsilon(x_1)$ et $\varepsilon(x_0) < \eta_2$ $\quad Q\,[\varepsilon(x_1) - \alpha\varepsilon(x_1)^2] \leq Q\,[g(\varepsilon(x_0)) - \alpha\, g(\varepsilon(x_0))^2\,]$

comme $\dfrac{||P_\mu(x_1)||}{||P\,x_1||} \geq f(\varepsilon_{(x_0)}) \cdot Q(\varepsilon_{(x_0)} - \alpha\varepsilon_{(x_0)}^2)$

Il suffit de vérifier la condition : (on note $\varepsilon_0 = \varepsilon_{(x_0)}$)

$$\phi_{(\varepsilon_0)} = \frac{f(\varepsilon_0) \cdot Q \cdot (\varepsilon_0 - \alpha\varepsilon_0^2)}{Q\,[g_{(\varepsilon_0)} - \alpha g_{(\varepsilon_0)}]^2} \geq 1$$

$$\phi(\varepsilon) = \frac{[|\mu|(1-\varepsilon)(1-\alpha\varepsilon) - \frac{K}{Q}\varepsilon]\,[1-\varepsilon-K\varepsilon^2]^2}{(1-\varepsilon+K\varepsilon^2)(|\mu|+K\varepsilon)(1-\varepsilon(1+\alpha|\mu|)-\varepsilon^2(K+\alpha K)}$$

On fera un développement limité à l'ordre 1 autours de $\varepsilon = 0$

$$\phi(\varepsilon) = \frac{1 + \varepsilon[-(\alpha+2) - \frac{K}{Q|\mu|}]}{1 + \varepsilon[\frac{K}{|\mu|} - (1+\alpha|\mu|)]} + 0(\varepsilon)$$

Ce développement montre que :

1) $\lim_{\varepsilon \to 0} \phi_{(\varepsilon)} = 1$

2) si $\quad -(\alpha+2) - \dfrac{K}{Q|\mu|} > \dfrac{K}{|\mu|} - (1 + \alpha|\mu|)$

soit, si $\quad \alpha > \dfrac{\frac{K}{|\mu|}(1+\frac{1}{Q}) + 1}{|\mu| - 1}$

la fonction $\phi(\varepsilon)$ est supérieure à 1 sur un intervalle

$$\varepsilon \in\,]0,\ \eta_3[$$

donc, tant que $\varepsilon(x_k) < \eta = \text{Min}\,(\eta_1, \eta_2, \eta_3)$

Nous pouvons appliquer récursivement le résultat :

$$\frac{||P_\mu(x_k)||}{||P\ (x_k)||} \geq A^k \frac{||P_\mu(x_0)||}{||P\ (x_0)||}$$

et :

$$\epsilon(x_k) = \frac{||(1-P)x_k||}{||P\ (x_k)||} \cdot \frac{||P(x_k)||}{||x_k||} \geq \frac{||P_\mu(x_k)||}{||P(x_k)||} \cdot (1-\eta) \geq (1-\eta) \frac{||P_\mu x_0||}{||P\ x_0||} \quad A^k$$

la suite x_N ne peut vérifier la condition :

$$\epsilon(x_N) < \eta$$

car alors $\lim_{N\to\infty} \epsilon(x_N) = +\infty \qquad (A > 1)$

Remarques :

1) la démonstration précédente reste valable dans le cas non symétrique, à condition que la valeur propre μ soit semi-simple. On aura encore :

$$||P_\mu\ C\ x\ || = |\mu|\ ||P_\mu(x)||$$

2) Dans un cas plus général, notre démonstration cesse d'être valable, et il n'est sans doute pas simple de l'y adapter. Néanmoins, il s'agit là d'un problème technique, et la nature du problème ne change probablement pas dans le cas non symétrique.

3) Nous conclurons donc que des convergences vers des valeurs propres intermédiaires ne sont pas théoriquement exclues, mais qu'elles semblent très improbables. Dans la partie concernant les essais numériques, nous avons essayé d'initialiser l'algorithme avec un vecteur "très voisin" du sous-espace propre associé à λ_2 .

4) Le problème de l'optimisation des paramètres α et β intervenant dans la définition de l'itération reste entier, aussi bien d'un point de vue local, soit : x_n étant donné, calculer α_n et β_n tels que μ_{n+1} soit minimum,
que d'un point de vue asymptotique : minimiser la valeur propre sous dominante de la matrice $C(\lambda_1) = I - (\alpha T_A + \beta\ T_B)^{-1}\ (A - \lambda_1\ B)$

C - EXPERIENCES NUMERIQUES

Etant donné le nombre important de paramètres intervenant dans la définition des méthodes étudiées, il a été nécessaire d'effectuer des essais préliminaires sur des exemples de petite taille.

Il en a résulté un certain nombre de choix, que nous avons fait lors de l'implémentation pratique de ces méthodes dans un programme d'éléments finis : le code DELTA, développé par A. PONCET [ ], au laboratoire IMAG de l'Université de Grenoble.

C1 - Essais préliminaires

Ces essais ont porté sur un exemple de petite taille, dont les éléments propres sont connus explicitement :

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & & & \\ & & \ddots & & \\ & & & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 4 & 1 & & & \\ 4 & 4 & & & \\ & & \ddots & & \\ & & & 4 & 1 \\ & & & 1 & 2 \end{pmatrix}$$

avec n = 13.

Les éléments propres sont :

$$\left.\begin{aligned} \lambda_k &= \left[1-\cos(2k-1)\frac{\pi}{2n}\right] \Big/ \left[2+\cos(2k-1)\frac{\pi}{2n}\right] \\ x_k^i &= \sin\left[(2k-1)\,i\frac{\pi}{2n}\right] \qquad i = 1\ldots n \end{aligned}\right\} \quad k = 1,\ldots,n$$

donc λ_1 est simple et le problème est bien conditionné puisque :

$$\begin{cases} \lambda_2 \neq 9.09\ \lambda_1 \\ K = \dfrac{\lambda_{13} - \lambda_1}{\lambda_2 - \lambda_1} \neq 100 \end{cases}$$

Les calculs ont été effectués sur l'IRIS 80 du C.I.C.G. (Centre Interuniversitaire de Calcul de Grenoble) en double précision.

Nous avons testé les méthodes suivantes :

1) Méthode SOR avec différentes valeurs de ω
2) Méthode SOR par blocs, avec des blocs de taille deux
3) "Méthode 1", soit :

$$r_N = (A - \mu_N B)\ x_N$$

$$p_N = -\left[\alpha_N\ T_A(\omega) + \beta_N\ T_B(\omega)\right]^{-1}\ r_N$$

$$x_{\mu+1} = \left[x_N + p_N\right] / \left|\left|x_N + p_N\right|\right|_B$$

avec $T_A(\omega) = \dfrac{1}{\omega} D_A - E_A$.

Rappelons que la convergence est assurée dès que :

$$\begin{cases} \alpha_N \geq 1 \\ \beta_N \geq 0 \end{cases}$$

D'autre part, un certain nombre de variantes et d'améliorations seront suggérées et testées.

C11 - Une première série de tests a consisté à déterminer pour cette dernière méthode, l'influence du choix des paramètres α et β, choisis au début de l'itération et, à cette étape, maintenus constants, sur la convergence. Les tableaux C1 et C1 bis indiquent, pour un même vecteur initial, la précision obtenue, mesurée par les quantités :

$$||r_N||_2 \quad \text{et} \quad \frac{\mu_N - \lambda_1}{\lambda_1}$$

dans les différents cas, après un nombre fixé d'itérations.

Nous constatons, à la lecture de ces tableaux, que la convergence la plus rapide a lieu pour des valeurs n'appartenant pas au domaine dans lequel elle est établie théoriquement. Ceci semble indiquer que des valeurs optimales compatibles avec la condition théorique de convergence, sont :

$$\alpha = 1, \ \beta = 0$$

Il est certainement difficile de démontrer une telle propriété. Nous pouvons néanmoins formuler la remarque suivante :

Si on considère l'application de la méthode précédente à la recherche de la plus grande valeur propre de B ; nous supposons donc :

$$A = I$$

et que l'on fixe la valeur de β et ω

$$\beta = 0 \ ; \ \omega = 1$$

Alors la valeur optimale de α est certainement inférieure à 1 :

$$\alpha_{opt} < 1$$

β \ α	0.8	0.9	1	1.1
0.1			7.10^{-5} / $1.7\ 10^{-5}$	10^{-4} / $3\ 10^{-5}$
0	25 itérations / $1{,}4\ 10^{-3}$ / 2.10^{-4}	$3.1\ 10^{-9}$ / $< 10^{-8}$	$2.2\ 10^{-8}$ / $\sim 10^{-8}$	10^{-7} / $\sim 10^{-8}$
-0.05			$2.3\ 10^{-10}$ / 10^{-8}	
-0.1			17 itérations / $6.8\ 10^{-4}$ / 5.10^{-4}	

TABLEAU C1

. 50 itérations

. vecteur initial :
$x_o^t = (1, -1, 1, -1, \ldots)$

. $\omega = 1.5$

. un test d'arrêt portant sur la décroissance de r_N a stoppé le programme dans deux cas.

. lecture du tableau :

$||r_{50}||_2$ / $\dfrac{\mu_{50} - \lambda_1}{\lambda_1}$

β \ α	0.9	0.95	1	1.05	1.1	1.5	2
0.05			$1.4\ 10^{-2}$ / 0.72				
0	$6.4\ 10^{-3}$ / 10^{-1}	$7\ 10^{-3}$ / $1.3\ 10^{-1}$	$7.7\ 10^{-3}$ / $1.6\ 10^{-1}$	$8\ 10^{-3}$ / $1.9\ 10^{-1}$	9.10^{-3} / 0.5	$1.3\ 10^{-2}$	$1.6\ 10^{-2}$ / 0.76
-0.05		$2.4\ 10^{-3}$ / $4\ 10^{-3}$	$2.5\ 10^{-3}$ / $5\ 10^{-3}$				
-0.1			$6.8\ 10^{-3}$ / $4.5\ 10^{-2}$				

TABLEAU C1 bis : sur 10 itérations avec $x_o^t = (1, 1 \ldots 1)$

En effet, pour $\omega = 1$, le taux de convergence assymptotique est gouverné par la valeur propre sous dominante de la matrice :

$$C(\alpha, \beta, \lambda_1) = I - M^{-1}(A - \lambda_1 B)$$
$$\text{où } M = \alpha\, T_A + \beta\, T_B$$

désignons par $q^i_{\alpha\beta}$ $\quad i = (p+1),\ldots,n$ les valeurs propres de cette matrice différentes de 1 (p est la multiplicité de λ_1 dans le problème $Ax = \lambda Bx$).

Il est facile de voir que, si $\beta = 0$, les vecteurs propres de $C(\alpha, 0, \lambda_1)$ sont indépendants de α :

$$C(\alpha, 0, \lambda_1)x^i = q^i_{\alpha,0}\, x^i$$

$$\alpha(1-q^i_{\alpha,0})T_A\, x^i = (A - \lambda_1 B)x^i$$

donc : l'optimum (dans le cas $\beta = 0$; $\omega = 1$) est solution de :

$$\inf_{\alpha \geq 1} \max_{i=(p+1)\ldots n} \left| \frac{q^i_{10} - 1}{\alpha} - 1 \right|$$

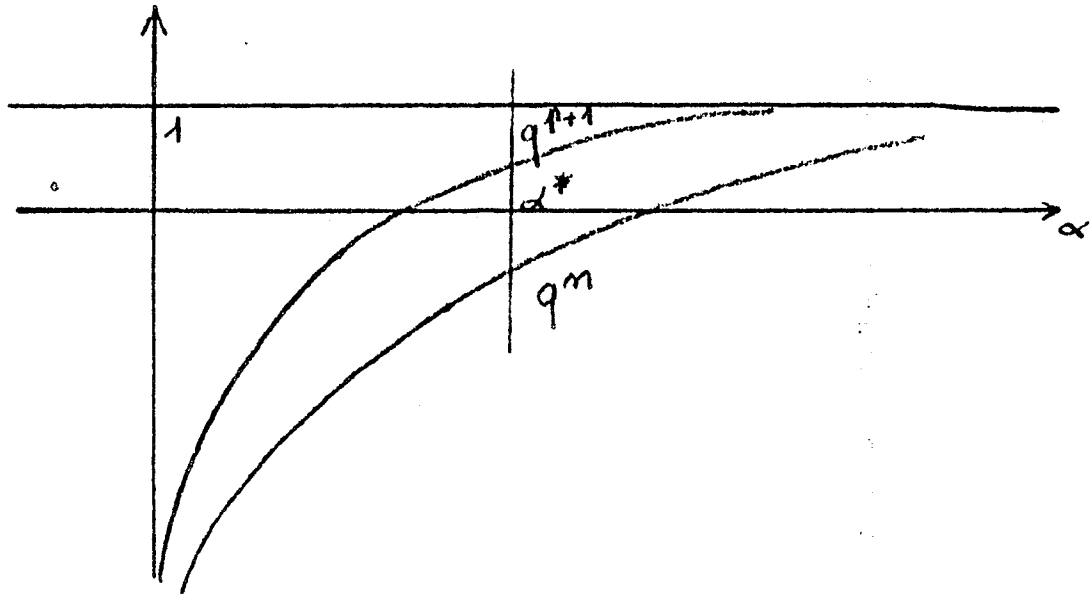
la simplification supplémentaire $A = I$ permet d'aller plus loin (dans ce cas, la méthode n'est rien d'autre qu'une itération polynomiale, puisque :

$$x_{N+1} = \frac{\alpha_N}{\alpha} B\, x_N + (1 - \frac{1}{\alpha})\, x_N)$$

B et C admettent les mêmes vecteurs propres :

$$(I - \frac{1}{\lambda_i} B)\, x_i = 0 \qquad C\, x_i = [1 + \frac{1}{\alpha}(\frac{\lambda_1}{\lambda_i} - 1)]\, x_i$$

les variations des valeurs propres de C en fonction de α dans l'intervalle $]0, +\infty[$ sont représentées par les courbes :



l'optimum sur cet intervalle est donc :

$$\alpha^* = 1 - \frac{\lambda_1}{2}\left(\frac{1}{\lambda_2} + \frac{1}{\lambda_n}\right)$$

il en résulte que $0 < \alpha^* < 1$. Sur l'intervalle $[1, +\infty[$, la valeur optimale est :

$$\alpha_{opt} = 1$$

Remarque :

Le problème de l'optimisation "locale" des paramètres α_N et β_N est compliqué :
soit à trouver α_N et β_N tels que : μ_{N+1} soit minimum, sachant que $x_{N+1} = x_N - (\alpha_N T_A(\omega) + \beta_N T_B(\omega))^{-1} r_N$.

Par contre, l'optimisation de α_N, dans le cas où β_N est fixé égal à zéro, est simple, puisque la dépendance devient linéaire :

$$x_{N+1} = x_N + \frac{1}{\alpha_N} T_A(\omega)^{-1} r_N$$

Les tests de cette variante sont reportés sous le nom de "méthode 1 avec α optimisé".

C12 - Le tableau C12 donne une comparaison des taux de convergence mesurés par la quantité :

$$||r_N||^{1/N}$$

pour différentes valeurs admissibles de ω. Les courbes C12 bis illustrent cette dépendance.

Nous n'avons pas reporté les temps de calcul dans ces tableaux. On trouvera en annexe une comparaison des coûts en nombre de multiplications de chaque méthode.

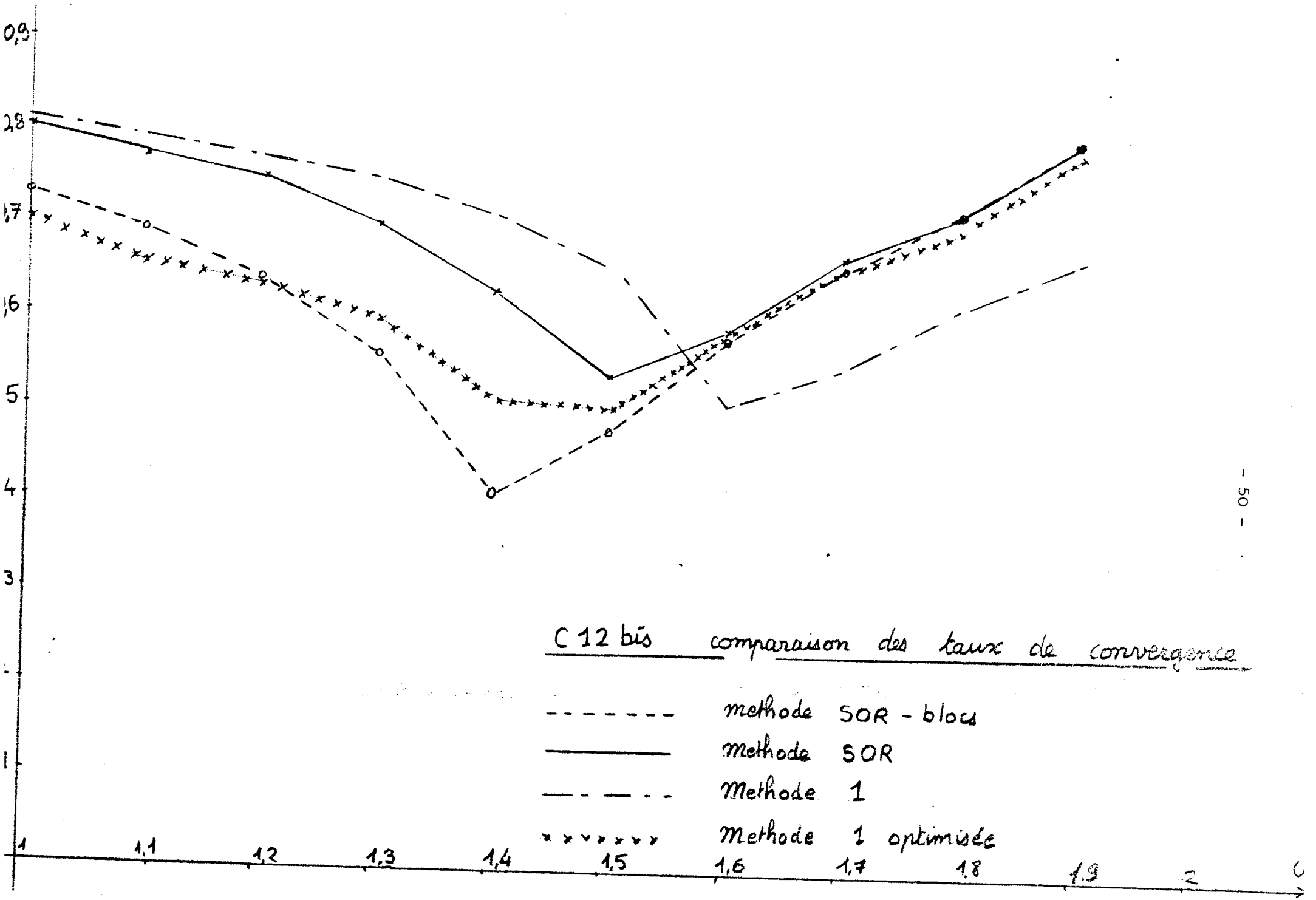
C13 - Le troisième test a consisté à initialiser les différentes méthodes avec un vecteur "proche" du sous-espace associé à λ_2 (pratiquement, nous avons gardé trois chiffres significatifs à chacune des composantes).

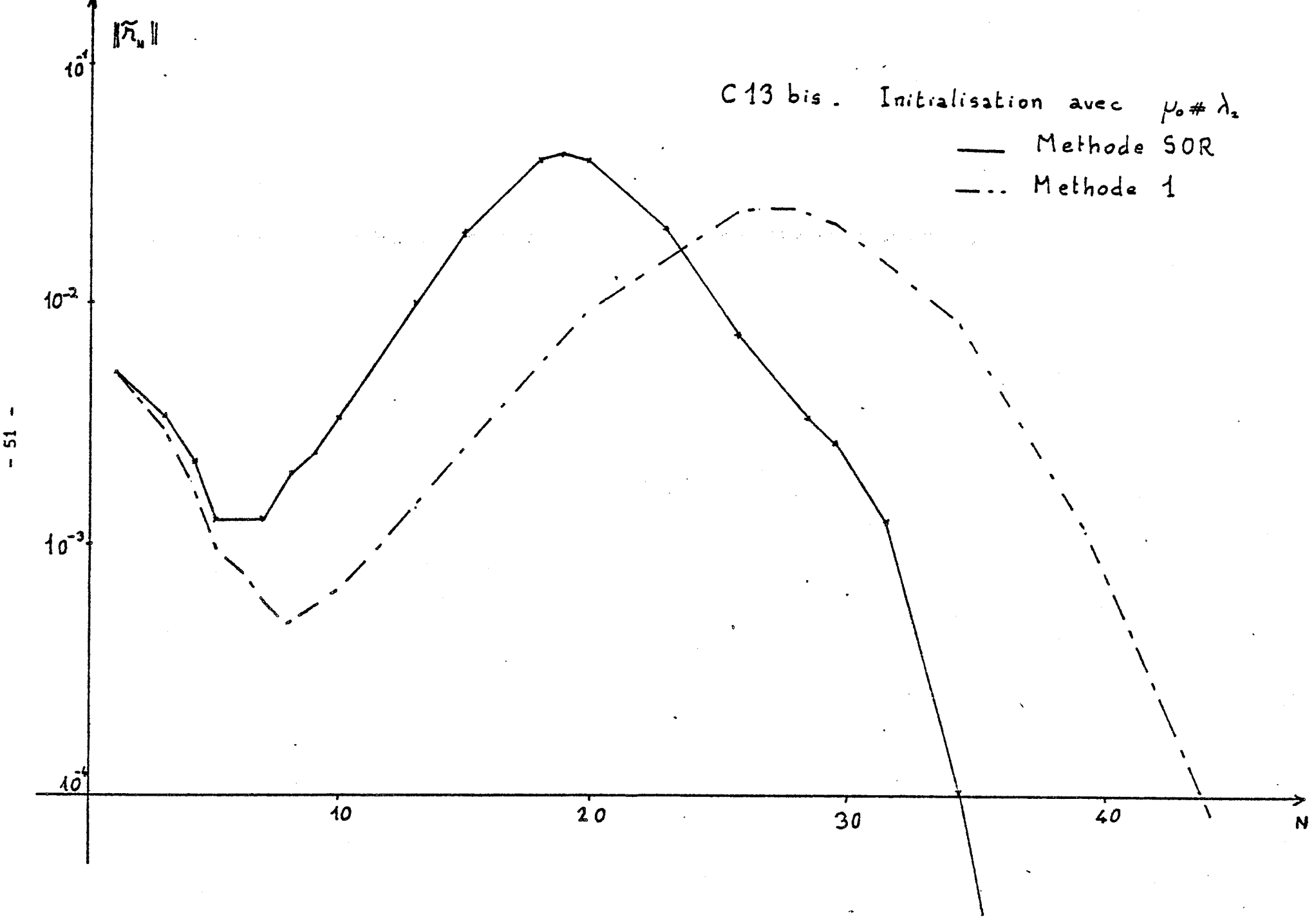
Nous reportons, dans le graphe C13 l'évolution du quotient de Rayleigh au cours de l'itération pour la méthode de SOR et la méthode 1. Le graphe C13 bis montre l'évolution de $||r_N|| = ||(A - \mu_N B)\hat{x}_N||$.

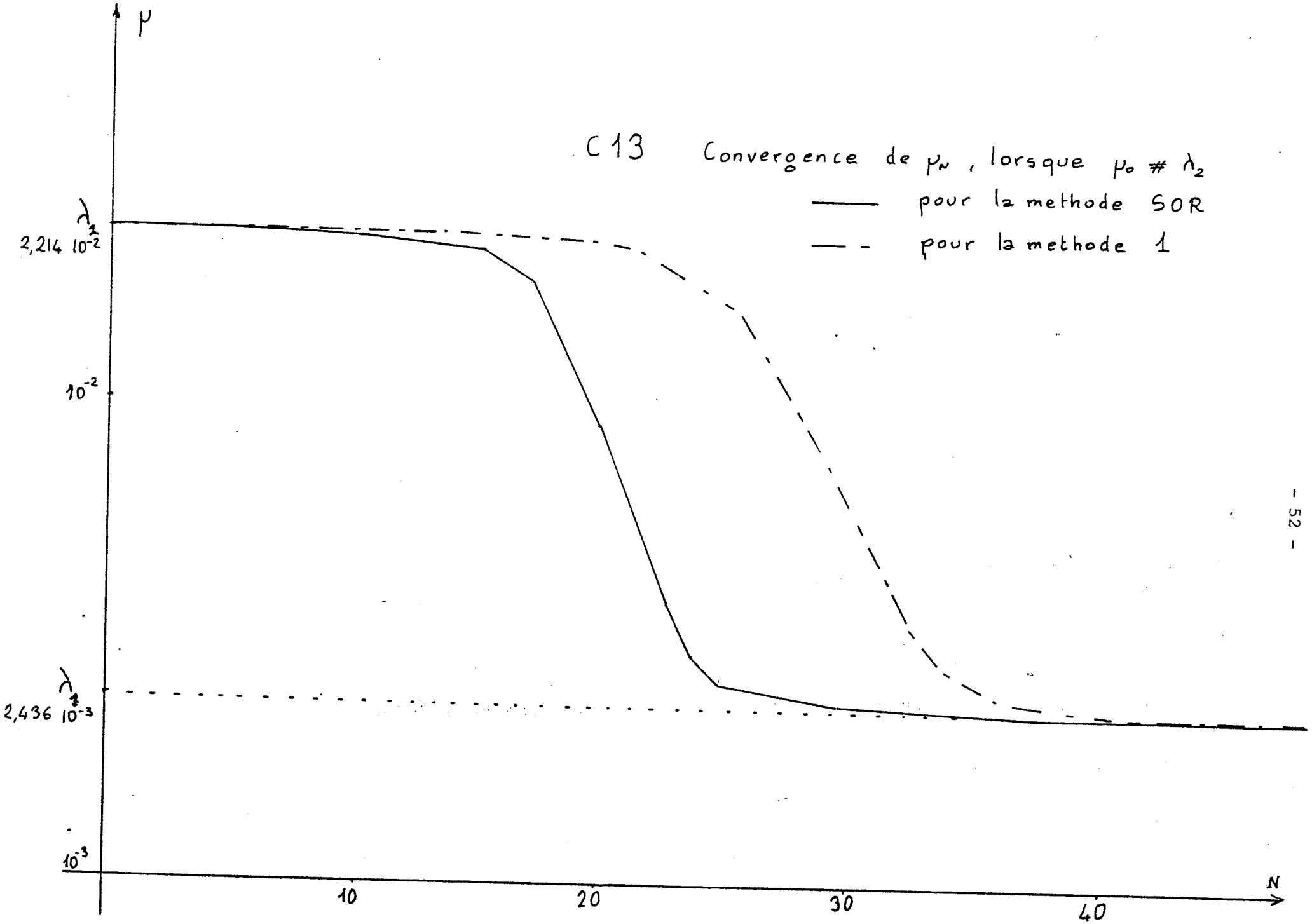
La méthode SOR a donc ici un comportement meilleur, puisque 15 itérations sont suffisantes, contre 21 dans l'autre cas, pour observer une décroissance significative de μ_N.

ω	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
SOR	0.8 n = 40 0.81 n = 50	0.77 n = 40 0.79 n = 50	0.75 n = 40 0.76 n = 50	0.70 n = 40 0.72 n = 50	0.63 n = 40 0.64 n = 50	0.54 n = 36 —	0.59 n = 40 —	0.67 n = 40 0.69 n = 50	0.74 n = 40 0.77	0.8 0.83
SOR blocs	0.73 n = 40 0.75 n = 50	0.69 n = 40 0.71 n = 50	0.64 n = 40 0.65 n = 50	0.56 n = 40 —	0.41 n = 25 —	0.48 n = 32 —	0.58 n = 39 —	0.66 0.67	0.74 0.75	0.8 0.8
Méthode 1	0.81 n = 40 0.83 n = 50	0.79 0.81	0.77 n = 40 0.79 n = 50	0.75 0.76	0.71 0.73	0.66 0.67	0.51 n = 31 —	0.55 n = 36 —	0.62 n = 40 0.64 n = 47	0.67 0.69
Méthode 1 optimisée	0.7 n = 40 0.71 n = 50	0.65 n = 40 0.66 n = 48	0.64 n = 40 0.63 n = 44	0.6 n = 40 —	0.52 n = 33 —	0.51 n = 32 —	0.6 n = 40 —	0.66 0.67	0.73 0.74	0.79 0.80

C 12 : taux de convergence, mesuré par $||r_n||_2^{1/n}$, pour un vecteur initial $x_o^t = (1, 1, \ldots, 1)$


0,9
2,8
1,7
1,6
5
4
3
1
1
1,1
1,2
1,3
1,4
1,5
1,6
1,7
1,8
1,9
2
C 12 bis comparaison des taux de convergence
methode SOR - blocs
methode SOR
Methode 1
Methode 1 optimisée


$\|\tilde{r}_k\|$
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10
20
30
40
N
C 13 bis. Initialisation avec $\mu_0 \# \lambda_2$
Methode SOR
Methode 1


μ
C 13 Convergence de μ_N , lorsque $\mu_0 \neq \lambda_2$
pour la methode SOR
pour la methode 1
λ_2
2,214 10^{-2}
10^{-2}
λ_3
2,436 10^{-3}
10^{-3}
10
20
30
40
N

C2 - Calcul des valeurs propres suivantes :

Nous ne faisons ici qu'ébaucher ce problème. Il y a, à priori, deux directions possibles :

a) soit l'utilisation de techniques de déflation :
si x_1 a été déterminé tel que :

$$Ax_1 = \lambda_1 Bx_1 \quad \text{et} \quad x_1^t B x_1 = 1$$

alors, l'application de l'une quelconque des méthodes décrites ici avec :

$$A' = A + p(Bx_1)\ (Bx_1)^t \quad ; \quad B' = B$$

permet de calculer λ_2 et x_2, si $p > \lambda_2 - \lambda_1$.

Il est, en principe, possible de calculer toutes les valeurs propres que l'on désire par un tel procédé. Cependant :

- le choix du paramètre p est toujours délicat en partique, car, d'une part, il doit vérifier une condition qui dépend de λ_2, inconnue, et, d'autre part, il ne doit pas être choisi trop grand, sous peine de ralentir la convergence [11] ;

- ensuite, l'accumulation des erreurs ne permet pas en pratique d'aller très loin.

b) Un deuxième choix possible consiste à travailler avec plusieurs vecteurs simultanément, et à utiliser une méthode de projection sur le sous-espace engendré par ces vecteurs. Le principe est indiqué ici sous aucune justification théorique de convergence.

soient x_N^i (i = 1,..., p) p vecteurs indépendants. On utilisera (par exemple) la méthode 1 sur chacun de ces vecteurs :

$$\forall i = 1,\ldots, p \begin{cases} \mu_N^i = [(x_N^i)^t A x_N^i] / (x_N^i)^t B x_N^i \\ y_N^i = [I - T_A(\omega)^{-1} (A - \mu_N^i B)] x_N^i \end{cases}$$

soit Y_N la matrice n x p engendrée par les vecteurs y_N^i :

$$Y_N = [y_N^1, y_N^2, \ldots, y_N^p]$$

Le problème projeté est alors défini par :

$$[Y_N^t A Y_N]\eta = \lambda[Y_N^t B Y_N]\eta$$

si on désigne par (η_N^i, λ_N^i) les p solutions de ce problème, telles que $(\eta_N^i)^t \cdot \eta_N^j = \overset{N}{\delta}_{ij}$,

alors $x_{N+1}^i = Y_N \ \eta_N^i \quad \forall i = 1,\ldots,p$

C'est donc une méthode d'itérations simultanées, où l'espace d'approximation est généré par l'une des techniques développées dans cette thèse (et qui ont l'avantage d'éviter une factorisation de B ou A sous la forme LL^t).

Nous avons testé cette méthode avec seulement deux vecteurs. (dans un cas plus général, il faut utiliser une des méthodes pour la résolution d'un problème Ax = λBx, avec des matrices pleines et petites (QZ par exemple)) en l'utilisant conjointement avec la méthode 1.

Les tableaux et graphes (C41 et C41 bis montrent la dépendance en ω du taux de convergence, mesuré par $||\hat{r}_N||_2^{1/N}$, pour les deux vecteurs, avec l'initialisation suivante :

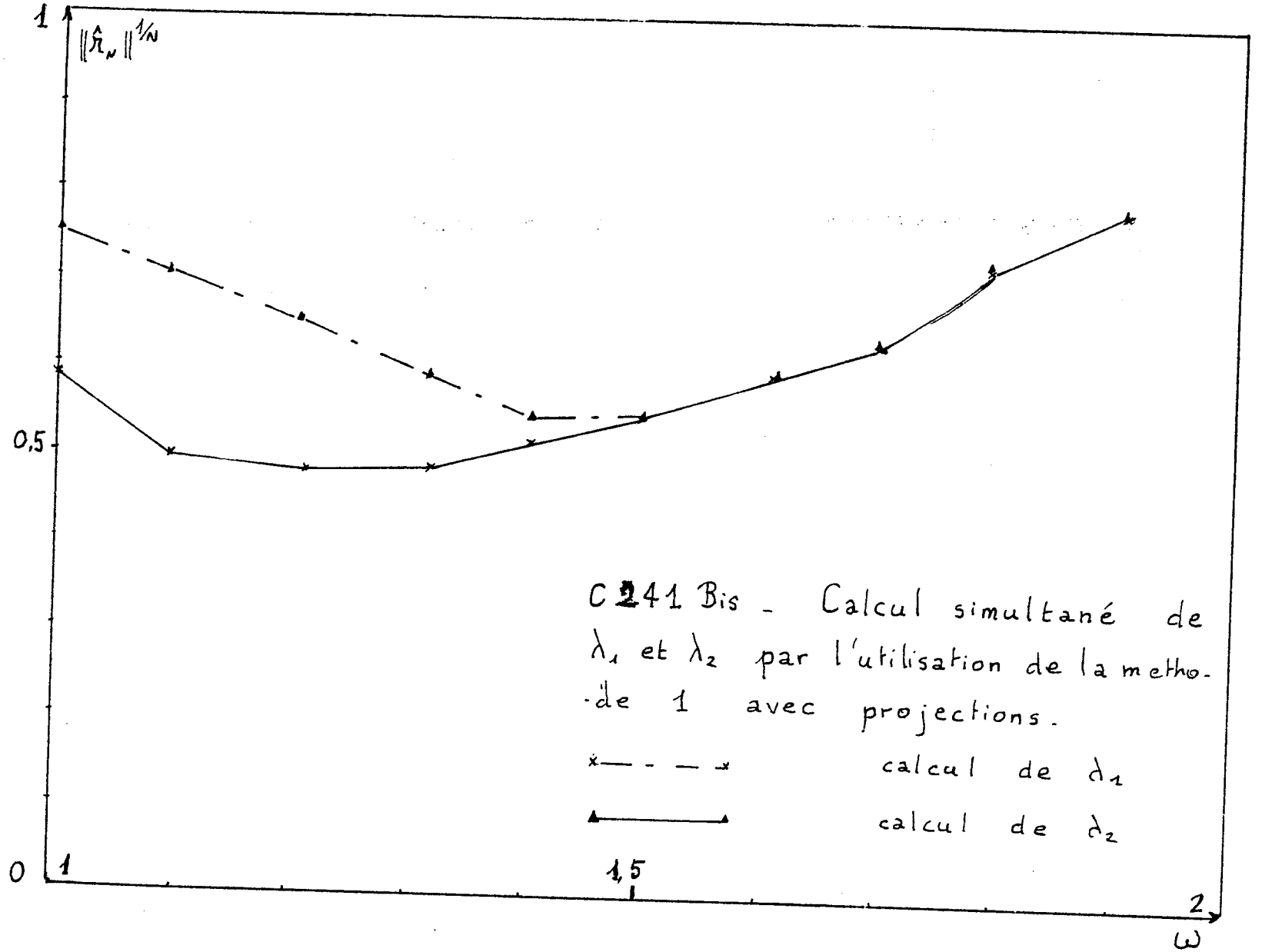
$$x_o^i = 1$$
$$y_o^i = (-1)^i \qquad i = 1 \ldots 13$$

ω	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
calcul λ_1	0.59 / $0.57.10^{-9}$	0.51 / $.17\ 10^{-11}$	0.5 / $0.63\ 10^{-12}$	0.5 / $.83\ 10^{-12}$	0.53 / $.73\ 10^{-11}$	0.55 / $.39\ 10^{-10}$	0.59 / $.92\ 10^{-9}$	0.65 / $.38\ 10^{-7}$	0.73 / $.34\ 10^{-5}$	0.82 / $.28\ 10^{-3}$
calcul λ_2	0.75 / $0.13\ 10^{-4}$	0.71 / $.13\ 10^{-5}$	0.65 / $0.44\ 10^{-7}$	0.6 / $.97\ 10^{-9}$	0.54 / $.18\ 10^{-10}$	0.55 / $.39\ 10^{-10}$	0.59 / $.75\ 10^{-9}$	0.65 / $.39\ 10^{-7}$	0.73 / $.32\ 10^{-5}$	0.82 / $.29\ 10^{-3}$

C 241 : Comparaison des taux de convergence pour une méthode de type itérations simultanées basée sur la méthode 1.

$x_o^i = 1$; $y_o^i = (-1)^i$ mesure par $||\hat{r}_N||^{1/N}$ et par $||\hat{r}_N||_2$

(N = 40 dans tous les cas).



C 241 Bis - Calcul simultané de λ_1 et λ_2 par l'utilisation de la méthode 1 avec projections.

x— - — x calcul de λ_1

▲——— calcul de λ_2

nous constatons que :

- la convergence vers λ_1 est grandement accélérée ;
- le deuxième vecteur produit bien une approximation du sous-espace correspondant à λ_2, comme on pouvait l'espérer.

Bien que des tests plus approfondis soient nécessaires pour évaluer véritablement les possibilités de cette variante, les résultats obtenus sont très prometteurs.

C3 - Comparaison des coûts d'une itération pour les différentes méthodes

Les comparaisons que nous allons faire se limitent à dénombrer le nombre de produits nécessaires dans une itération pour chaque méthode. Le calcul de la norme du résiduel $||(A - \mu B)\ \hat{x}||_2$ sera inclus dans cette évaluation, étant donné qu'il intervient dans le test d'arrêt.

On notera n la dimension des matrices, et M le nombre d'éléments non nuls contenus dans chacune des demi-matrices A et B.

C31 - Calcul d'un produit r = Ax

Il s'effectue de la manière suivante :

pour i = 1,..., n

$$r_i = \sum_{j \geq i} a_{ij}\ x_j + tab(i)$$

$$\text{et} \quad tab(j) = tab(j) + a_{ij}\ x_i \qquad \forall j = i+1, \ldots, n$$

En tout, 2M-n produits, et nécessité d'utiliser un vecteur auxiliaire dont la dimension peut être réduite à la demi-longueur de la bande de A.

C32 - Calcul de $x^t Ax$

Sous la forme : $x^t Ax = \sum_{i=1}^{n} x_i \left[a_{ii} x_i + 2 \left(\sum_{j>i} a_{ij} x_j \right)\right]$

il nécessite M+n produits.

C33 - Méthode 1

Le calcul peut être effectué de la manière suivante :

- calcul de $[Ax_N]$ et $[Bx_N]$
- calcul de $P = x_N^t (A\, x_N)$ et $\eta = x_N^t (B\, x_N)$
- résiduel $r_N = (A\, x_N) - (P/\eta)\ (B\, x_N)$
- erreur $\varepsilon = [\Sigma (r_N^i)^2]^{1/2}/\sqrt{\eta}$
- résolution de $T_A\, p_N = - r_N$
- $x_{N+1} = (x_N + \omega . p_N)/\sqrt{\eta}$

sous cette forme, il requiert :

5M + 3n produits

avec utilisation de 3 vecteurs auxiliaires.

Il est possible de réduire à 2 le nombre de vecteurs auxiliaires par une programmation différente, mais dans ce cas, il faut 6M + 2n produits par itération.

C34 - Méthode SOR

Programmée sous la forme suivante :

$$-\mu_N = (x_N^t Ax_N) \;/\; (x_N^t B\, x_N)$$

pour $i = 1, \ldots, n$

$$x_i^{N+1} = \left[\left(\frac{\omega}{a_{ii} - \mu_N b_{ii}}\right) \left(\sum_{j>i} (\mu_N b_{ij} - a_{ij})\, x_j + tab_1(i)\right)\right.$$

$$\left. + (1-\omega) x_i^N\right] \;/\; [x_N^t Bx_N]^{1/2}$$

$$\varepsilon^2 = \varepsilon^2 + \sum_{j>i} [(a_{ij} - \mu_N\, b_{ij}) x_j + tab_2(i)]^2$$

$$tab_1(j) = tab_1(j) + x_i^{N+1} (\mu_N\, b_{ij} - a_{ij}) \quad \forall j = i+1, \ldots, n$$

$$tab_2(j) = tab_2(j) + x_i^N (a_{ij} - \mu_N\, b_{ij}) \quad \forall j = i+1, \ldots, n$$

cette itération requiert $6M + 5n$ produits et elle utilise 2 vecteurs auxiliaires.

C35 - Méthodes par blocs

On suppose que le découpage régulier en blocs de taille p est possible. C'est à dire :

$$\exists k \text{ entier avec } k.p = n$$

Le coût supplémentaire est celui de la résolution des k sous-systèmes (pleins) de table p.

soit : $\frac{n}{3}(p^2 + 3p)$ produits supplémentaires.

C36 - Méthode 1 optimisée

Les calculs supplémentaires occasionnés par la détermination de sont :

- calcul de Ap_N et Bp_N
- de $x_N^t(Ap_N)$, $x_N^t(B\ p_n)$, $p_N^t(Ap_N)$ et $p_N^t(Bp_N)$
- de $x_{N+1} = x_N +$ opt p_N

soit 4M + n produits supplémentaires.

Comme les suites constituées par les valeurs optimales de ne sont pas régulières, il est nécessaire de refaire ce calcul à chaque itération.

TABLEAU RECAPITULATIF

	Expression	Exemple n = 500 M = 5000
Méthode 1	5M + 3n	26.500
SOR	6M + 5n	32.500
SOR par blocs	$6M+5n+\frac{n}{3}(p^2+3p)$	p = 5 40.000 p = 10 54.000 p = 20 109.000
Méthode 1 optimisée	9M + 4n	47.000

C 4 - Matrices provenant de l'utilisation de méthodes d'éléments finis

Les trois méthodes suivantes ont été implémentées dans le programme d'éléments finis DELTA :

- SOR
- SOR par blocs, en se limitant à des blocs de taille 2
- Méthode 1 ($\alpha = 1$, $\beta = 0$).

Il en a résulté un certain nombre de programmes spécifiques, chargés de :

* calculer l'encombrement des matrices stockées sous la forme compacte suivante :

- un tableau de nombres réels, pour stocker les coefficients non nuls de chaque demi-matrice ;
- un tableau d'entiers, pour les indices de colonnes correspondants
- un tableau d'entiers, pour le nombre de coefficients non nuls dans chaque ligne.

Pour illustrer ceci, la matrice :

$$A = \begin{pmatrix} 21 & 0 & 2 & 0 & -1 \\ 0 & 7.5 & 0 & 1.1 & 0 \\ 2 & 0 & 50,1 & 0 & 0 \\ 0 & 1.1 & 0 & 1.2 & 1 \\ -1 & 0 & 0 & 1 & 75 \end{pmatrix}$$

serait stockée sous la forme :

tableau des coefficients

21 2 -1 7,5 1,1 50,1 12 1 75

tableau des indices de colonnes

1 3 5 2 4 3 4 5 5

tableau des encombrements par ligne

3 2 1 2 1

Ce travail est effectué au moment de la constitution du maillage (commande "Domaine")

* d'assembler et de stocker les matrices A et B sous la forme décrite précédemment.

Ce travail est effectué dans la boucle sur les triangles du programme principal de la commande "résultat".

* de gérer, en mode intéractif, le calcul de valeur propre proprement dit ; une possibilité est laissée à l'utilisateur de changer de méthodes, ou de modifier la valeur du paramètre de relaxation, en cours d'itération.

Le problème test que nous avons utilisé est le calcul de vibration de membranes : trouver les fonctions u, définies sur un ouvert régulier donné Ω (de frontière Γ), et les scalaires λ tels que :

$$\begin{cases} \Delta u = -\lambda \ u & \text{dans } \Omega \\ u = 0 & \text{sur } \Gamma \end{cases}$$

on a utilisé la formulation faible suivante :

$$\left.\begin{array}{l} \text{trouver } u \in H_o^1(\Omega) \\ \text{et } \lambda \in \mathbb{R} \text{ tels que} \end{array}\right\} \forall v \in H_o^1(\Omega) \qquad \int_\Omega \Delta u \Delta v dx = \lambda \int_\Omega u \, v dx$$

Ω étant le carré unité : $]0,1[\times]0,1[$ sur lequel les valeurs propres exactes sont connues :

$$u = \sin(\pi k_1 x).\sin(\pi k_2 y) \qquad \lambda = \pi^2(k_1^2 + k_2^2)$$
$$k_1, k_2 = 1, 2, \ldots$$

Nous avons comparé les trois méthodes avec des maillages réguliers et des interpolations sur des éléments triangulaires à trois et six noeuds, par des polynomes de degré un et deux.

Les courbes C21 à C25 indiquent les variations du taux de convergence, mesuré par $||r_n||^{1/n}$, en fonction de ω. Le vecteur initial a été dans tous les cas, calculé par la procédure décrite au paragraphe A28 .

On remarque, dans tous les cas, un avantage plus ou moins prononcé de la méthode par blocs sur la méthode SOR classique. Le passage à des blocs de taille supérieure augmenterait certainement cet avantage, mais au prix de calculs beaucoup plus lourds (initialisation, inversion des blocs diagonaux).

La méthode 1 est très compétitive par rapport au deux autres, et elle semble moins sensible au choix du paramètre de relaxation.

D'autre part, nous avons essayé de profiter des outils dont nous disposons pour comparer l'efficacité des deux procédés d'interpolation cités. Une telle comparaison est rendue possible par l'utilisation de maillages réguliers, ce qui permet d'avoir le même nombre d'inconnues.

Le graphe C26 reporte la précision, mesurée par $(\mu - \lambda_1)/\lambda_1$ dans les différents cas. Le calcul itératif a été poussé jusqu'à ce que la condition :

$$||r_N|| = ||(A - \mu_N B)\ \hat{x}_N|| < 10^{-5}$$

soit remplie, ce qui fait que l'erreur introduite par ce calcul est négligeable devant l'erreur due à la méthode d'éléments finis.

Le graphe indique un net avantage en faveur de l'utilisation d'une interpolation d'ordre élevé , et ceci même avec le maillage le plus grossier.

Cet avantage est certainement lié à la grande régularité des fonctions propres.

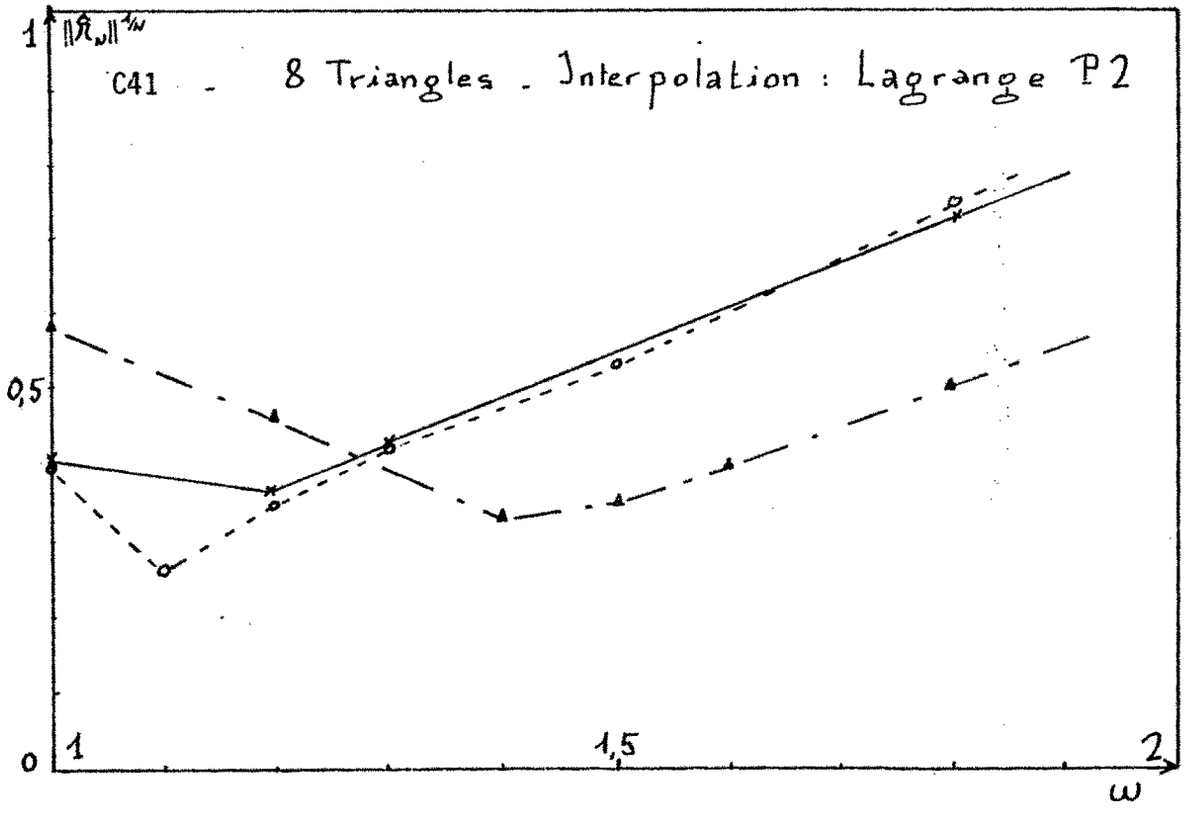
Les courbes C41 à C45 montrent la variation du taux de convergence, $\|\hat{R}_N\|^{1/n}$, pour differentes valeurs de ω, en utilisant une triangulation régulière sur le carré unité.

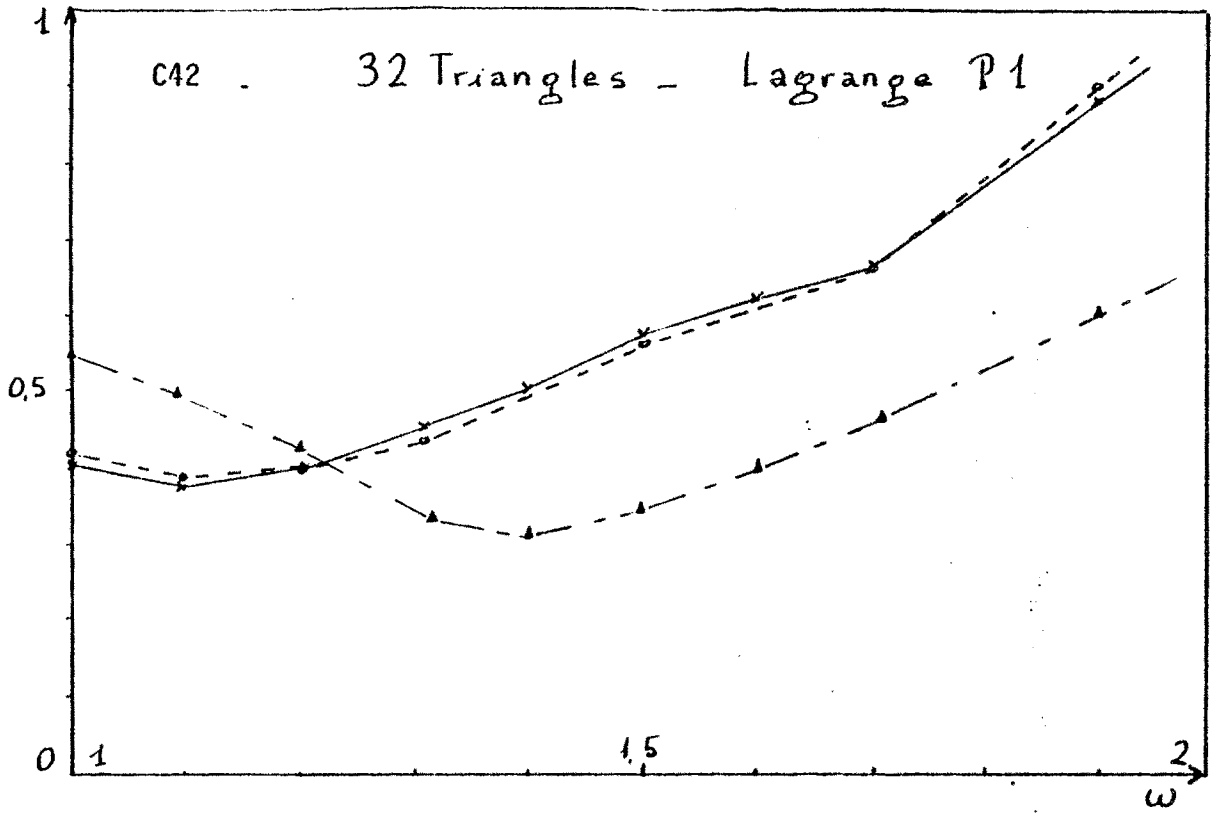
Legende :

× ——— × Methode SOR

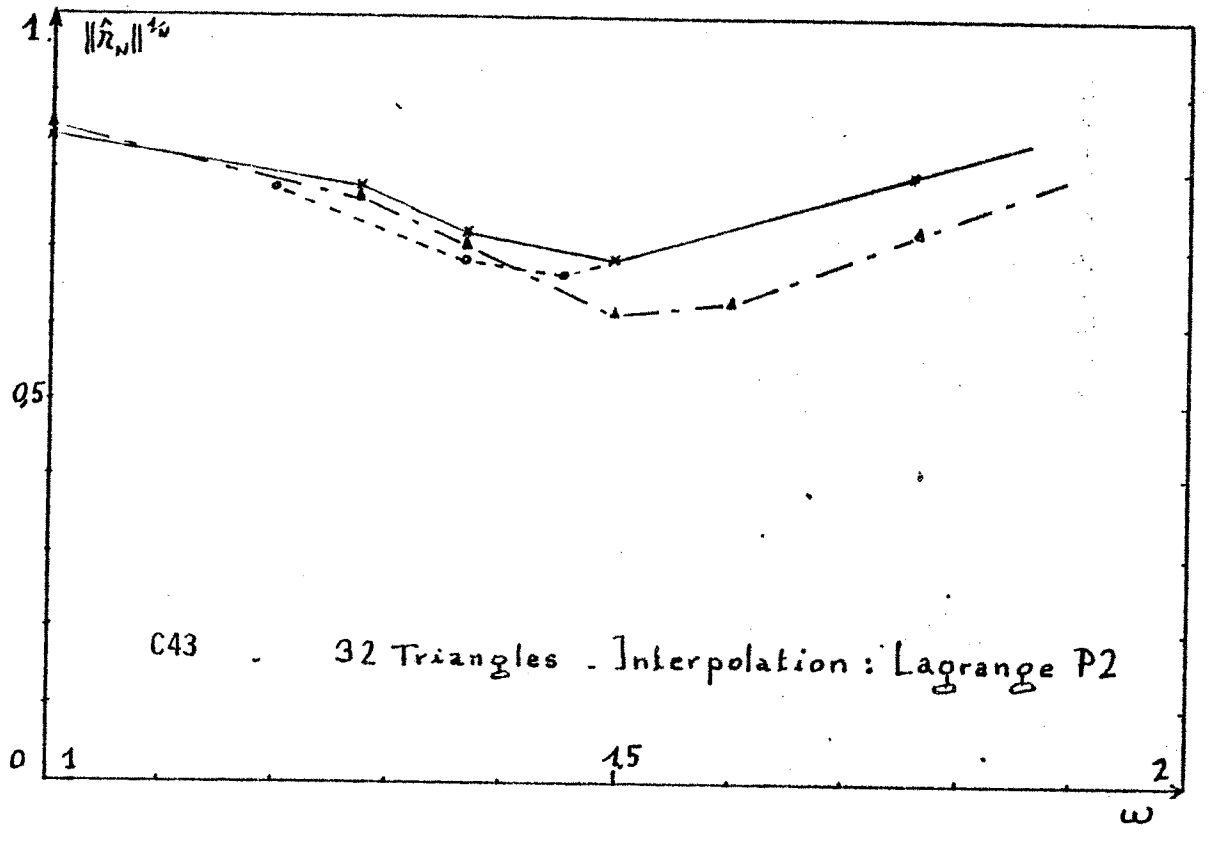
o - - - - - o Methode SOR par blocs

▲ — · — · ▲ Methode 1

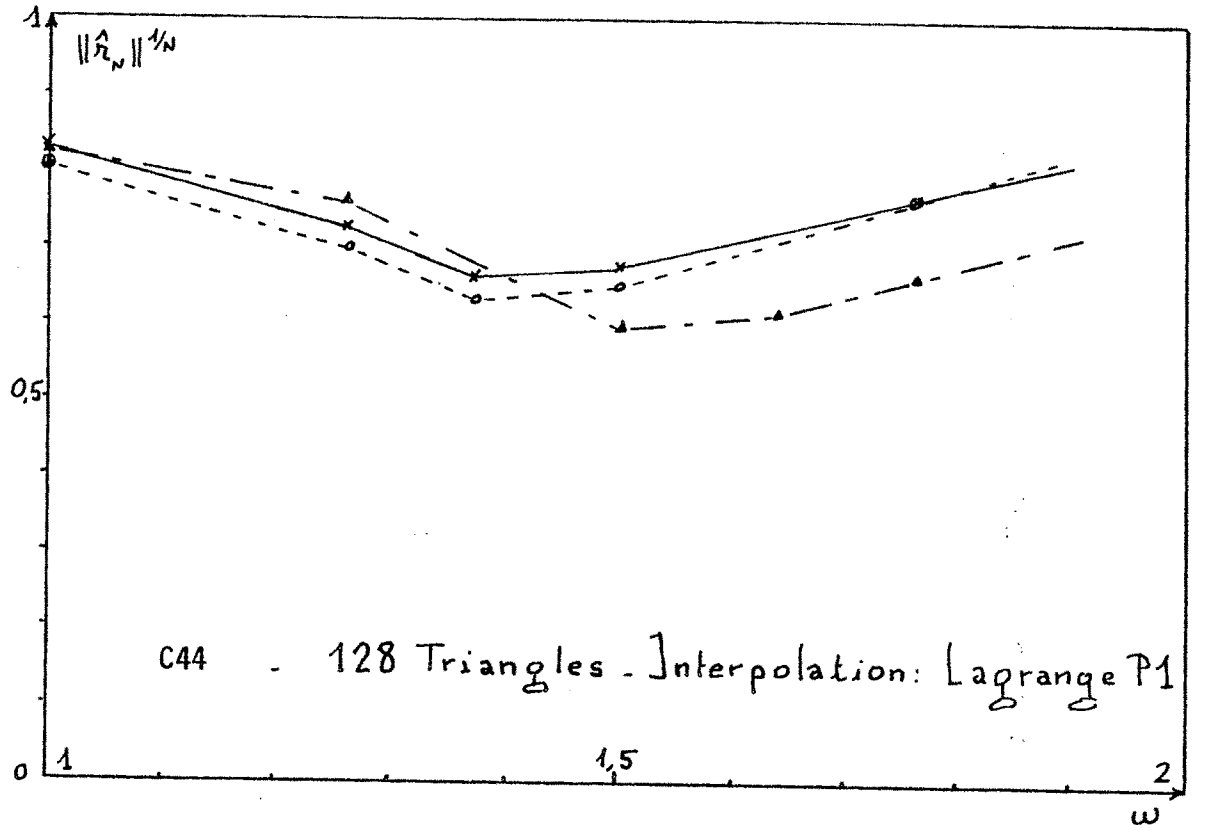




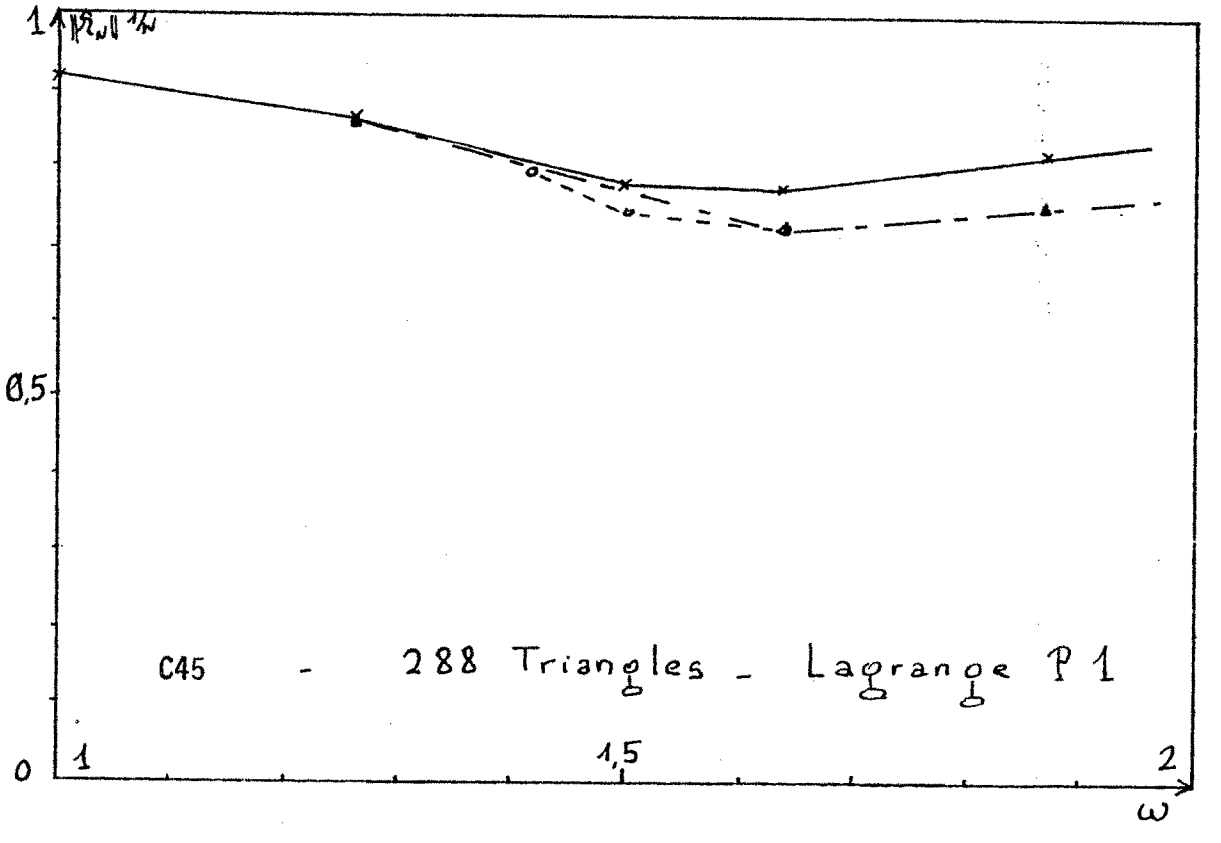
C42 . 32 Triangles - Lagrange P1



C43 . 32 Triangles . Interpolation : Lagrange P2



C44 - 128 Triangles - Interpolation: Lagrange P1

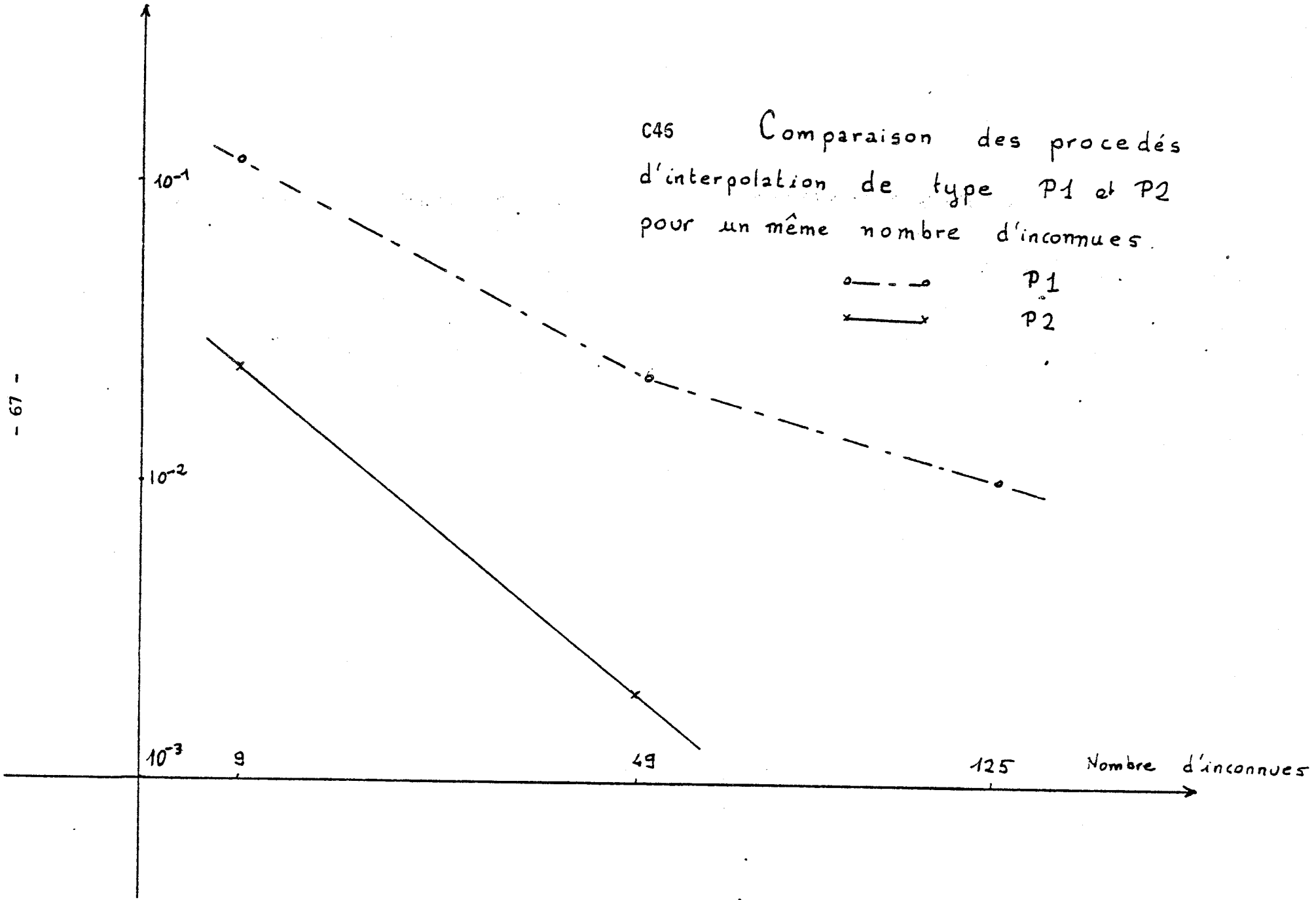


C45 - 288 Triangles - Lagrange P 1

C45 Comparaison des procedés d'interpolation de type P1 et P2 pour un même nombre d'inconnues.

o— - —o P1

×——× P2



REFERENCES

[1] BUFFONI

"Evaluation of eigenvalues of discrete space equation"
Calcolo 4, 169-177 (1967)

[2] T. KATO

"Perturbation theory for linear operators"
Springer Verlag (1976)

[3] T. KATO

"On the upper and lower bounds of eigenvalues"
J. Phys. Soc. Jap. 4, 334-339 (1949)

[4] G. PETERS et J.M. WILKINSON

"Ax = λBx and the generalised eigenvalue problem"
Siam J. num. anal. vol 7 n° 4, 479-492 (1970)

[5] C. REINSCH et J.M. WILKINSON

"Handbook of automatic computation"
lin. alg. (vol 2). Springer Verlag (1972)

[6] G. RODRIGUE

"A gradient method for the eigenvalue problem Ax = λBx"
Num. Maths. 22, 1-16 (1973)

[7] A. RUHE

"SOR methods for the eigenvalue problem with large sparse matrix"
Math. of Comp. vol 28 n° 127 (July 1974)

[8] A. RUHE

"Iterative Eigenvalue Algorithms based on convergent splittings"
Report UMINF 43 (1973)

[9] Y. SAAD

"On the rate of convergence of the block Lanczos methods"
Rap. Rech. de l'Université de Grenoble n° 123 (Janv. 1978)
également à paraitre dans SIAM. J. num. anal.

[10] M.R. SCHWARTZ

"La méthode de surrelaxation en coordonnées pour (A-λB)x"
Sem. Anal. num. Grenoble n° 223 (Mars 1975)

[11] M.R. SCHWARTZ

"Two algorithms for treating Ax = λBx"
Com. maths. in Applied Mech. and Eng. 12, 181-199 (1977)

[12] G.W. STEWART

"A bibliographical tour of the large Sparse Generalised Eigenvalue Problem"
Sparse Matrix Comput. Bunch Acad

DEUXIEME PARTIE

TABLE DES MATIERES DE LA DEUXIEME PARTIE

INTRODUCTION

Le problème que nous avons étudié dans cette deuxième partie nous a été communiqué par Monsieur André Ronvaux, du Laboratoire de Physique Mathématiques et Physique du Solide de l'Université Notre Dame de la Paix à Namur, à l'occasion du Colloque d'Analyse Numérique 1978. [8].

Son origine se trouve dans l'intérêt, en Physique des phénomènes "de surface", de la connaissance de l'énergie associée à la présence d'un certain nombre (voir d'un réseau), de cavités près de la surface d'un matériau.

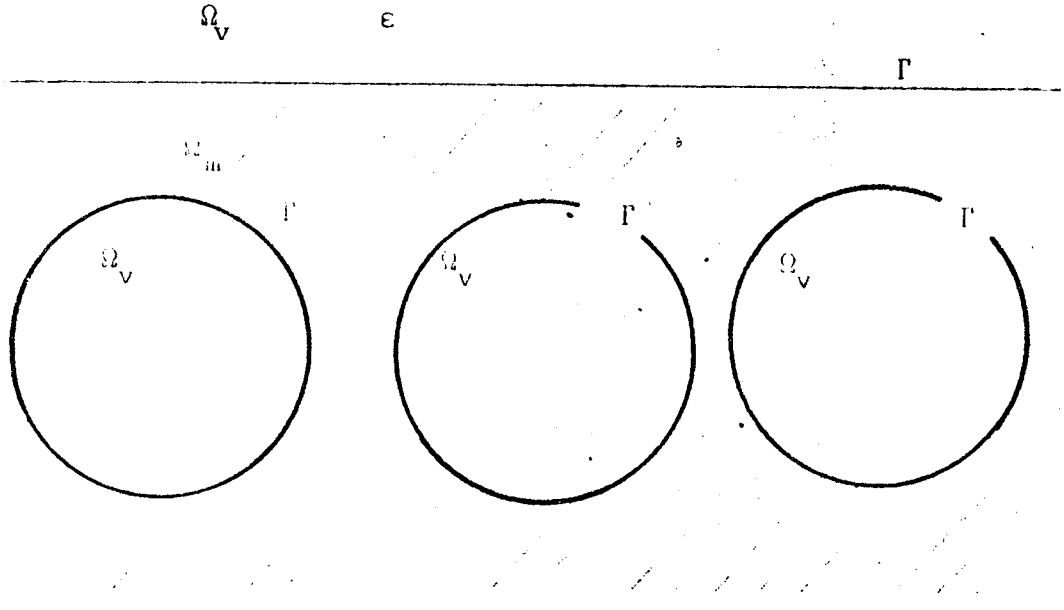
La géométrie du problème étant donnée, le calcul de cette énergie peut se ramener à un problème d'électrostatique dans lequel la fonction potentiel (notée U) et surtout la constante diélectrique du milieu, ε, sont à déterminer.

On distinguera deux domaines ouverts, en général non connexes de $\mathbb{R}^3$:

Ω_v domaine du vide, de constante diélectrique ε_o connue

Ω_m domaine du matériau, de constante diélectrique ε inconnue

Γ désignera l'ensemble des surfaces de séparation.



Le problème peut se poser de la manière suivante :

trouver les fonctions U , définies sur IR^3 , et continues ainsi que les valeurs ε telles que :

$$\begin{cases} \Delta U = 0 & \text{dans } \Omega_v \text{ et } \Omega_m \\ U \to 0 & \text{à l'infini} \\ \left.\frac{\partial u}{\partial n}\right|_v \cdot \varepsilon_o = \left.\frac{\partial u}{\partial n}\right|_m \cdot \varepsilon & \text{sur } \Gamma \end{cases}$$

Il est possible dans certains cas de résoudre directement le problème posé de cette manière en cherchant des solutions sous la forme d'un développement en série de fonctions harmoniques dans un sytème, de coordonnées adapté à la géométrie du problème.

Ronvaux et al [10]ont étudié le cas de N sphères alignées de même rayon, et régulièrement espacées. Ils ont utilisé N systèmes de coordonnées sphériques $(r_i, \theta_i, \varphi_i)$, ayant pour origine les centres des N sphères pour l'écriture des équations.
Le développement fait intervenir des termes du type :

$$r_i^e \; P_e^m[\cos \phi_i] \quad \cos m \phi_i$$

et le théorème d'addition des harmoniques sphériques joue un rôle fondamental dans l'écriture des conditions de continuité, et de transmission, puisqu'il permet le passage d'un système de coordonnées à un autre.

Les auteurs ont obtenus des résultats intéressants dans le cas où le rapport entre le diamètre des sphères et leur écartement est suffisamment petit pour qu'il soit licite de ne conserver dans les équations que la plus petite puissance de ce paramètre.

D'autres géométries ont également été étudiées par les mêmes auteurs [9], avec des techniques voisines.

Le but de notre travail est d'examiner une possibilité de traitement numérique de ce problème. Ceci nécessite de formuler le problème de manière plus adaptée.

On désignera par la suite par $q(x)$ la fonction suivante, définie pour $x \in \Gamma$.

$$q(x) = \frac{\partial u}{\partial n}\Big|_v - \frac{\partial u}{\partial n}\Big|_m$$

Alors, moyennant certaines hypothèses sur la décroissance de U à l'infini, on sait que U admet la représentation intégrale suivante :

$$U(x) = \frac{1}{4\pi} \int_\Gamma \frac{1}{|x-y|}\, q(y)\, dy$$

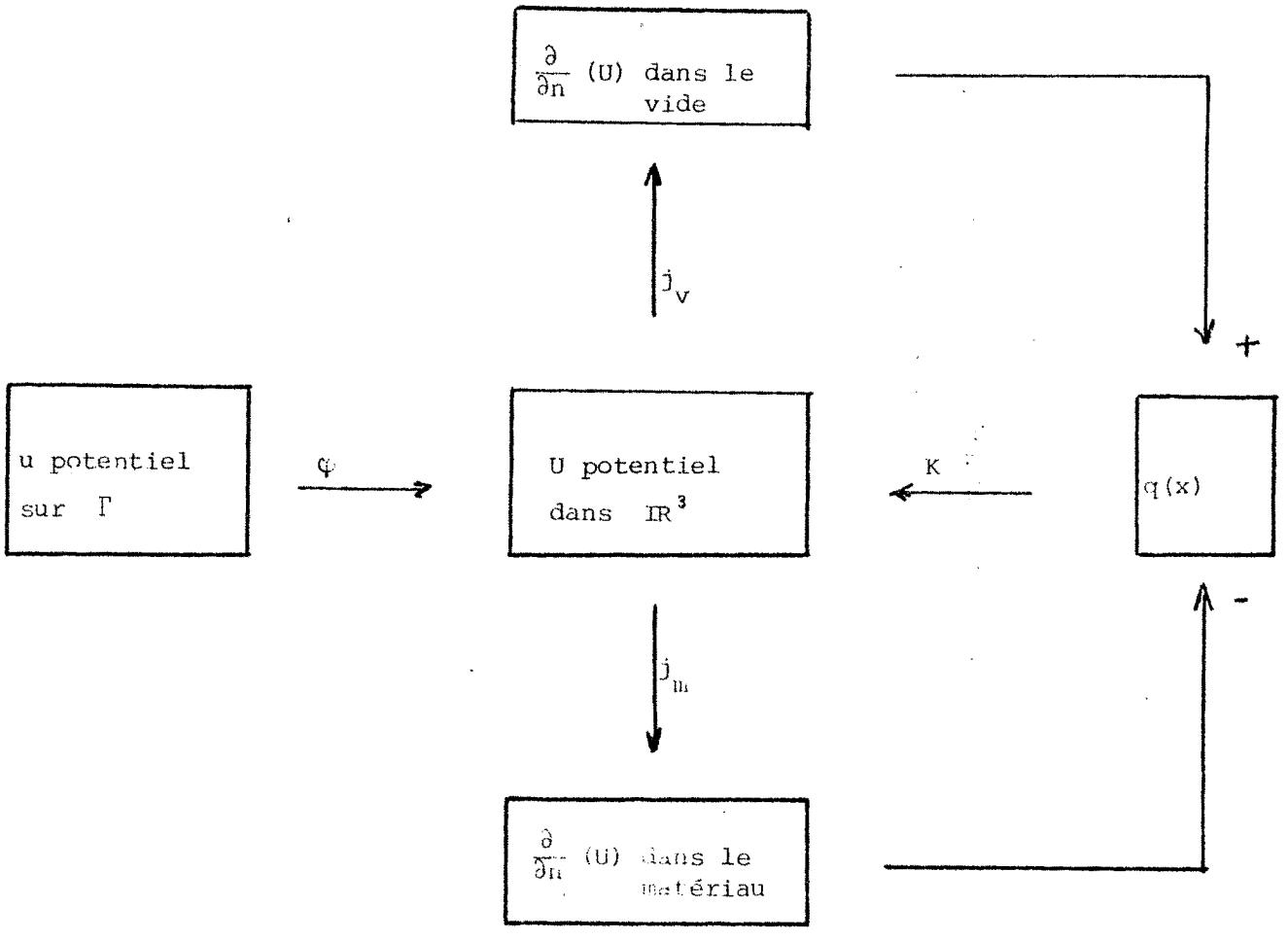
D'autre part, si on connait la valeur de u sur Γ seulement, et en supposant le problème "bien posé" (en mettant les bonnes conditions pour le compartiment de U à l'infini, et en supposant que les surfaces Γ présentent toute la régularité nécessaire), on peut résoudre le "problème de Dirichlet" non homogène suivant :

$$\begin{cases} \Delta U = 0 & \text{dans} \quad \mathbb{R}^3 \setminus \Gamma \\ U = u & \text{(donnée)} \quad \text{sur} \quad \Gamma \end{cases}$$

On désignera par φ l'opérateur qui, à u associe la solution (unique) du problème ; soit :

$$u \xrightarrow{\varphi} U$$

on a alors la situation suivante :



Deux approches sont, à priori en concurrence :

1) une première manière de prendre le problème consiste à écrire :

$$\varepsilon_m (j_m \circ K) q = \varepsilon_o [j_v \circ K] q$$

en utilisant l'expression explicite de K , on a :

$$(j_V \circ K)q(y) = \frac{1}{2} q(y) + \frac{1}{4\pi} \int_\Gamma q(x) \frac{\partial}{\partial n_y} \left| \frac{1}{|x-y|} \right| dx$$

$$(j_m \circ K)\ q(y) = -\frac{1}{2} q(y) + \frac{1}{4\pi} \int_\Gamma q(x) \frac{\partial}{\partial n_y} \left(\left| \frac{1}{|x-y|} \right| \right) dx$$

d'où la première formulation du problème :

$$(P) \quad \left.\begin{array}{l} \text{trouver } q(y) \\ \\ \text{et } \mu \text{ tels que :} \end{array}\right\} \quad \mu . q(y) = \int_\Gamma \frac{\partial}{\partial n_y} \left| \frac{1}{|x-y|} \right| q(x)\, dx$$

on a posé : $\mu = 2\pi \dfrac{\varepsilon+\varepsilon_o}{\varepsilon-\varepsilon_o}$

Signalons également une formulation faible associée à cette approche, mieux adaptée à l'utilisation de méthodes d'éléments finis sur Γ :

$$(P') \quad \left.\begin{array}{l} \text{trouver } q \in H^{-1/2}(\Gamma) \\ \text{et } \mu \text{ , tels que} \\ \forall\ \omega \in H^{1/2}(\Gamma) \text{ ,} \end{array}\right\} \quad \mu < q, \omega > = < \int_\Gamma \frac{\partial}{\partial n_x} \frac{1}{|x-y|} q(y)\, d\gamma_y, \omega >$$

avec $< q, \omega > = \int_\Gamma q.\omega \ d\gamma$

Cette approche est évidemment très séduisante, puisqu'elle permet de travailler avec des fonctions de deux variables. En fait il faut noter que :

- la description des domaines d'intégration n'est, en général pas facile ; ces domaines sont d'ailleurs en général infinis, à cause de la présence du plan/séparation

de

- le calcul des intégrales intervenant dans la formulation n'est pas simple non plus, en particulier en ce qui concerne les termes diagonaux

à cause de la présence de $\frac{\partial}{\partial n}\left|\frac{1}{|x-y|}\right|$

2) une seconde manière de poser le problème consiste à choisir pour fonction inconnue la restriction de la fonction potentiel à Γ .

$$(j_v \circ \varphi)u = \lambda (j_m \circ \varphi)u \qquad (Q_1)$$

on a posé $\lambda = \varepsilon/\varepsilon_o$

On aura, sur chaque exemple, à spécifier les espaces fonctionnels, V (sur Γ) et X (sur $\Omega_m \cup \Omega_v$) tels que :

1) si $u \in V$, il existe une solution φ unique dans X au problème suivant :

$$\begin{cases} \Delta\varphi = 0 \\ \varphi|_\Gamma = u \end{cases} \quad \text{dans } \Omega_m \text{ et } \Omega_v$$

2) les formes bilinéaires suivantes :

$$a(u,v) = \int_{\Omega_v} \nabla\varphi(u)\ \nabla\varphi(v)\ dx$$

$$b(u,v) = \int_{\Omega_m} \nabla\varphi(u)\ \nabla\varphi(v)\ dx$$

soient continues sur $V \times V$; l'une des deux sera coercive.

3) les solutions du problème suivant :

$$\boxed{\left.\begin{array}{l}\text{trouver } u \text{ et } \lambda \in V \times \mathbb{R} \\ \text{tels que } \forall\, v \in V\end{array}\right\} \ a(u,v) = \lambda\, b(u,v)} \qquad (Q)$$

puissent s'interpréter comme des solutions de (Q1). (Ceci étant possible par l'utilisation d'une formule de Green).

Une approximation par éléments finis comprend plusieurs étapes :

1) Approximation de la surface Γ par Γ_h, constituée de "faces" d'éléments finis de $\mathbb{R}^3$. (Si on utilise par exemple des tétraèdres Γ_h sera simplement une triangulation de Γ).

2) Approximation de l'espace V par un espace de dimension finie V_h, constitué de fonctions polynomiales par morceaux sur la surface Γ_h.

3) Résolution numérique des problèmes de Dirichlet par une méthode d'éléments finis
On a alors $\varphi_h(u_h)$ approximation de $\varphi(u_h)$.

4) On a alors les formes bilinéaires approchées :

$$a_h(u_h,v_h) = \int_{\Omega_h^v} \nabla\varphi_h(u_h)\ \nabla\varphi_h(v_h)\ dx$$

$$b_h(u_h,v_h) = \int_{\Omega_h^m} \nabla\varphi_h(u_h)\ \nabla\varphi_h(v_h)\ dx$$

et le problème approché :

$(\mathcal{Q}_h)$ trouver $(u_h,\lambda_h) \in V_h \in \mathbb{R}$ tels que $\forall\ v_h \in V_h$ $\Big\}$ $a_h(u_h,v_h) = \lambda_h\ ,\ b_h(u_h,v_h)$

Remarques :

Les étapes 1 et 3 précédentes font intervenir des approximations sur des domaines infinis.

En ce qui concerne le premier point, on pourra lever ce problème en imposant des conditions de périodicité. (Ce qui d'ailleurs est le moyen pratique de traiter des réseaux infinis).

En ce qui concerne l'étape 3, une première approximation consiste à imposer un potentiel nul sur une surface "assez éloignée" du domaine d'intérêt. Dans un calcul plus précis, on pourra utiliser une des techniques exposées par Céa et Grisvard au Colloque d'Analyse Numérique 1978. [1].

La suite de ce travail est consacrée à l'étude de la méthode que nous venons de décrire.

(A) ETUDE DE QUELQUES MODELES SIMPLIFIES

Nous avons, dans ce chapitre, rassemblés un certain nombre de problèmes simplifiés, que nous présentons par ordre de difficulté croissante. Cette étude est essentiellement destinée à donner une idée sur les questions suivantes :

1) - Quels sont les types de conditions nécessaires en ce qui concerne le comportement du potentiel à l'infini ?

2) - Quelle "taille" faut-il donner à un domaine tronqué, pour qu'il permette un calcul avec une précision suffisante des valeurs propres du problème posé dans un domaine infini ?

3) - Comment se comportent les solutions du problème lorsque le nombre de cavités augmente, et, en particulier, y-a-t-il convergence des valeurs propres vers une limite lorsque ce nombre tend vers l'infini ?

A1) Problèmes dans $\mathbb{R}$:

La valeur d'un problème modèle unidimentionnel est évidemment limitée, dans la mesure où le passage d'une dimension à une autre modifie grandement la structure du problème : Ces solutions élémentaires de l'équation :

$$\Delta u = 0$$

ne peuvent se déduire simplement les unes des autres, et les "bonnes" conditions de décroissance du potentiel à l'infini ne sont pas les mêmes.

Remarquons d'abord que, les seules fonctions continues et, à dérivée continue, satisfaisant à l'équation :

$$\forall\, x \in D \quad , \quad u''(x) = 0 \qquad (D \text{ partie connexe})$$

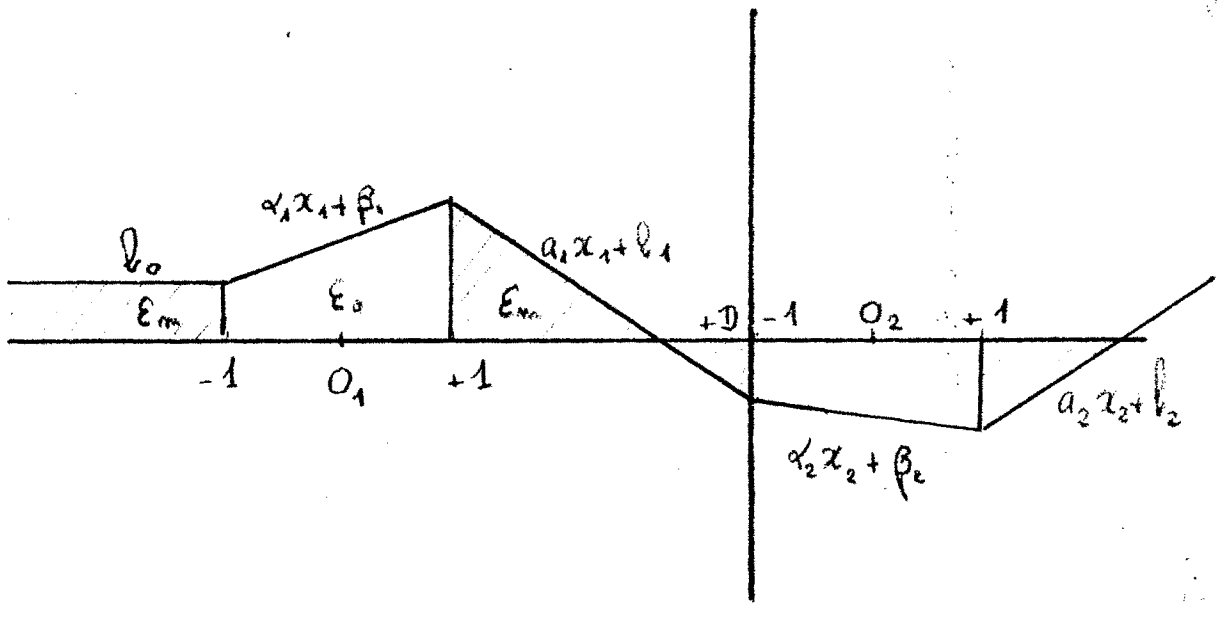
sont affines dans D . Il en résulte que, si nous imposons une condition limite du type :

$$\lim_{x \to \infty} u(x) = 0$$

Alors, les seules solutions possibles dans le cas d'un seul segment, sont identiquement nulles.
De manière moins restrictive, on imposera $\lim_{x \to \infty} |u(x)| < + \infty$
On envisagera deux cas :

A11) Système de N segments de même longueur et équidistants dans un domaine infini :

Pour faciliter l'écriture, on utilisera N système de coordonnées centrées sur les milieux des segments :



les conditions de transmissions s'écrivent de la manière suivante :

$$\alpha_k + \beta_k = a_k + b_k$$

$$\varepsilon_o \alpha_k = -\varepsilon_m a_k$$

$$a_k p + b_k = -\alpha_{k+1} + \beta_{k+1}$$

$$\varepsilon_m a_k = -\varepsilon_o \alpha_{k+1}$$

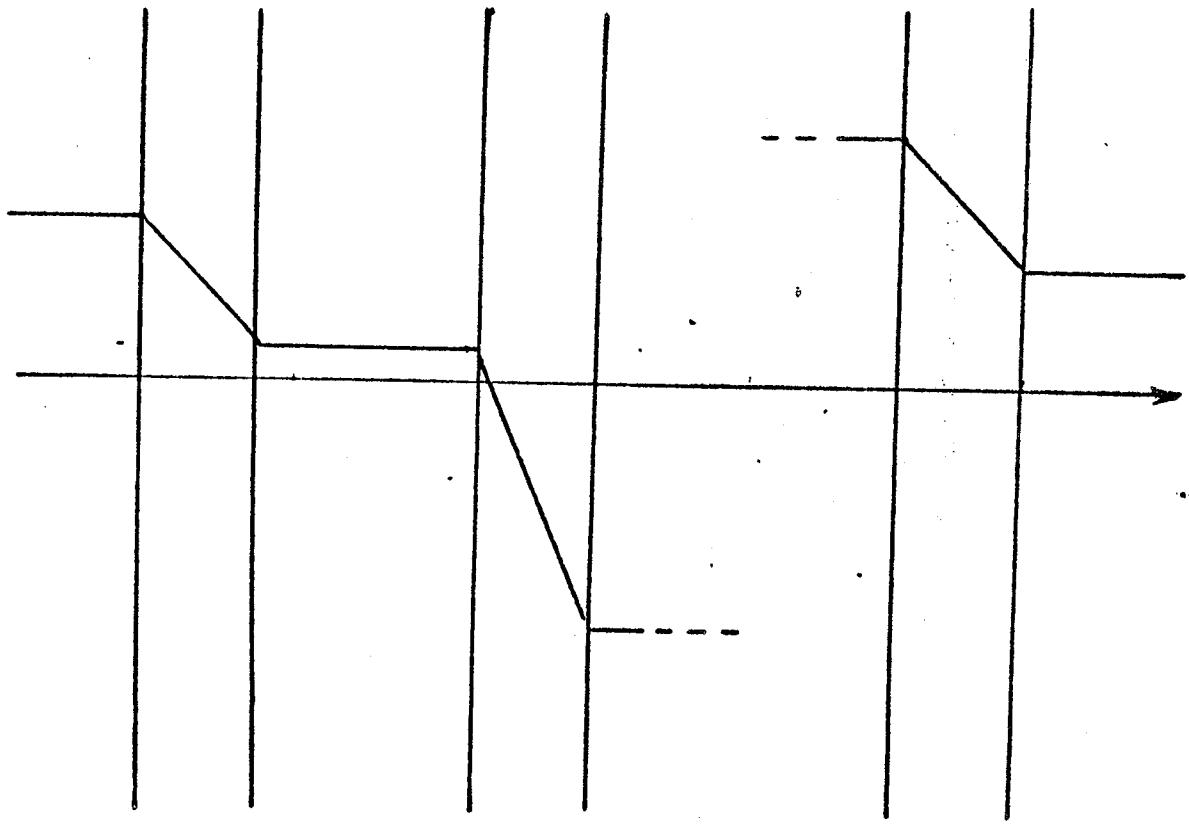
Les seules solutions (non constantes) possibles sont donc telles que :

$$[\varepsilon_m / \varepsilon_o]^{-1} = 0 \qquad \text{(non physique)}$$

les potentiels possibles sont alors tels que

$$a_i = 0 \qquad \forall\, i = 0,\ldots,N$$

les autres coefficients ayant des valeurs telles que la continuité soit assurée. Toutes les fonctions du types suivant conviennent :



A12 Système de N segments équidistants et de même longueur dans un domaine fini.

On imposera donc :

$$u\left(x_1 = -M\right) = 0 = u\left(x_N = +M\right)$$

soit :

$$-a_o M + b_o = a_N M + b_N = 0$$

en plus des conditions précédentes :

$$\begin{cases} \alpha_k + \beta_k = a_k + b_k \\ \varepsilon_o \alpha_k = -\varepsilon_m a_k \\ a_k D + b_k = -\alpha_{k+1} + \beta_{k+1} \\ \varepsilon_m a_k = -\varepsilon_o \alpha_{k+1} \end{cases}$$

il y a donc deux solutions évidentes :

$\varepsilon_o / \varepsilon_m = 0$. solution du cas précédent

$\varepsilon_m / \varepsilon_o = 0$ alors $\alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$

Dans le cas contraire, nous avons :

$$\alpha = \alpha_1 = \ldots = \alpha_N$$

$$a = a_o = \ldots = a_N$$

L'élimination des a_i et b_i donne le système suivant en $\alpha, \beta_1, \ldots, \beta_N$.

On pose $\lambda = \varepsilon_m / \varepsilon_o$

$$\begin{vmatrix} M-1-\lambda & \lambda & & & & \\ D-1-2\lambda & -\lambda & \lambda & & & \\ D-1-2\lambda & & \lambda & \lambda & & \\ \vdots & & & \ddots & \ddots & \\ D-1-2\lambda & & & & -\lambda & \lambda \\ M-1-\lambda & & & & & -\lambda \end{vmatrix} \begin{vmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \\ \beta_N \end{vmatrix} = 0$$

Le déterminant est facile à développer :

$$[(M-1-\lambda) + \sigma_{(2)}(-1)(D-1-2\lambda] + \ldots + \sigma_{(k)}(-1)^{k-1}(D-1-2\lambda)$$

$$+ \ldots + \sigma_{(N+1)}(-1)^{N}(M-1-\lambda)]\,\lambda^{N} = 0$$

où $\sigma_{(k)}$ désigne le signe de la permutation :

$$k,1,2,\ldots,k-1,k+1,\ldots,N+1$$

$$\sigma_{(k)} = (-1)^{k-1}$$

et $$P(\lambda) = \lambda^{N}[2(M-1-\lambda)+(N-1)(D-1-2\lambda)] = 0$$

comme $\lambda \neq 0$

$$\lambda = \frac{2(M+1)+(D-1)(N-1)}{2N}$$

Cette valeur propre est simple, et, admet la limite suivante lorsque le nombre de segments augmente tandis que M reste fixe :

$$\lambda_{\ell} = \frac{D-1}{2}$$

Sur cet exemple très simple, on constate toute l'importance du choix du type de conditions limites.

A2 Etude de cas simples dans $\mathbb{R}^2$

Les solutions de l'équation

$$\Delta u = 0$$

en coordonnées polaires, s'écrivent :

$$V = a_o \log r + b_o + \sum_{n>0} [a_n \cos(n\theta) + b_n \sin(n\theta)] \cdot [\alpha_n r^n + \beta_n r^{-n}]$$

A21 Problème avec une cavité circulaire centrée à l'origine dans un domaine infini

A l'intérieur, $\Delta u = 0$ et u reste bornée, donc le potentiel est de la forme :

$$V_{int} = b_o^i + \sum_{n>0} [a_n^i \cos(n\theta) + b_n^i \sin(n\theta)] r^n$$

A l'extérieur, l'expression du potentiel tendant vers zéro à l'infini est :

$$V_{ext} = \sum_{n>0} [c_n^e \cos(n\theta) + d_n^e \sin(n\theta)] r^{-n}$$

L'écriture des conditions de transmission , sur la circonférences $r = R$ détermine les coefficients, en utilisant les relations d'orthogonalité suivantes :

$$\int_0^2 \cos n\theta \sin m\theta \, d\theta = 0$$

$$\int_0^2 \cos n\theta \cos m\theta \, d\theta = \int_0^2 \sin n\theta \sin m\theta \, d\theta = \pi . \delta_{mn}$$

relations de "continuité" :

$$\begin{cases} b_o^i = 0 \\ a_n^i = C_n^e \, R^{-2n} \\ b_n^i = d_n^e \, R^{-2n} \end{cases}$$

Conditions de transmission :

$$- V'_{r\,int} = \lambda \, V'_{r\,ext}$$

$$\begin{cases} a_n^i \, R^{n-1} = + \lambda \, C_n^e \, R^{-(n+1)} \\ b_n^i \, R^{n-1} = + \lambda \, d_n^e \, R^{-(n+1)} \end{cases}$$

d'où les solutions :

$\lambda = +1$

$$V_{int} = \sum_{n>0} [a_n \cos(n\theta) + b_n \sin(n\theta)] \, r^n$$

$$V_{ext} = \sum_{n>0} [a_n \cos(n\theta) + b_n \sin(n\theta)] \, R^{2n} \, r^{-n}$$

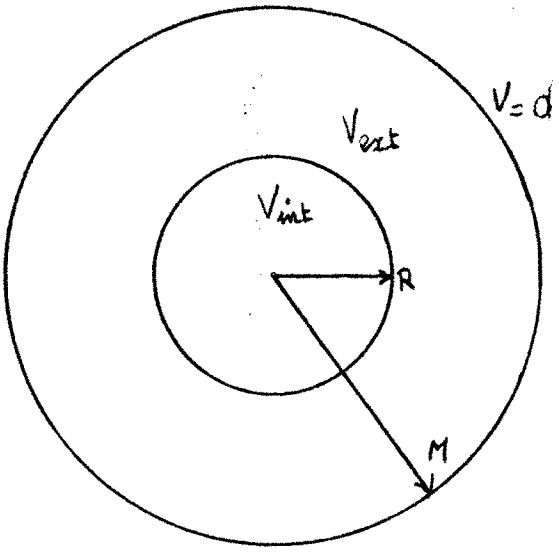
La valeur propre $\lambda = +1$ a donc une multiplicité infinie les coefficients a_n et b_n étant arbitraires (sous réserves que les séries convergent).

A22 Problème avec un cercle dans un domaine fini :

$$V_{ext} = \sum_{n>0} r^n (C_n \cos(n\theta) + d_n \sin(n\theta))$$

$$+ r^{-n} (\alpha_n \cos n\theta + \beta_n \sin n\theta)$$

$$+ a_o \log r + b_o$$



La condition $V(M) = 0$ s'écrit :

$$\begin{cases} M^n c_n + M^{-n} \alpha_n = 0 \\ M^n d_n + M^{-n} \beta_n = 0 \\ a_o \log M + b_o = 0 \end{cases}$$

donc :

$$V_{ext} = \sum_{n>0} c_n \cos(n\theta) [r^n - M^{2n} r^{-n}] + d_n \sin(n\theta) [r^n - M^{2n} r^{-n}]$$

$$+ a_o \log \frac{r}{M}$$

$$V_{int} = \sum_{n>0} [a_n \cos n\theta + b_n \sin n\theta] r^n + c_o$$

par continuité :

$$\begin{cases} c_n (R^n - M^{2n} R^{-n}) = a_n R^n \\ d_n (R^n - M^{2n} R^{-n}) = b_n R^n \\ a_o \log \left(\frac{R}{M}\right) = c_o \end{cases}$$

La condition de transmission donne :

$$n \lambda_n c_n [R^{n-1} + M^{2n} R^{-(n+1)}] = -n a_n R^{n-1}$$

$$n \lambda_n d_n [R^{n-1} + M^{2n} R^{-(n+1)}] = -n b_n R^{n-1}$$

$$\lambda_n = \frac{(M/R)^{2n} - 1}{(M/R)^{2n} + 1} \qquad n \neq 0$$

d'où :

$$V_{int} = [a_n \cos(n\theta) + b_n \sin(n\theta)] r^n$$

$$V_{ext}^{(n)} = a_n \left(\frac{R^{2n}}{R^{2n} - M^{2n}}\right) \cos(n\theta) \cdot (r^n - M^{2n} r^{-n})$$

$$+ b_n \left(\frac{R^{2n}}{R^{2n} - M^{2n}}\right) \sin(n\theta) \cdot (r^n - M^{2n} r^{-n})$$

Chaque valeur propre λ_n est "donc" double, sauf λ_o :

$$\begin{cases} \lambda_o = 0 \\ V_{int}^{(o)} = a_o \log \frac{R}{M} \\ V_{ext}^{(o)} = a_o \log \frac{r}{M} \end{cases}$$

Remarque : en'tronquant' l'espace, on ajoute la valeur propre 0 , et on "déplace" substantiellement la valeur propre +1.

Par exemple, si $M = 3R$ $\quad \lambda_1 = + 0.8$

$$\lambda_2 = + \frac{80}{82}$$

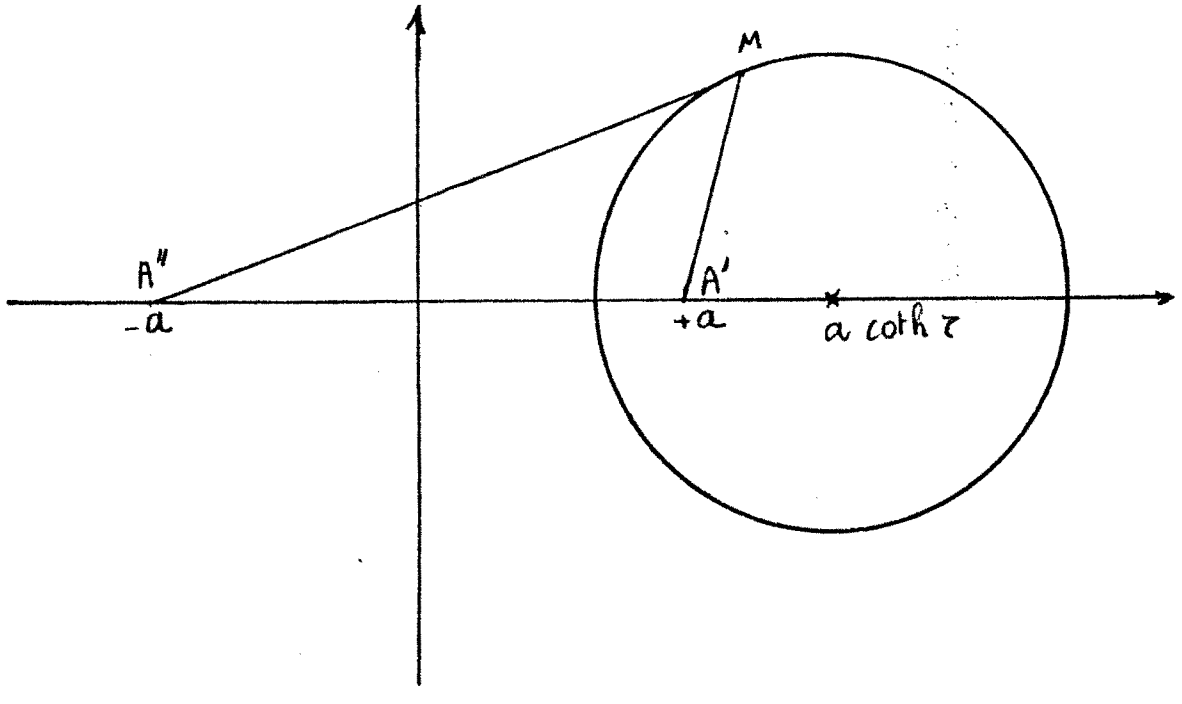
pour avoir, par exemple une erreur inférieure ou égale à 1 % , il faut donc prendre au moins $M = 14\ R$.

A23 Problème avec deux cercles de même diamètre

On utilisera un système de coordonnées permettant d'exprimer de manière commode les conditions de transmission. [ ]

L'équation d'un faisceau de cercles, dont les points limites sont $-a$ et $+a$ est donnée par :

$$(x - a \coth \tau)^2 + y^2 = \frac{a^2}{\mathrm{sh}^2 \tau}$$



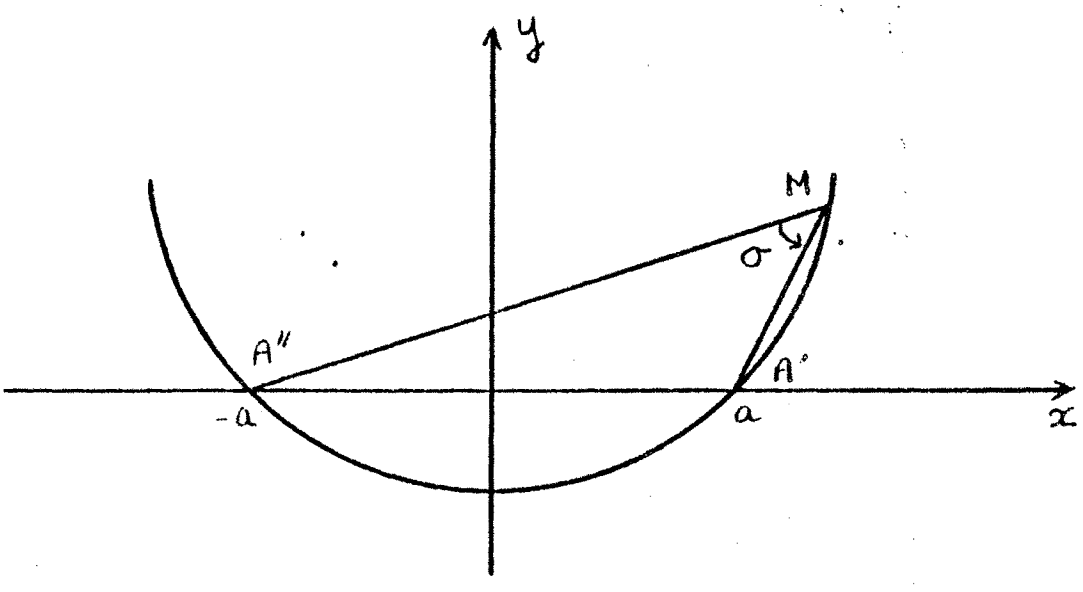
Le paramètre τ pouvant être déterminé géométriquement par la relation :

$$\frac{MA''}{MA'} = e^{2\tau}$$

l'équation du faisceau conjugué (orthogonal) s'écrit :

$$x^2 + (y - a \operatorname{cotg} \sigma)^2 = \frac{a^2}{\sin^2 \sigma}$$

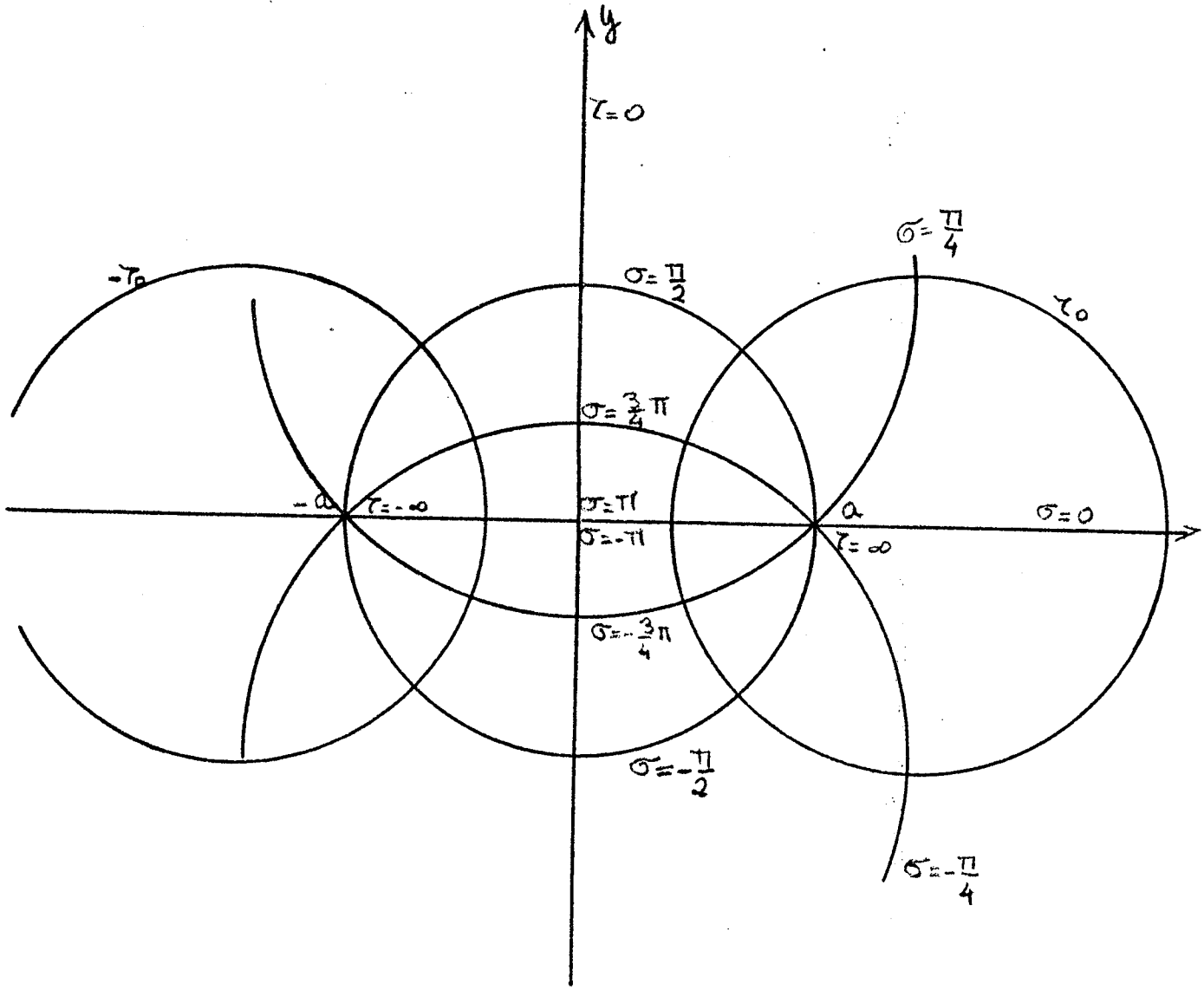
ici, σ désigne "l'angle" (MA",MA') (Mod 2π)



On utilisera le système de coordonnées défini par les paramètres σ et τ . L'intersection d'un cercle σ = cte avec un cercle τ = cte ayant, en général deux points, on limitera les ensembles σ = cte > 0 (σ compris entre 0 et π) , en ne conservant que les parties au-dessus de l'axe x'Ox . Les valeurs négatives de y correspondront à σ compris entre $-\pi$ et 0 .

Le changement de coordonnées est donné par les formules :

$$\begin{cases} x = a \dfrac{\operatorname{sh} \tau}{\operatorname{ch} \tau - \cos \sigma} \\ y = a \dfrac{\sin \sigma}{\operatorname{ch} \tau - \cos \sigma} \end{cases} \qquad \begin{cases} \sigma = \dfrac{i}{2} \operatorname{Log} \dfrac{x^2 + (y-ia)^2}{x^2 + (y+ia)^2} \\ \tau = \dfrac{1}{2} \operatorname{Log} \dfrac{(x+a) + y}{(x-a) + y} \end{cases}$$



les points $\begin{cases} x = \pm a \\ y = 0 \end{cases}$ sont des points de discontinuité pour σ .

De même, le franchissement du segment A" O A' s'accompagne du passage de $\sigma = -\pi$ à $\sigma = +\Gamma$, les fonctions $F(\sigma,\tau)$ seront donc obligatoirement périodique en σ , avec une périodicité $T = 2\pi$.

L'expression du Laplacien dans ce système de coordonnées est :

$$\Delta\phi = \frac{1}{a^2}(\operatorname{ch}\tau - \cos\sigma)^2\left(\frac{\partial^2\phi}{\partial\sigma^2} + \frac{\partial^2\phi}{\partial\tau^2}\right)$$

Remarque : la quantité

$$\operatorname{ch}\zeta - \cos\sigma$$

s'annulle uniquement pour $\tau = \sigma = 0$, qui est un "point" à l'infini.

On aura donc à résoudre :

$$\frac{\partial^2\phi}{\partial\sigma^2} + \frac{\partial^2\phi}{\partial\tau^2} = 0$$

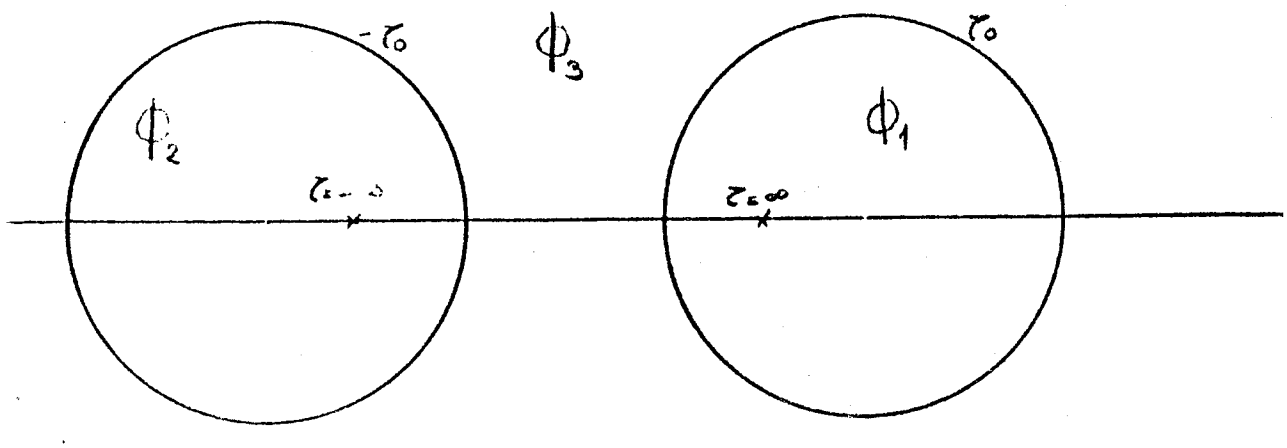
les solutions, périodiques en σ , et de la forme :

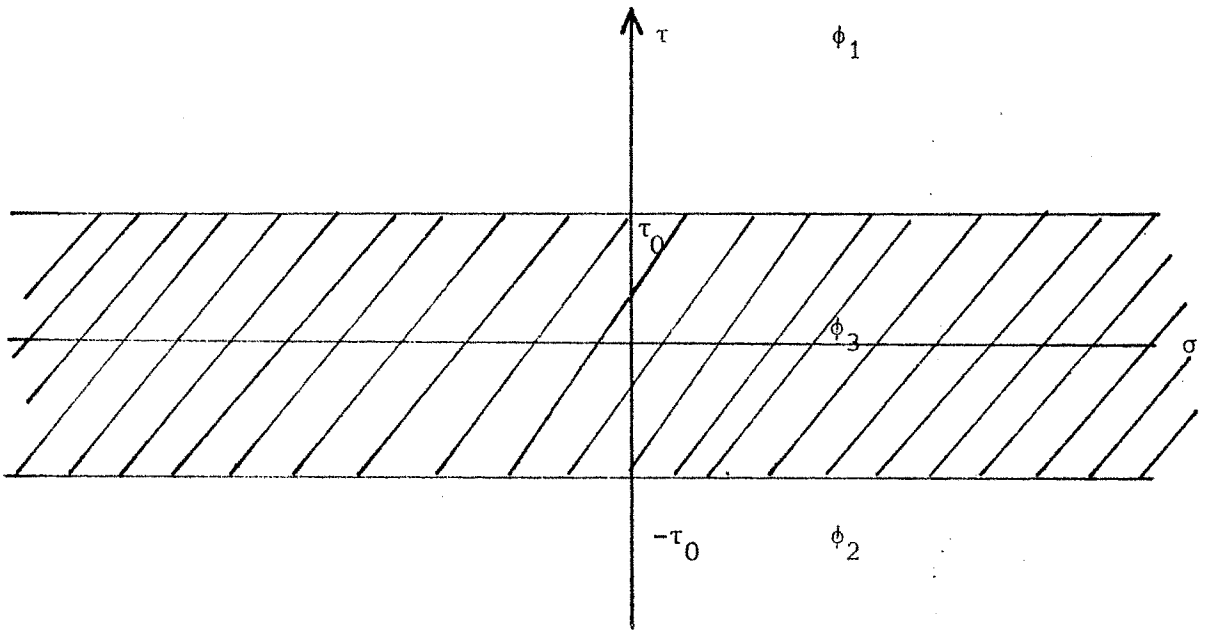
$$\phi(\sigma,\tau) = \sum_{n>0} f_n(\sigma) \cdot g_n(\tau)$$

sont facilement identifiables :

$$\phi = \alpha_o\tau + \beta_o + \sum_{n=1}^{\infty}\left[a_n\cos(n\sigma) + b_n\sin(n\sigma)\right](\alpha_n e^{-n\tau} + \beta_n e^{n\tau})$$

la géométrie du problème est la suivante :





Les fonctions, harmoniques en chaque secteur, et bornées pour $\tau = \pm\infty$ s'écrivent :

$$\phi_1 = \beta_o^1 + \sum_{n>0} e^{-n\tau}[a_n^1 \cos(n\sigma) + b_n^1 \sin(n\sigma)]$$

$$\phi_2 = \beta_o^2 + \sum_{n>0} e^{n\tau}[a_n^2 \cos(n\sigma) + b_n^2 \sin(n\sigma)]$$

$$\phi_3 = \alpha_o^3 \tau + \beta_o^3 + \sum_{n>0} e^{n\tau}[a_n^3 \cos(n\sigma) + b_n^3 \sin(n\sigma)] +$$

$$+ e^{-n\tau}[a_n^4 \cos(n\sigma) + b_n^4 \sin(n\sigma)]$$

L'écriture des conditions de continuité donne :

* sur Γ_{13}

$$\begin{cases} \beta_o^1 = \alpha_o^3 \tau_o + \beta_o^3 \\ a_n^1 = e^{2n\tau_o} a_n^3 + a_n^4 \\ b_n^2 = e^{2n\tau_o} b_n^3 + b_n^4 \end{cases}$$

* sur Γ_{23}

$$\begin{cases} \beta_o^2 = -\alpha_o^3 \tau_o + \beta_o^3 \\ a_n^2 = a_n^3 + e^{2n\tau_o} a_n^4 \\ b_n^2 = b_n^3 + e^{2n\tau_o} b_n^4 \end{cases}$$

les conditions de transmission portent sur les dérivées de ϕ par rapport à la variable τ ,

* sur $\Gamma_{1/3}$

$$\begin{cases} \alpha_o^3 \cdot \lambda = 0 \\ a_n^1 = \lambda\,[-a_n^4 + e^{2n\tau_o} a_n^3] \\ b_n^1 = \lambda\,[-b_n^4 + e^{2n\tau_o} b_n^3] \end{cases}$$

* sur Γ_{23}

$$\begin{cases} \alpha_o \lambda = 0 \\ a_n^2 = \lambda(-a_n^3 + e^{2n\tau_o} a_n^4) \\ b_n^2 = \lambda(-b_n^3 + e^{2n\tau_o} b_n^4) \end{cases}$$

① valeur propre $\lambda = 0$

Il est facile de voir que, dans ce cas

$$a_n^i = 0 \quad , \quad \forall\, i = 1,2,3,4 \quad , \quad \forall\, n \geq 1$$

donc : $\phi_1^{(0)} = \alpha_o^3 \tau_o + \beta_o^3$

$\phi_2^{(0)} = - \alpha_o^3 \tau_o + \beta_o^3$

$\phi_3^{(0)} = \alpha_o^3 \tau + \beta_o^3$

la condition de décroissance à l'infini pouvant être obtenue si on choisi

$$\beta_o^3 = 0$$

② Un ensemble de valeurs a_n^i , b_n^i , i=1,...,4 , peut être choisi satisfaisant aux conditions écrites, si on a la condition :

$$\begin{vmatrix} 1 & 0 & -\bar{e} & -1 \\ 0 & 1 & -1 & -\bar{e} \\ 1 & 0 & -\lambda\bar{e} & +\lambda \\ 0 & 1 & \lambda & -\lambda\bar{e} \end{vmatrix} = 0$$

On a posé $\bar{e} = e^{2n\tau_o}$

le déterminant est facile à développer :

$$\lambda^2 - 2\lambda \frac{\bar{e}^2+1}{\bar{e}^2-1} + 1 = 0$$

ses racines sont :

$$\lambda_1 = \frac{(1-\bar{e})^2}{\bar{e}^2-1} \qquad \lambda_2 = \frac{(\bar{e}+1)^2}{\bar{e}^2-1}$$

d'où $\lambda_1^n = 2 \dfrac{[\text{sh } n\,\tau_o]^2}{\text{sh}(2n\tau_o)}$ $\lambda_2^n = 2 \dfrac{[\text{ch}(n\,\tau_o)]^2}{\text{sh } (2n\,\tau_o)}$

avec $\tau_o = \text{Arg ch } (\frac{D}{R})$

où D est la moitié de la distance entre les deux centres des cercles, et R leur rayon .

Comparaison des premières valeurs propres pour différentes valeurs de D/R

D/R	1.5	3	10	100
τ_o	0,962	1,763	2,993	5,298
$\lambda_1^{(1)}$	0.7453	0.94	0.995	
$\lambda_2^{(2)}$	1,34	1,06	1,005	
$\lambda_1^{(2)}$	9.58			
$\lambda_2^{(2)}$	1.04			

A3) Etude dans $\mathbb{R}^3$:

A31 Cas d'une seule sphère :

L'expression en coordonnées sphériques d'un potentiel harmonique, borné à l'intérieur, et tendant vers zéro à l'infini est donnée par [10]

$$V^{int} = \sum_{\ell=0}^{\infty} \sum_{m=0}^{\ell} r^{\ell} \left[a_{\ell m} Y^{m,1}(\theta,\varphi) + b_{\ell m} Y^{m,2}(\theta,\varphi) \right]$$

$$V^{ext} = \sum_{\ell=0}^{\infty} \sum_{m=0}^{\ell} r^{-(\ell+1)} \left[c_{\ell m} Y^{m,1}(\theta,\varphi) + d_{\ell m} Y^{m,2}(\theta,\varphi) \right]$$

avec :

$$Y_{\ell}^{m,1}(\theta,\varphi) = P_{\ell}^{m}(\cos\theta) \cos m\phi$$

$$Y_{\ell}^{m,2}(\theta,\varphi) = P_{\ell}^{m}(\cos\theta) \sin m\phi$$

ces fonctions vérifiant les conditions d'orthogonalité suivantes :

$$\int_0^{2\pi} \int_0 Y_\ell^{m,i}(\theta,\varphi)\ Y_{\ell'}^{m',j}(\theta,\varphi)\quad \sin\theta\, d\theta\, d\varphi = K(\ell,m)\quad \delta_{\ell\ell'} \cdot \delta_{mm'} \cdot \delta_{ij}$$

avec $K(\ell,m) = \dfrac{4\pi}{\varepsilon(m)}\ \dfrac{(\ell+m)!}{(\ell-m)!}\ \dfrac{1}{2\ell+1}$; $\varepsilon(m) = 2 \quad (m\neq 0)$; $\varepsilon(0) = 1$

L'écriture des conditions de continuité et de transmission pour une sphère de rayon R donne les valeurs propres suivantes :

$$\lambda_\ell = \frac{\ell}{\ell+1} \qquad \ell = 0,1\ldots$$

Ces valeurs s'accumulent en 1, et ont la multiplicité $2\ell+1$. Les vecteurs propres s'écrivent :

$$V_\ell^{int} = r^\ell\ G_\ell(\theta,\varphi)$$

$$V_\ell^{ext} = r^{-(\ell+1)}\ R^{2\ell+1}\ G_\ell(\theta,\varphi)$$

Où $G_\ell(\theta,\varphi)$ est une quelconque combinaison linéaire des Y_ℓ^m :

$$G_\ell(\theta,\varphi) = \sum_{m=0}^{\ell} a_{\ell_m}\ Y_\ell^{m1}(\theta,\varphi) + b_{\ell_m}\ Y_\ell^{m2}(\theta,\varphi)$$

A32 une cavité sphérique dans un domaine limité par une sphère de rayon M , avec condition

$$V = 0 \quad \text{pour} \quad r = M$$

$$V^{int} = \sum_{\ell\geq 0} \sum_{m=0}^{\ell} r^\ell\ [a^1_{\ell m}\ Y_\ell^{m1}(\theta,\varphi) + \ell^1_{\ell m}\ Y_\ell^{m2}(\theta,\varphi)]$$

$$V^{ext} = \sum_{\ell\geq 0} \sum_{m=0}^{\ell} r^\ell\ [a^2_{\ell m}\ Y^{m1}(\theta,\varphi) + \ell^2_{\ell m}\ Y_\ell^{m2}(\theta,\varphi)] +$$

$$+ \sum_{k\leq 0} \sum_{m=0}^{k} r^{-(l+1)} [c_{\ell m}\ Y^{m1}(\theta,\varphi) + d_{\ell m}\ Y^{m2}(\theta,\varphi)]$$

$$V^{ext} = 0 \ , \ r = M \Rightarrow \quad \begin{cases} M^{\ell} \, a^2_{\ell m} = M^{-(\ell+1)} \, c_{\ell m} \\ M^{\ell} \, b^2_{\ell m} = M^{-(\ell+1)} \, d_{\ell m} \end{cases}$$

continuité pour $r = R$

$$\begin{cases} R^{\ell} \, a^1_{\ell m} = R^{\ell} \, a^2_{\ell m} + R^{-(\ell+1)} \, c_{\ell m} \\ R^{\ell} \, b^1_{\ell m} = R^{\ell} \, b^2_{\ell m} + R^{-(\ell+1)} \, d_{\ell m} \end{cases}$$

condition de transmission :

$$\begin{cases} \ell \, R^{\ell-1} \, a_{\ell m} = - \lambda \, \ell \, R^{\ell-1} \, a^2_{\ell m} + \lambda \, (\ell+1) \, R^{-(\ell+2)} \, c_{\ell m} \\ \ell \, R^{\ell-1} \, b_{\ell m} = - \lambda \, \ell \, R^{\ell-1} \, b^2_{\ell m} + \lambda \, (\ell+1) \, R^{-(\ell+2)} \, d_{\ell m} \end{cases}$$

on a donc les valeurs propres :

$$\lambda_{\ell} = \frac{[1+(R/M)^{2\ell+1}] \, \ell}{\ell[1-(R/M)^{2\ell+1}]+1}$$

Les multiplicités sont les mêmes que précédemment.

Nous donnons les valeurs numériques des premières valeurs propres pour différentes valeurs de M/R dans le tableau suivant :

	domaine infini	M/R=2	=4	6	8	10
$\ell=0$	0	0	0	0	0	0
$\ell=1$	0,5	0,6	0,512	0,5035	0,5015	0,5007
$\ell=2$	0,66666	0,702	0,668	0,6668	0,6667	0,66668
$\ell=3$	0,75	0,760	0,75008	0,750005		
$\ell=4$	0,8	0,803	$0{,}8+10^{-6}$			
$\ell=5$	0,8333...	0,834				
$\ell=6$	0,85714	0,85719				
$\ell=7$	0,875	0,87502				

(B) FORMULATION FAIBLE, CONDITIONS LIMITES

Ce chapitre a deux objectifs :

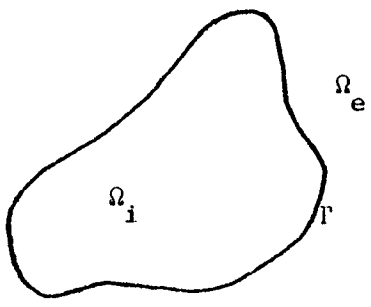
- D'une part, nous avons essayés de donner un sens à la formulation faible introduite précédemment, en précisant les espaces fonctionnels dans un cas particulier simple.

Les domaines que l'on a à considérer sont, en général non bornés, et il en résulte que les cadres fonctionnels ne sont pas forcément standards. Nous allons détailler ceci, en se limitant à un problème simplifié dans $\mathbb{R}^3$, pour lequel un tel cadre existe déjà.

- D'autre part, on a cherché à définir une méthode, permettant de résoudre pratiquement la difficulté liée à la présence du domaine infini ; dans un cas géométrique plus réaliste, c'est à dire dans le cas d'un réseau infini périodique.

B1 Problème avec une seule cavité dans $\mathbb{R}^3$

Soit donc Ω_i un ouvert borné de $\mathbb{R}^3$ de frontière Γ "assez régulière" et Ω_e le complémentaire de $\overline{\Omega_i}$.



Nous devons spécifier :

- un espace V (Hilbertien) de fonctions définies sur Γ ,
- un espace X_i de fonctions définies sur Ω_i , et X_e sur Ω_e , tels que $\forall\ u \in V$, les solutions des problèmes :

$$\Delta\, \varphi^i = 0 \quad (\Omega_i) \qquad\qquad \Delta\, \varphi^e = 0 \quad (\Omega_e)$$
$$\varphi^i|_\Gamma = u \qquad\qquad \varphi^e|_\Gamma = u$$

existent, soient uniques dans X_i et X_e , et dépendent continuement de u

nous devons pour ces solutions, pouvoir définir les dérivées normales au voisinage de Γ , et pouvoir utiliser les formules de Green suivantes :

$$\forall\ \omega \in X_i\ ,\quad \int_{\Omega_i} \nabla\varphi^i\ \nabla\omega\ dx = \int_\Gamma \frac{\partial\varphi^i}{\partial n} \cdot \omega\ \ d\gamma$$

$$\forall\ \omega \in X_e\ ,\quad \int_{\Omega_e} \nabla\ \varphi^e\ \nabla\omega\ \ dx = \int_\Gamma \frac{\partial\varphi^e}{\partial x}\ \ \omega\ \ d\gamma$$

* pour le problème dans Ω_i , avec :

$X_i = H^1(\Omega_i)$, muni de la norme habituelle

$V = H^{1/2}(\Gamma)$, espace des "traces" des fonctions de $H^1(\Omega_i)$ muni de la norme suivante :

$$|u|_V = \inf_{\substack{\Psi\in H^1(\Omega_i) \\ \gamma_o^i(\Psi)=u}} ||\Psi||_{1,\Omega_i}$$

(ici et par la suite, γ_o^i désignera l'application trace), alors l'existence et la continuité de $\varphi_i(u)$ est acquise ; et, pour tout u dans V , $\frac{\partial}{\partial n}\ \varphi_i(u)$ est dans $H^{-1/2}(\Gamma)$, dual de $H^{1/2}(\Gamma)$, avec :

$$< \frac{\partial\varphi_i}{\partial n}\ (u), v > = \int_\Gamma \frac{\partial\varphi_i(u)}{\partial n}\ v\ d\gamma = \int_{\Omega_i} \nabla\ \varphi_i(u)\ \nabla\varphi_i(v)\ \ dx = a(u,v)$$

Par l'utilisation d'une formule de Green. La forme ainsi définie est continue sur $V \times V$.

* Le problème dans Ω_e est plus délicat, il nécessite l'utilisation d'espace de Sobolev avec Poids, dont on trouvera une description détaillée dans [5] , [7] . On se limitera ici à rappeler des résultats utiles dans notre contexte.

Définitions :

1) $\mathcal{D}(\bar{\Omega}_e)$ est l'ensemble des restrictions à Ω_e des fonctions indéfiniement différentiables et à support compact dans $\mathbb{R}^3$.

2) $W_o^1(\Omega^e)$ est alors défini comme le complèté de $\mathcal{D}(\bar{\Omega}_e)$ pour la norme suivante :

$$\| \Psi \|_{1 o \Omega^e} = \left[\int_{\Omega^e} \left[\left| \frac{\Psi}{\sqrt{1+r^2}} \right|^2 + |\nabla \Psi|^2 \right] dx \right]^{1/2}$$

Muni de cette norme $W_o^1(\Omega^e)$ est un espace de Hilbert.

Remarque : Les propriétés locales de $W_o^1(\Omega^e)$ sont les mêmes que pour les espaces de Sobolev usuels. Le théorème des traces s'applique donc encore.

En particulier, $V = H^{1/2}(\Gamma)$ coïncide encore avec l'espace des traces des fonctions de $W_o^1(\Omega^e)$ sur Γ , et, on peut montrer que la norme :

$$\|v\|' = \inf_{\substack{\Psi \in W_o^1(\Omega^e) \\ \gamma^e(\Psi) = v}} \|\Psi\|_{1,o,\Omega^e}$$

est équivalente à celle définie sur V par le problème dans Ω^i .

D'autre part, v étant donnée dans $H^{1/2}(\Gamma)$, on peut choisir un relèvement IR_o de v dans $W_o^1(\Omega^e)$, tel que l'application :

$$v \to IR_o^e(v)$$

soit continue.

Définition : On notera $\mathring{W}_o^1(\Omega^e)$ le complété de $\mathcal{D}(\Omega^e)$ pour la norme $\| \ \|_{1,o,\Omega^e}$.

C'est encore le sous espace des fonctions de $W_o^1(\Omega^e)$ dont la "trace" sur Γ est nulle.

Propriété : La forme :

$$(\Psi_1, \Psi_2)_1 = \int_{\Omega e} \nabla \Psi_1 \nabla \Psi_2 \, dx$$

est coercive sur l'espace $W_o^1(\Omega^e)$
c'est-à-dire, :

$$\exists\, C > 0 \text{ tel que } \forall \Psi \in W_o^1(\Omega^e), \ (\Psi,\Psi)_1 \geq C \| \Psi \|^2_{1,o,\Omega^e} .$$

Remarque : Dans IR^2, cette propriété n'est vraie que pour $\mathring{W}_o^1$.
Alors, le problème suivant :

trouver $\Psi \in \mathring{W}_o^1(\Omega^e)$ tel que $\forall \omega \in \mathring{W}_o^1(\Omega^e)$

$$\int_{\Omega^e} \nabla \Psi \nabla \omega \, dx = - \int_{\Omega^e} \nabla R_o(u) \nabla \omega \, dx$$

admet une solution unique et

$$\varphi^e(u) = \Psi + R_o^e(u)$$

La continuité de $\varphi^e : V \to X^e$ ainsi définie est donc assurée.

Propriété : $\forall u \in V$ et $\forall \omega \in W_o(\Omega^e)$ on a la formule de Green suivante :

$$\int_{\Omega^e} \nabla \varphi(u) \nabla \omega \, dx = \int_\Gamma \frac{\partial \varphi_{(u)}}{\partial n} \cdot \omega \, d\gamma$$

en effet, elle est vraie pour toute fonction $\omega \in \mathcal{D}(\overline{\Omega^e})$, et nous pouvons l'étendre à $\omega \in W_o^1(\Omega^e)$ par continuité et densité

* Nous venons donc, sur cet exemple, de justifier l'utilisation de la formulation :

. trouver $u, \lambda \in V \times \mathbb{R}$ tels que :

$$\forall\, v \in V \qquad a(u,v) = \lambda\, b(u,v)$$

On remarquera que la forme $b(u,v)$ est coercive sur V :

$$b(u,u) = \int_\Omega |\nabla\, \varphi(u)|^2\, dx \geq c\, \|\varphi(u)\|^2_{1,0,\Omega^e} \geq c \cdot \|u\|^2_{1/2}$$

La première inégalité résulte d'une propriété déjà énoncée, et la seconde de la définition de la norme de $H^{1/2}(\Gamma)$.

Les deux formes $a(u,v)$ et $b(u,v)$ sont continues sur $V \times V$:

$$a(u,v) \leq |\varphi_i(u)|_{1,\Omega_i} \times |\varphi_i(v)|_{1,\Omega_i}$$

$$b(u,v) \leq |\varphi_e(u)|_{1,\Omega_e} \times |\varphi_e(v)|_{1,\Omega_e}$$

Comme $\varphi_i(u)$ et $\varphi_e(u)$ sont les uniques solutions des problèmes :

$$(1) \quad \begin{cases} \gamma_o^i\, \varphi_i = u \\ \forall\, \omega \in H^1_o(\Omega_i)\ ,\ \int_{\Omega_i} \nabla\, \varphi_i\ \nabla\, \omega\, dx = 0 \end{cases}$$

$$(2) \quad \begin{cases} \gamma_o^e\, \varphi_e = u \\ \forall\, \omega \in \mathring{W}_o(\Omega_e),\ \int_{\Omega_e} \nabla\, \varphi_e\ \nabla\, \omega\, dx = 0 \end{cases}$$

donc $$|\varphi_i(u)|_{1,\Omega_i} = \underset{\substack{\gamma_o^i \Psi = u \\ \Psi \in H^1(\Omega_i)}}{\text{Min}} |\Psi|_{1,\Omega_i} \leq |R_o(u)|_1 \leq \|u\|_{1/2} \times K$$

$$|\varphi_e(u)|_{1,\Omega_e} = \underset{\substack{\gamma_o^e \Psi = u \\ \Psi \in W_o^1(\Omega_e)}}{\text{Min}} |\Psi|_{1,\Omega_e} \leq R_o(u) \leq \|u\|_{1/2} \times K$$

ce qui montre la propriété.

Donc $b(u,u)^{1/2}$ est une norme sur V, équivalente aux deux normes précédemment envisagées. Il en résulte que nous pouvons définir l'opérateur T :

$$T : V \to V$$

par : Tu solution unique du problème :

$$\forall\, v \in V \qquad b(Tu,v) = a(u,v)$$

T ainsi défini est continu, et autoadjoint. Les solutions du problème variationnel sont les éléments propres de T. Nous allons donner une interprétation simple de T :

On remarquera d'abord que, par une propriété donnée précédemment, il est possible de résoudre le problème suivant dans Ω^e :

$$\begin{cases} \Delta \Psi = 0 \text{ dans } \Omega^e \\ \dfrac{\partial}{\partial n} \Psi = z \text{ sur } \Gamma \quad (z \text{ donnée}) \end{cases}$$

soit Ψ^e l'application ainsi définie :

$$\Psi^e : H^{-1/2}(\Gamma) \to W_o^1(\Omega^e)$$

Nous allons montrer que :

Propriété :

$$T = -\gamma_o^e \circ \psi^e \circ \frac{\partial}{\partial n}_i \circ \varphi^i$$

Vérification :

Il suffit de vérifier que :

$$\forall\ (u,v) \in V^2 \quad b[(\gamma_o^e \circ \psi^e \circ \frac{\partial}{\partial n_i} \circ \varphi^i)u, v] = -a(u,v)$$

comme $\varphi^e \circ \gamma_e^i \circ \psi^e = \psi^e$, on a :

$$b[(\gamma_o^e \circ \psi^e \circ \frac{\partial}{\partial n_i} \circ \varphi^i)u, v] = \int_{\Omega^e} \nabla(\psi^e \circ \frac{\partial}{\partial n_i} \circ \varphi^i)u \nabla \varphi^e v \, dx$$

$$= -\int_\Gamma (\frac{\partial}{\partial n_i} \circ \varphi^i)u \, v \, d\gamma$$

$$= -\int_{\Omega^i} \nabla \varphi^i(u) \nabla \varphi^i v \, dx = -a(u,v)$$

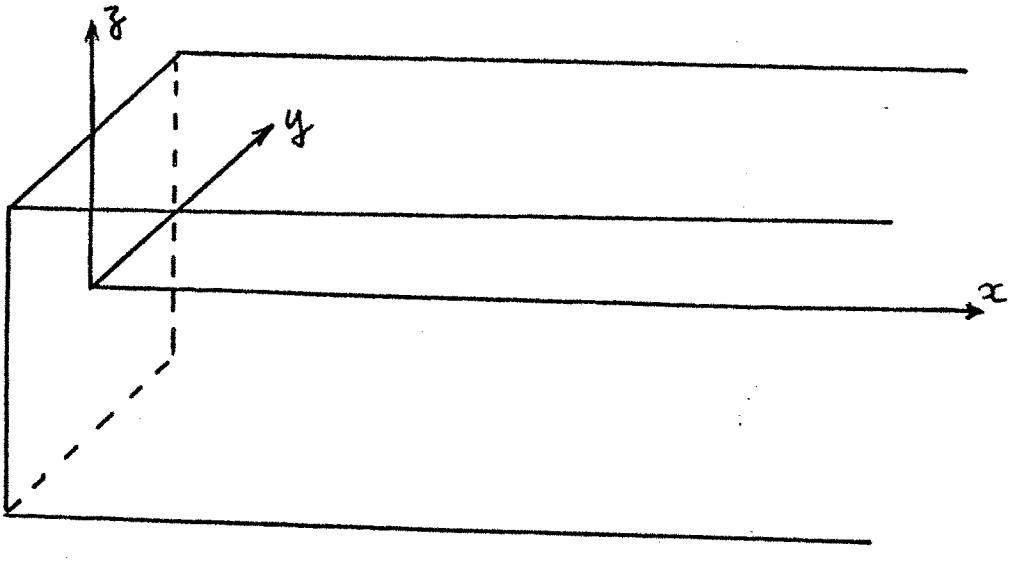
B2 Problème périodique

B21 Présentation

La situation décrite précédemment est évidemment assez éloignée des situations que l'on rencontre dans la pratique (présence d'un plan séparateur, et réseau infini de cavités près de ce plan). Une modélisation plus réaliste pourra être obtenue en introduisant des conditions de périodicité dans les deux directions déterminées par le réseau de cavités. On a représenté schématiquement un tel domaine, dans le cas où la périodicité serait de deux.

Nous allons maintenant nous intéresser à la résolution d'un problème posé dans un domaine parallélipipédique infini dans une direction

$\Omega =]0,+\infty[\times]-1,+1[\times]-1,+1[$



Soit donc le problème :

trouver U tel que :

$$\begin{cases} \Delta\ U = 0 \quad (\Omega) \\ \\ U(0,y,z) = u \quad \text{donnée} \end{cases}$$

avec les conditions suivantes :

$$\text{(I)} \qquad \begin{cases} U(x,-1,z) = U(x,1,z) \\ \\ U(x,y,-1) = U(x,y,+1) \end{cases}$$

$$\text{(II)} \qquad \begin{cases} U'_y(x,-1,z) = U'_y(x,1,z) \\ \\ U'_z(x,y,-1) = U'_z(x,y,1) \end{cases}$$

Ces conditions permettent de prolonger U en une fonction périodique ainsi que ses dérivées d'ordre 1 dans tout le demi-espace.

La fonction u donnée doit, bien sur, satisfaire aux conditions déduites de (I) pour $x = 0$.

Ce problème est intéressant dans notre problème pour deux raisons :

* Il permettrai de définir un cadre fonctionnel adapté à la résolution du problème périodique, c'est à dire un espace de Hilbert, dont les fonctions satisfont aux conditions (I) (les conditions (II) seraient alors des conditions naturelles), et tel que l'expression :

$$\int_\Omega \nabla\Psi\nabla\phi \, dx$$

soit continue et coercive sur un sous-espace de fonctions nulles pour x=0. Nous n'avons pas abordé l'étude de la définition de ces espaces.

* La résolution analytique de ce problème permet de trouver les "bornes" conditions limites pour un domaine tronqué :

Les solutions "générales" satisfaisant aux conditions (I) et (II), et telles que :

$$\lim_{x \to +\infty} u(x,y,z) = 0$$

peuvent s'écrire sous la forme :

$$U(x,y,z) = \sum_{m,n=0}^{\infty} e^{-\pi \sqrt{n^2+m^2}\, x} \left[A_1^{m,n} \sin ny \sin mz + A_2^{m,n} \sin ny \cos mz + A_3^{m,n} \cos ny \sin mz + A_4^{m,n} \cos ny \cos mz\right]$$

en posant :

$$T_1^{m,n}(y,z) = \sin(ny) \sin(mz) \qquad T_2^{m,n}(y,z) = \sin(ny) \cos(mz)$$

$$T_3^{m,n}(y,z) = \cos(ny) \sin(mz) \qquad T_4^{m,n}(y,z) = \cos(ny) \cos(mz)$$

On cherchera un développement de la donnée $u(y,z)$ en série :

$$u(y,z) = \sum_{m,n} \sum_{i=1}^{4} A_i^{m,n} T_i^{m,n} (y,z)$$

par l'utilisation des relations :

$$A_i^{m,n} = \int_{-1}^{1} \int u(y,z) \; T_i^{m,n} (y,z) \; dy \; dz$$

La solution U est alors connue explicitement sur tout le domaine, et la dérivée normale au plan $x=0$ est donnée par :

$$\frac{\partial U}{\partial n} = \sum_{m,n=0}^{\infty} \pi \sqrt{n^2+m^2} \sum_{i=1}^{4} A_i^{m,n} T_i^{m,n} (y,z)$$

(en orientant, par convention, la normale vers l'extérieur du domaine). On a donc une expression explicite d'un opérateur B, qui à une donnée u sur le carré, associe la dérivée normale du potentiel solution dans Ω du problème :

$$\begin{cases} \Delta U = 0 \\ U_{|x=0} = u \end{cases}$$

Il en résulte que nous pouvons découper le domaine initial, en trois parties :

- la partie au-dessus du plan séparateur métal/vide (P)
- une "tranche" finie, contenant les cavités, et sur laquelle on devra faire une résolution numérique. Cette tranche est limitée au-dessus par le plan séparateur, et au-dessous par un plan fixé arbitrairement ; soit Σ ce plan.
- une partie (infinie) au-dessous de Σ .

B22 Résolution d'un problème dans le Matériau

On désignera par :

Γ l'ensemble des surfaces de séparation, en distinguant

Γ_1 partie de Γ sur le plan séparateur P

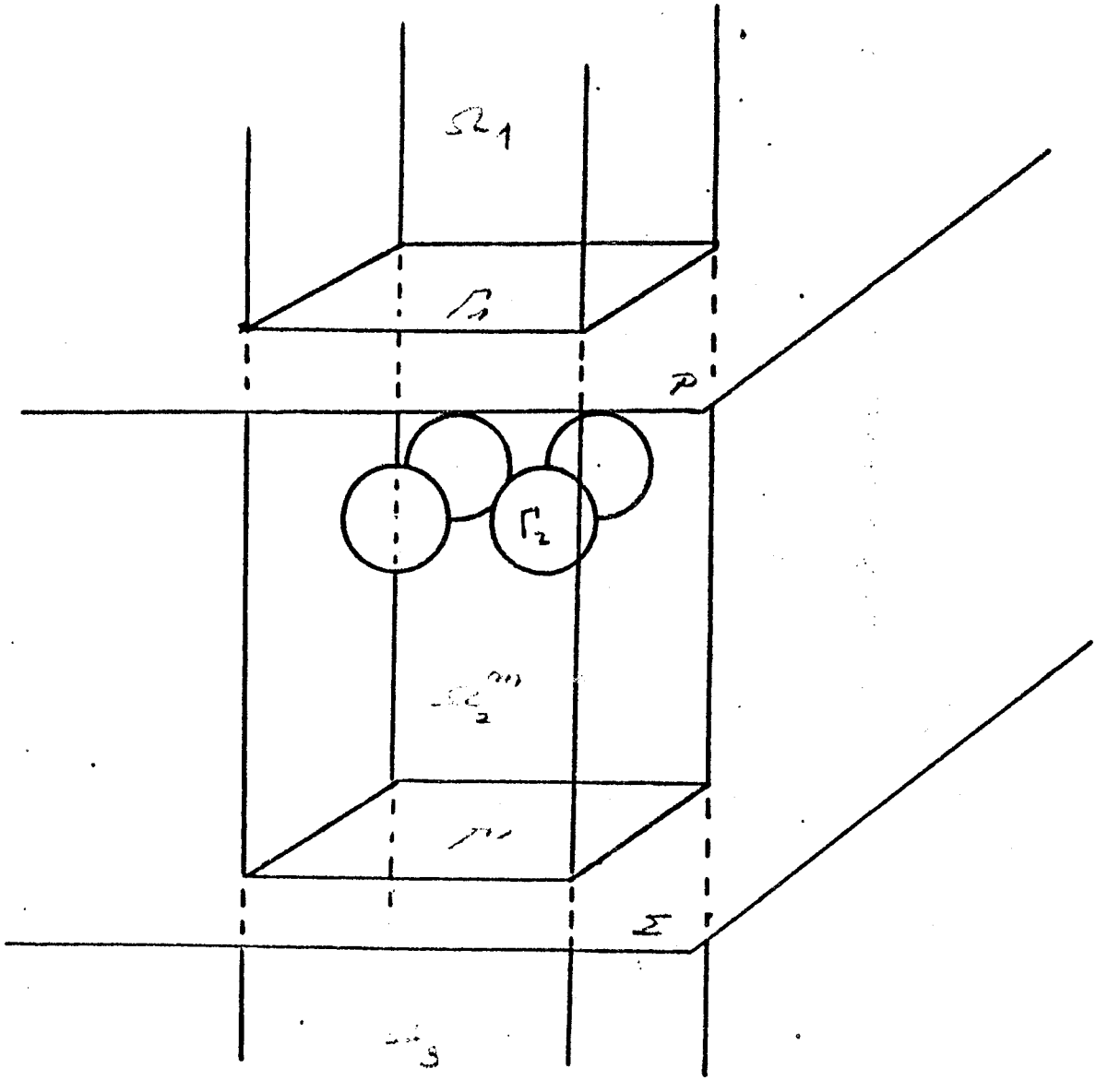
Γ_2 partie de Γ relative aux cavités

Γ_y^- partie du plan $y = -1$ servant de frontière au domaine

Γ_y^+ , Γ_z^- , Γ_z^+ , définis de manière analogue

Γ' partie de Σ commune avec le domaine,

enfin Ω_2^m désignera l'ouvert du domaine constitué par la partie extérieure aux cavités dans la tranche



On considérera alors X , sous-espace (Hilbertien) des fonctions de $H^1(\Omega_2^m)$ qui vérifient les conditions (I) et $\mathring{X}$ le sous-espace de X des fonctions "nulles" sur Γ ($= \Gamma_1 \cup \Gamma_2$)

L'espace V sera alors la "trace" sur Γ des fonctions de X . Si u est donné dans V , le problème du calcul de φ sera "résolu" de la manière suivante :

$R_o(u)$ sera choisie dans X , telle que $R_o(u)|_\Gamma = u$
on cherchera alors Ψ dans $\mathring{X}$ telle que :

$$\forall\, \omega \in \mathring{X}\ ,\ \int_{\Omega_2^m} \nabla\Psi\ \nabla\omega\ dx + \int_{\Gamma'} (B \circ \gamma')\Psi .\ \omega d\gamma = -\int_{\Omega_2^m} \nabla R_o(u)\ \nabla\omega\ dx$$

$$(B21) \qquad - \int_{\Gamma'} (B \circ \gamma')(R_o u)\ .\ \omega\ d\gamma$$

où γ' désigne l'application trace de X dans Γ' et B l'opérateur décrit précédemment.

Remarques :

- Nous ne vérifions pas la coercivité de la forme bilinéaire définie par le terme de gauche sur l'espace $\mathring{X}$.
- Cependant, cette forme n'est autre (dans des espaces appropriés, c'est-à-dire avec les bonnes conditions lorsque $x \to \infty$) que l'intégrale de $\nabla\Psi\nabla\omega$ étendue au domaine $\Omega_2^m \cup \Omega_3$
- $B(u)$ sera approché dans la résolution numérique en ne conservant que les premiers termes du développement de u :

$$u = \sum_{n,n} \sum_{i=1,4} A_i^{m,n}\ \gamma^i$$

Vérifions que cette formulation donne bien les solutions cherchées :

$$\varphi(u) = \Psi + R_o(u)$$

$$\forall\, \omega \in \mathring{X}\ ,\ \int_{\Gamma'} (B \circ \gamma')(\Psi + R_o u)\ \omega\, d\gamma + \int_{\Gamma' \cup \Gamma_{yz}^{\pm}} \frac{\partial}{\partial n}(\Psi + R_o(u))\ .\ \omega\, d\gamma =$$

$$= \int_{\Omega_2^m} \Delta(\Psi + R_o(u)\ .\ \omega\, dx$$

(nous supposons la solution Ψ assez régulière pour pouvoir utiliser la formule de Green).

Donc $\quad \Delta[\Psi + R_o(u)] = 0 \Rightarrow \Delta\, \varphi(u) = 0$

D'autre part, comme ω vérifie les conditions (I), nous déduisons de l'intégrale sur $\Gamma_{yz}^{\pm} = \Gamma_y^- \cup \Gamma_y^+ \cup \Gamma_z^- \cup \Gamma_z^+$ que la solution $\varphi(u)$ vérifie les conditions (II).

Il reste $\quad \frac{\partial}{\partial n}\Big|_{\Gamma'} [\Psi + R_o(u)] + (B \circ \gamma')(\Psi + R_o u) = 0$

ce qui nous assure de la continuité de la dérivée normale de $\varphi(u)$ à la traversée de Γ'

B23 Expression des formes bilinéaires $a(u,v)$ et $b(u,v)$

Ω^1 désignera le domaine au-dessus du plan P
Ω_2^v désignera le domaine intérieur aux cavités
r_1 la restriction d'une fonction de V à Γ_1

alors $\quad a(u,v) = \int_{\Omega_2^v} \nabla\varphi(u)\ \nabla\varphi(v)\ dx - \int_{\Gamma_1} (B \circ r_1)\ u.v d\gamma$

$$b(u,v) = \int_{\Omega_2^m} \nabla\varphi(u)\ \nabla\varphi(v)\ dx + \int_{\Gamma'} (B \circ \gamma')\ \varphi(u)\ .\ \varphi(v)\ d\gamma\ .$$

Remarques :

La construction d'un sous-espace satisfaisant aux conditions (I) dans Ω_2^m impose que l'on prenne la précaution d'avoir des maillages symétriques deux à deux sur les frontières suivantes :

Γ_y^- et Γ_y^+ d'une part

Γ_z^- et Γ_z^+ d'autre part

(C) ETUDE DE L'APPROXIMATION NUMERIQUE

On fera, pour cette étude, un certain nombre de simplifications techniques :
tout d'abord, on supposera que nous étudions des problèmes posés dans des domaines finis, par exemple en imposant un potentiel nul sur une surface "suffisamment éloignée" du domaine d'intérêt.
Désignons par Γ' cette surface :

alors : $X_m = \{u \in H^1(\Omega_m) \ / \ u|_{\Gamma'_m} = 0\}$

$X_v = \{u \in H^1(\Omega_v) \ / \ u|_{\Gamma'_v} = 0\}$

(X_m et X_v seront munis des normes induites).

Une seconde simplification consiste à supposer que Γ et Γ' sont d'un type particulier, permettant d'avoir :

$$\Gamma_h = \Gamma \qquad\qquad \Gamma'_h = \Gamma'$$

ce qui nous autorisera à avoir :

$$V_h \subset V \qquad\qquad X_{m_h} \subset X_m \qquad X_{v_h} \subset X_v$$

Rappelons que V est un espace de Hilbert, muni de la norme :

$$\|u\| = b(u,u)^{1/2}$$

alors, la continuité de la forme $a(u,v)$ nous permet de considérer l'opérateur T défini par :

$$\forall\, v \in V \qquad\qquad b(Tu,v) = a(u,v)$$

T est continu, autoadjoint. Ses éléments propres sont les solutions du problème suivant :

$$a(u,v) = \lambda\, b(u,v) \qquad \forall\, v \in V$$

En ce qui concerne le problème approché, l'introduction de l'hypothèse suivante :

H_o : $\|u_h\|_h = [b_h(u_h,u_h)]^{1/2}$ est une norme sur V_h

nous permet de définir T_h : $V_h \to V_h$ par :

$$b_h(T_h\, u_h, v_h) = a_h(u_h,v_h) \qquad \forall\ v_h \in V_h$$

Le problème posé est alors de montrer la convergence des éléments propres de T_h vers ceux de T . Deux points de vue sont alors possibles :

* on muni V_h de la norme induite par b , en utilisant l'inclusion $V_h \subset V$.

Nous perdons alors le caractère autoadjoint de T_h :

$$b(T_h u_h, V_h) \neq b(u_h, T_h v_h)$$

* on utilise sur V_h la norme $\|u_h\|_{b_h}$ et T_h est alors autoadjoint.

On se place donc dans le cadre de l'approximation discrète.

On adoptera ici le second point de vue, sachant que son utilisation sera nécessaire par la suite, si on veut, par exemple, prendre en compte le fait que $\Gamma_h \neq \Gamma$.

Nous allons dans les pages qui vont suivre étudier le problème de la convergence des valeurs propres. On rappellera d'abord un résultat classique, valable dans le cas d'opérateurs autoadjoints :

C1 Un résultat sur l'approximation des valeurs propres d'opérateurs linéaires autoadjoint

C11 Notations définition

Soit V , un espace de Hilbert réel, muni de la norme $\| \ \|$ et du produit scalaire $(\ , \ )$.

Soit V_h $(h \in]0\ 1[)$ une famille d'espace de Hilbert réels, avec $\| \ \|_h$ et $(\ , \ )_h$ pour norme et produit scalaire.

On suppose qu'il existe une famille d'applications linéaies :

$$p_h \ : \ V \to V_h \ ,$$

continues $\forall\ h$: $\exists\ C_h$ constante positive avec :

$$\forall\ x \in V \ , \quad \|p_h\ x\|_h \leq C_h \|x\|$$

et vérifiant l'hypothèse de compatibilité suivante :

(H) $$\forall\ x \in V \ , \quad \lim_{h \to 0} \|p_h x\|_h = \|x\|$$

C12 Rappel :

Théorème de la borne uniforme (6) soient φ_h des applications de B , espace de Banach dans $\mathbb{R}^+$

continues pour tout h dans l'intervalle $I \subset \mathbb{R}^+$ et telles que :

$$\forall\, h \in I\ ,\ \forall\ (u,v) \in B^2 \qquad \begin{cases} \varphi_h(u+v) \leq \varphi_h(u) + \varphi_h(v) \\ \varphi_h(-u) = \varphi_h(u) \end{cases}$$

$\forall\, u$, $\exists\, K_u$, constante réelle positive, telle que

$$\forall\, h \in I \qquad \varphi_h(u) \leq K_u$$

alors , il existe une constante M , telle que :

$$\forall\, u \in B\ ,\ \|u\| \leq 1\ ,\ \forall\, h \in I$$

$$\varphi_h(u) \leq M\ .$$

Une conséquence immédiate de ce théorème est que les applications p_h considérées précédemment sont uniformément bornées.

L'existence des p_h permet de donner un sens à l'approximation d'éléments de V par des éléments de V_h , la difficulté provenant du fait que, en général, $V_h \not\subset V$.

De nombreux auteurs ont introduit dans ce but, la notion de convergence discrète : [2] , [6] .

C13 Définitions :

a) On dira que la famille de vecteurs $x_h \in V_h$, $h \in]0,1]$, tend vers x , au sens de l'approximation discrète, lorsque le paramètre h tend vers zéro ssi :

$$\lim_{h \to 0} \|x_h - p_h(x)\|_h = 0$$

et on notera par la suite : $x_h \to\to x$.

b) Convergence discrète d'opérateurs :
Soit T un opérateur linéaire de V dans lui-même, et T_h une famille d'opérateurs de V_h dans V_h .
De même que précédemment, on dira que T_h converge ponctuellement (fortement) vers T , au sens de l'approximation discrète si, pour toute suite x_h telle que :

$$x_h \to\to x$$

alors $$T_h x_h \to\to T x$$

On notera cette propriété de la manière suivante :

$$T_h \to\to T \quad .$$

Remarque :

$T_h \to\to T$ entraine que pour tout $x \in X$:

$$\lim_{h \to 0} \|(p_h T - T_h p_h)x\|_h = 0$$

C15 Théorème classique sur la convergence des valeurs propres [ ], [ ].

Si T et T_h sont des opérateurs fermés et autoadjoints, soit :

$\forall\, x\,,\ \forall\, y \in V\ ,\quad (Tx,y) = (x,Ty)$

$\forall\, h\,,\ \forall\, x_h\,,\ \forall\, y_h \in V_h\ ,\quad (T_h x_h, y_h)_h = (x_h, T_h y_h)_h$

et si, de plus, on a la propriété de convergence suivante :

$$\forall\, x\,,\qquad \|T_h p_h x - p_h T x\|_h \to 0$$

Alors, toutes les valeurs propres isolées de T, de multiplicité finie, sont approchées par des valeurs propres de T_h.
Plus précisément, pour tout λ_o, valeur propre isolée de T, de multiplicité finie, et $\forall\, \varepsilon > 0$.
$\exists\ H(\varepsilon) > 0$ tel que $\forall\, h < H(\varepsilon)$, $\exists\ \lambda_h \in \sigma(T_h)$

avec $$|\lambda_h - \lambda_o| \leq \varepsilon$$

Démonstration :

A) On montrera d'abord la proposition suivante :

$\forall\, \varepsilon > 0$, $\exists\ h \in I$ et $\lambda_h \in \sigma(T_h)$ avec $\lambda_h \in D(\lambda_o, \varepsilon)$

Pour cela, on supposera, en raisonnant par l'absurde que λ_o, valeur propre isolée, n'est pas approchée. Il est alors possible de choisir $\eta > 0$ tel que le disque fermé : $D(\lambda_o, \eta) \subset \mathbb{C}$, de centre λ_o et de rayon η, soit tel que :

$D \setminus \{\lambda_o\} \subset \rho(T)$, ensemble résolvant de T

$D \subset \rho(T_h)$ pour tout $h \in I$.

Il est alors possible de borner uniformément $R(z)$ et $R_h(z)$ uniformément sur le contour $\partial D'[\lambda_o, \eta/2]$ en effet, T_h autoadjoint montre que :

$$\forall\, z \in (T_h) \quad , \quad \|R_h(z)\|_h = \frac{1}{\text{dist}[z, \sigma(T_h)]}$$

donc

$$\forall\, h \in I \quad , \quad \forall\, z \in \partial D' \quad , \quad \|R_h(z)\|_h \leq \frac{2}{\eta}$$

et, de même $\|R(z)\| \leq \frac{2}{\eta}$.

On va montrer que, pour tout $x \in X$, on a la proposition :

$$\forall\, \varepsilon > 0 \quad , \quad \exists\, H \in I \text{ tel que } \forall\, h \in I \quad h \leq H$$

$$\sup_{z \in \partial D'} \|(R_h(z) p_h - p_h R(z))\, x\|_h \leq \varepsilon$$

soit z_i un point de la frontière $\partial D'$ du disque D' il est possible de trouver $\eta_{i1} > 0$ indépendant de h tel que :

$$|z - z_i| < \eta_{i1} \Rightarrow \|p_h [R(z) - R(z_i)] x\|_h \leq \frac{\varepsilon}{3}$$

en effet, R est analytique dans un voisinage de $\partial D'$ et les p_h sont uniformément bornés.

On prendra d'autre part η_{i2} tel que :

$$|z - z_i| < \eta_{i2} \Rightarrow \|[R_h(z_i) - R_h(z)]\, p_h\, x\|_h \leq \frac{\varepsilon}{3}$$

Cela est possible puisque :

$$\|[R_h(z) - R_h(z_i)]\, p_h\, x\|_h \leq |z - z_i|\, \|R_h(z)\|_h \cdot \|R_h(z_i)\|_h \cdot \|p_h\, x\|_h$$

et que $\|R_h(z_i)\|_h \leq \frac{2}{\eta}$; $\|R_h(z)\|_h \leq \frac{1}{\frac{\eta}{2} - \eta_i}$

$$\|p_h x\|_h \leq C(x)$$

il suffit pour cela de choisir :

$$\eta_{i2} = \text{Min } [\frac{\eta}{4} ; \frac{\varepsilon}{24} \frac{\eta^2}{C(x)}]$$

On associera à chaque point z_i le disque ouvert suivant :

$$D(z_i, \eta_i) \quad \text{avec} \quad \eta_i = \text{Min } (\eta_{i1}, \eta_{i2})$$

Nous avons alors un recouvrement ouvert de ∂D , dont on extraira un recouvrement fini :

$$D(z_i, \eta_i) \quad , \quad i=1,\ldots,M$$

il suffit alors de choisir H tel que :

$$\begin{cases} h < H \\ \forall\ i = 1,\ldots,M \end{cases} \qquad \|(R_h(z_i)p_h - p_h R(z_i))x\|_h \leq \frac{\varepsilon}{3}$$

ce qui est possible :

$$\|(R_h(z_i)p_h - p_h R(z_i)x\|_h = \|R_h(z_i)(p_h T - T_h p_h)R(z_i)x\|_h$$

$$\leq \frac{2}{\eta} \|(p_h T - T_h p_h)R(z_i)x\|_h$$

et, en appliquant l'hypothèse de convergence ponctuelle à chaque vecteur $R(z_i)x$, $i=1,\ldots,M$, on a le rang H cherché.

alors : $\|(p_h R(z) - R_h(z) p_h)x\|_h \leq \|p_h [R(z) - R(z_i)]x\|_h +$

$$+ \|(p_h R(z_i) - R_h(z_i) p_h)x\|_h + \|(R_h(z_i) - R_h(z)) p_h x\|_h \leq \varepsilon$$

* Soit alors $x \neq 0$ avec $x \in P(X)$ où P est un projecteur sur le sous-espace invariant associé à λ_o :

$$P(x) = \frac{-1}{2i\pi} \int_{\partial D'} R(z)\ x\ dz = x$$

On suppose $\dim (\mathrm{Im}\ P) < +\infty$ (λ_o de multiplicité finie)
De même, on considérera :

$$P_h : V_h \to V_h \qquad P_h x_h = \frac{-1}{2i\pi} \int_{\partial D'} R_h(z)\ x_h\ dz$$

on a : $P_h(x_h) = 0$, $\forall\, x_h \in X_h$, puisque $R_h(z)$ est analytique dans D'

$$\text{mais } \|(P_h p_h - p_h P)x\|_h = \left\|\frac{1}{2i\pi} \int_{\partial D'} (R_h(z) p_h - p_h R(z))x\ dz\right\|_h$$

$$\leq \frac{1}{2i\pi} \text{ mes } |\partial D'| \ .\ \sup_{z \in \partial D'} \|(R_h(z) p_h - p_h R(z))x\|_h$$

le résultat précédent montre que :

$$\|(P_h p_h - p_h P)x\|_h \to 0$$

d'où $$\|(p_h P)x\|_h \to 0$$

ce qui est incompatible avec l'hypothèse sur p_h :

$$\|p_h P x\|_h \to \|(Px\| = \|x\| \neq 0$$

Donc il existe toujours une valeur propre de T_h dans tout disque de centre λ_o , et de rayon $\eta > 0$, arbitrairement petit.

B) On raisonnera encore par l'absurde pour passer du résultat démontré en A) au théorème :

Supposons $\eta > 0$ donné, tel que :

$\forall\, H \in]0,1]$, $\exists\, h < H$ avec $D(\lambda_o,\eta) \subset P(T_h)$

On appliquera cette hypothèse à la suite de valeurs :

$$H = \frac{1}{k} \quad , \quad k=1,2,\ldots,N,\ldots$$

nous obtenons donc une suite de valeurs :

$$h_{(k)} \to 0$$

et une suite d'approximation $T_{h_{(k)}}$, qui vérifient toujours l'hypothèse du théorème.

Nous pouvons encore appliquer à $T_{h_{(k)}}$ le raisonnement de la partie A) , ce qui montre l'impossibilité.

C2 Vérification des hypothèses

C21 Hypothèse H_o :

Le domaine Ω_m est connexe ; on supposera que sa frontière comprend une partie de Γ' , et qu'il existe au moins un noeud de la triangulation de Ω_m dans cette partie.

Alors, si : $\int_{\Omega_m} |\nabla \varphi_h(u_h)|^2 \, dx = 0$

avec l'hypothèse minimale : $X_m^h(\Omega_m) \subset C(\Omega_m)$ vérifiée déjà avec l'utilisation de polynômes de Lagrange de degré 1

$$\varphi_h(u_h) = Cte$$

comme $\varphi_h(u_h)|_{\Gamma'} = 0$

on a : $\varphi_h(u_h)|_{\Gamma} = u_h = 0$

d'où : $\| u_h \|_h = b_h(u_h, u_h)^{1/2} = [\int_{\Omega_m} [\nabla \varphi_h(u_h)]^2 dx]^{1/2}$

est bien une norme sur V_h .

C22

* On définira $p_h : V \to V_h$ comme la projection orthogonale sur V_h :

$u \in V$, $p_h u$ est défini par : $\begin{cases} p_h u \in V_h \\ b(p_h u - u_h, \omega) = 0 \quad \forall \omega_h \in V_h \end{cases}$

Nous utilisons, bien sur l'inclusion $V_h \subset V$ qui ne peut avoir lieu que si $\Gamma_h = \Gamma$.

On a alors :

$$\|p_h u\| \leq \|u\|$$

d'autre part, puisque V_h est de dimension finie ,

$\exists\, C_h$ tel que $\forall\ v_h \in V_h$ $\qquad \|v_h\|_h \leq C_h \|v_h\|$

d'où $\qquad \|p_h u\|_h \leq C_h \|u\|$

ce qui règle le problème de la continuité.

C23

* Avant d'aller plus loin, il est nécessaire de dire quelques mots sur les approximations de problème de "Dirichlet".

Ω désignera indifféremment Ω_m ou Ω_v et X sera X_m ou X_v selon le cas (on rappelle qu'il s'agit d'un sous-espace de $H^1(\Omega)$, avec une condition du type $u_{|\Gamma'} = 0$).

Approximation de $\varphi(u)$ par $\varphi_h(p_h u)$

$\varphi(u)$ peut être théoriquement obtenue de la manière suivante :

On désignera par $\bar{\Gamma}$ l'union de Γ et Γ'_m (ou Γ'_v). Alors, V a été choisi tel que :

si $v \in V$, $\bar{v}$, obtenue en prolongeant v par zéro sur Γ' est dans $H^{1/2}(\bar{\Gamma})$. Par le théorème des traces, nous savons qu'il existe un "relèvement" R(u) , satisfaisant :

$$\begin{cases} R(u) \in X \\ \gamma_o(u) = \bar{u} \end{cases}$$

où γ_o désigne la trace sur $\bar{\Gamma}$.

Ce relèvement peut être choisi continu [3] :

$\exists\, R_o$ tel que $\|R_o(u)\|_1 \leq K\|\bar{u}\|_{1/2} \leq K_1\|u\|_V$.

On introduira le problème suivant :

trouver $\Psi \in H_o^1(\Omega)$ telle que :

$$\forall\, \omega \in H_o^1(\Omega) \qquad \int_\Omega \nabla\Psi\, \nabla\omega\, dx = -\int_\Omega \nabla R_o(u)\, \nabla\omega\, dx$$

qui a une solution unique. Alors :

$$\varphi(u) = \Psi + R_o(u)$$

Une approximation par éléments finis construira un sous-espace de X :

$$X_h \subset X$$

avec : si $x_h \in X_h$ $\qquad x_h|_{\Gamma'} = 0$ et $x_h|_\Gamma \in V_h$

On désignera par X_h^0 le sous-espace de X_h des fonctions nulles sur Γ , et par π_h : $X \to X_h$ la projection orthogonale, au sens du produit scalaire de $H^1(\Omega)$, sur X_h .

La solution par éléments finis peut être calculée de la manière suivante :
Ψ_h étant l'unique fonction de X_h^0 telleque :

$$\forall\, \omega_h \in X_h^\circ \qquad \int_\Omega \nabla\Psi\, \nabla\omega_h\, dx = -\int_\Omega \nabla\pi_h\, R_o p_h u\, \nabla\omega_h\, dx$$

alors :

$$\varphi_h\, p_h\, u = \Psi_h + \pi_h\, R_o\, p_h\, u$$

Dans la pratique, $(R_{oh} \circ p_h)\, u$, unique fonction de X_h , qui vaut $p_h\, u$ sur Γ et 0 sur les noeuds intérieurs, remplace $\pi_h \circ R_o \circ p_h\, u$. Ceci ne change pas φ_h .

Nous allons borner $\| \varphi(u) - \varphi_h u \|_1$:

$$\|\varphi_h p_h u - \varphi u\|_1 \leq \|\psi_h - \psi\|_1 + \|R_o u - \pi_h R_o p_h u\|_1$$

on introduira alors ψ_h^* , unique solution dans X_h^0 du problème :

$$\forall \ \omega_h \in X_h^0 \qquad \int_\Omega \nabla \psi_h^* \nabla \omega_h \, dx = - \int_\Omega \nabla R_o u \nabla \omega_h \, dx$$

alors :

$$\|\psi_h - \psi\|_1 \leq \|\psi_h - \psi_h^*\|_1 + \|\psi_h^* - \psi\|_1$$

La coercivité du semi-produit scalaire sur $H_o^1(\Omega)$ montre que :

$$\|\psi - \psi_h^*\|_1 \leq C \inf_{\bar{v}_h \in X_h} |\psi - v_h|_1 \leq C \, |\psi - \pi_h \psi|_1$$

d'autre part :

$$\|\psi_h^* - \psi_h\|_1 \leq C \, |R_o u - \pi_h R_o p_h u|_1$$

$$\leq C \{ |(1-\pi_h) R_o u|_1 + |\pi_h R_o (1-p_h) u|_1 \}$$

avec : $|\pi_h R_o (1-p_h) u|_1 \leq \|R_o\| \, \|(1-p_h) u\|_{1/2}$

nous avons :

$$\boxed{\|\varphi(u) - \varphi_h p_h u\|_1 \leq C_1 \{ |\psi - \pi_h \psi|_1 + \|R_o u - \pi_h R_o u\|_1 + \|u - p_h u\|_{1/2} \}}$$

C24 Hypothèse de compatibilité

Soit à montrer que $\forall\, u \in V \qquad \|p_h\, u\|_h \to \|u\|$

$$\|p_h\, u\|_h^2 - \|u\|^2 = \int_\Omega \{|\nabla \varphi_h\, p_h\, u|^2 - |\nabla \varphi\, u|^2\}\, dx$$

$$\leq |\varphi_h\, p_h\, u - \varphi\, u|_{1,\Omega} \cdot |\varphi_h\, p_h\, u + \varphi\, u|_{1,\Omega}$$

en utilisant les résultats précédents, on va montrer que :

$$|\varphi_h\, p_h\, u - \varphi u|_1 \to 0$$

ce qui montrera que le second terme reste borné, du même coup

$$|\varphi_h\ p_h\, u - \varphi\, u|_1 \leq C_1\{|\Psi - \pi_h \Psi|_1 + \|R_o\, u - \pi_h\, R_o\, u\|_1 + \|u - p_h\, u\|_{1/2}\}$$

1) terme $\|u - p_h\, u\|_{1/2}$

avec $$\|u\|_{1/2} = \inf_{\substack{\Psi \in H^1(\Omega) \\ \gamma_o(\Psi) = \bar{u}}} \|\Psi\|_1,$$

qui est une norme équivalente à la norme définie par b :

$$\|u - p_h\, u\|_{1/2} \leq K\|u - p_h' u\|_b \leq K\|u - v_h\|_b \leq K'\|u - v_h\|_{1/2}$$

pour tout v_h dans V_h

mais : $$\|u - v_h\|_{1/2} = \inf_{\substack{\Psi \in H^1(\Omega) \\ \gamma_o \Psi = \overline{u - v_h}}} \|\Psi\|_1$$

en particulier, nous pouvons écrire cette majoration pour

$$v_h = \gamma_0\, \pi_h\, \varphi\, u|_\Gamma$$

(qui est bien dans V_h) , et pour $\Psi = \varphi\, u - \pi_h\, \varphi\, u$ d'où :

$$\|u - p_h u\|_{1/2} \leq \|\varphi\, u - \pi_h\, \varphi\, u\|_1$$

On est donc ramené à l'approximation d'un élément de $H^1(\Omega)$.

2) Termes du type :

$$\|\Psi - \pi_h \Psi\|_1$$

Plaçons-nous dans l'hypothèse minimale d'approximation avec des éléments du type Lagrange de degré 1.
On va montrer que :

$\forall\, \varepsilon > 0$, $\exists\, H(\varepsilon)$ tel que si $h < H$: $\|\Psi - \pi_h \Psi\|_1 \leq \varepsilon$

Comme $\mathcal{D}(\Omega)$ est dense dans $H^1(\Omega)$, nous pouvons trouver :

$\phi_k \in \mathcal{D}(\Omega)$ telle que $\|\Psi - \phi_k\|_1 \leq \frac{\varepsilon}{3}$

alors, de même : $\|\pi_h(\Psi - \phi_k)\|_1 \leq \frac{\varepsilon}{3}$

La théorie de l'interpolation par des éléments finis nous apprend que [ ]

$$\|\phi_k - \pi_h \phi_k\|_1 \leq C\, h |\phi_k|_{2,\Omega}$$

(avec une interpolation par des polynômes dans P1).
Il est donc facile de trouver H , tel que :

si $h < H$ alors $\|\phi_k - \pi_h \phi_k\|_1 \leq \frac{\varepsilon}{3}$

d'où le résultat.

Nous pouvons donc conclure que :

$$|\varphi_h\, p_h\, u - \varphi\, u|_1 \to 0$$

donc

$$\|p_h\, u\|_h \to \|u\| \qquad \forall\, u \in V\ .$$

C25 Hypothèse de convergence ponctuelle :

$$\|(p_h T - T_h p_h)u\|_h \to 0$$

Rappelons d'abord que de par les propriétés de p_h (et par l'utilisation du théorème de la borne uniforme), nous avons :

$\exists\, C > 0$ tel que $\forall\, h$, $\forall\, u$ $\quad \|p_h u\|_h \leq C\|u\| \leq C'\|u\|_{1/2}$

Il sera donc suffisant de démontrer que :

$$\|Tu - T_h p_h u\|_{1/2} \to 0$$

Cette démonstration se fera en plusieurs étapes :

① Définition et propriétés d'un opérateur $G : X_v \to X_m$

Soit f un élément de X_v . Alors le problème suivant :

Trouver $g \in X_m$ tel que $\forall\, \omega \in X_m$, on ait :

$$\int_{\Omega_m} \nabla g . \nabla \omega \, dx = \int_{\Omega_v} \nabla f \nabla [(R_{ov} \cdot \gamma_o^m)\omega] \, dx$$

où $\gamma_o^m : X_m \to V$ désigne l'application "trace" sur Γ et R_{ov} un relèvement continu, de V dans X_v .

Ce problème admet une solution unique, puisque :

- la forme : $\int_{\Omega_m} \nabla u \nabla v \, dx$ est coercive sur $X_m = \{u \in H^1(\Omega_m) \;/\; u_{|\Gamma'} = 0\}$

- l'application : $\omega \to \int_{\Omega v} \nabla f \nabla [(R_{ov} \cdot \gamma_m)\omega] \, dx$

est continue (pour f donnée) de X_m dans $\mathbb{R}$.

Soit donc G l'application définie par la résolution de ce problème : $G(f) = g$. Cette application est continue :

$$\|G(u)\|^2_{1,\Omega_m} \leq K \int_{\Omega_m} |\nabla G(u)|^2 dx \leq K \int_{\Omega_v} \nabla u \nabla [(R_{ov} \circ \gamma_o^m \quad G) u] dx$$

$$\leq K' |u|_{1,\Omega_v} \cdot \|Gu\|_{1,\Omega_m}$$

D'autre part, l'utilisation d'une formule de Green dans Ω_m et Ω_v mo n tre que :

$\forall\, u \in V$ $\qquad \Delta(G \circ \varphi_v) u = 0 \quad$ dans Ω_m

$$\frac{\partial}{\partial n_m} (G \circ \varphi_v) u = - \frac{\partial}{\partial n_v} (\varphi_v (u)$$

Il est alors facile de vérifier que :

$$\boxed{T = \gamma_o^m \circ G \circ \varphi_v}$$

$$b(\gamma_m \circ G \circ \varphi_v \, u, v) = \int_{\Omega_m} \nabla \, (G \circ \varphi_v) u \, \nabla \varphi_o^m (v) \, dx$$

$$= \int_{\Omega_v} \nabla \, \varphi_v (u) \, \nabla R_{ov} (v) \, dx = \int_{\Omega_v} \nabla \, \varphi_v (u) \, \nabla \varphi_v (v) dx$$

$$= a(u,v)$$

② Nous allons définir de même : $G_h : X_v \to X_m^h$ par le problème :

Trouver $G_h(f) \in X_m^h$ tel que $\forall\ \omega_h \in X_m^h$:

$$\int_{\Omega_m} \nabla G_h (f) \nabla \omega_h \, dx = \int_{\Omega_v} \nabla \, f \nabla \, (\varphi_{h_v} \circ \gamma_o^m) \, \omega_h \, dx$$

φ_{h_v} vérifie les propriétés :

$\exists H$, $\exists C$ tel que $\forall h, < H$, $|\varphi_{h_v} p_h u|_{1,\Omega} \leq C \|u\|_{1/2}$

$(\forall u)$ (théorème de la borne uniforme)

$$\gamma_v \circ \varphi_{h_v} v_h = v_h \qquad \forall v_h \in V_h$$

La première propriété assure de l'existence et unicité de $G_h(f)$. De plus, les G_h sont uniformément bornées en h :

$\exists K$ tel que $\forall h$, $\forall U_h \in X_v$, $\|G_h U_h\|_{1,\Omega_m} \leq K\|U_h\|_{1,\Omega_v}$

enfin, de même que précédemment, T_h admet la décomposition :

$$\boxed{T_h = \gamma_o^m \circ G_h \circ \varphi_{h_v}}$$

Remarquons d'abord que pour tout f_h dans X_v^h et v_h dans V_h

$$\int_{\Omega_m} \nabla(\varphi_{h_m} \circ \gamma_o^m \circ G_h) f_h \cdot \nabla \varphi_{h_m} v_h \, dx = \int_{\Omega_m} \nabla G_h \, f_h \nabla \varphi_{h_m} \, v_h \, dx$$

puisque $(\varphi_{h_m} \circ \gamma_o^m - 1)\, G_h\, f_h$ est nulle sur Γ

donc :

$$b_h(\gamma_m \circ G_h \circ \varphi_{h_v} u_h, v_h) = \int_{\Omega_m} \nabla(G_h \circ \varphi_{h_v}) u_h \nabla \varphi_{h_m} \, v_h \, dx$$

$$= \int_{\Omega_v} \nabla \varphi_{h_v} \, u_h \nabla \varphi_{h_v} \, v_h \, dx = a_h(u_h, v_h)$$

③ La convergence ponctuelle est donc assurée si :

$$\|(G \circ \varphi_v - G_h \circ \varphi_{h_v} p_h)u\|_{1,\Omega_m} \to 0$$

$$\|(G\,\varphi_v - G_h\,\varphi_{h_v} p_h)u\|_1 \leq \|(G-G_h)\varphi_v u\|_1 + \|G_h(\varphi_v - \varphi_{h_v} p_h)u\|_1$$

En ce qui concerne le deuxième terme, les résultats précédents :

$$\|(\varphi_v - \varphi_{h_v} p_h)u\|_{1,\Omega_v} \to 0$$

et $\|G_h u\|_{1,\Omega_m} \leq C\|u\|_{1,\Omega_v}$

montrent que

$$\|G_h(\varphi_v - \varphi_{h_v} p_h)u\|_{1,\Omega_m} \to 0$$

Remarquons d'abord que : $\forall\, V_h \in X_m^h$

$$\|(G-G_h)f\|^2_{1,\Omega_m} \leq K\,[\int_{\Omega_m} \nabla(G-G_h)f \cdot [\nabla(Gf-V_h) + \nabla(V_h-G_h f)]]dx$$

nous allons voir que la deuxième contribution est nulle si $f = \varphi_v u$.

Notons $\varepsilon_h = \gamma_o^m [G_h f - V_h]$. Le deuxième terme s'écrit :

$$\int_{\Omega_m} \nabla(G-G_h)\varphi_v u \cdot \nabla(V_h-G_h f)\,dx = \int_{\Omega_v} \nabla\varphi_v(u)\nabla[\varphi_{h_v}\varepsilon_h - R_{ov}\varepsilon_h]dx$$

$$= \int_\Gamma \frac{\partial}{\partial n}\varphi_v(u) \cdot \gamma_o^v[(\varphi_{h_v}\varepsilon_h - R_{ov}\varepsilon_h]\,d\gamma = 0$$

donc $\|(G-G_h)\varphi_v u\|_{1,\Omega_m} \leq \inf_{V_h \in X_m^h} |G\,\varphi_v\,u - V_h|_{1,\Omega_m}$

$$\|(G-G_h)\,\varphi_v\,u\|_{1,\Omega_m} \to 0$$

Nous venons donc de montrer que :

$$\forall\, u \in V \qquad \|(p_h T - T_h p_h)u\|_h \to 0$$

Ce qui garanti la convergence des valeurs propres, au sens du théorème C.15.

Remarque : Ce résultat reste assez partiel, puisque, en particulier il n'assure ni la convergence des vecteurs propres, ni la conservation des ordres de multiplicité.

Des résultats plus fort que ceux du théorème C15, dans le cas d'opérateurs non compacts ont été donnés par Descloux Nassif et Rappaz [4]. Leurs résultats ne semblent pas adaptables facilement au cas de l'approximation discrète.

REFERENCES

[1] J.CEA, GRISVARD

Colloque d'Analyse Numérique - 1978

[2] F. CHATELIN

"Théorie de l'approximation des opérateurs linéaires"
Cours de DEA. Université de Grenoble 1976-77

[3] P.G. CIARLET

"Introduction à l'analyse numérique de la méthode des éléments finis"
Cours DEA - Université Paris 6 1976-77

[4] J. DESCLOUX, N. WASSIF, J. RAPPAZ

"On the spectral approximation"
RAIRO - Analyse Numérique - vol 12 n°2 1978 p. 97 à 119

[5] J. GIROIRE

"Formulation variationnelle par équations intégrales de problèmes aux limites extérieures".
Ecole polytechnique - Centre de Mathématiques Appliquées - rapport n°6.

[6] T. KATO

"Perturbation theory for linear operation"
Springer Verlag (1966).

[7] J.C. NEDELEC, J. PLANCHARD

"Une méthode variable d'éléments finis pour la résolution numérique d'un problème extérieur dans R^3"
RAIRO - Analyse Numérique (décembre 1973) p. 105-129

[8] A. RONVEAUX

Colloque d'Analyse Numérique 1978

[9] A. RONVEAUX

" Survey of the integral equation. Governing the surface Plasmon modes"
International Congress of Maths. Helsinki - Aout 1978

[10] A. RONVEAUX, A. MOUSSIAUX, A. LACAS

"Interpolation dans une chaine de sphères ou de cavités alignées"
Can. Jour. of Physique. Voll 55 n° 16 (1977) p. 1407-1422

Dernière page d'une thèse

VU

Grenoble, le 10 juin 1980

Le Président de la thèse

F. Mathelin

Vu, et permis d'imprimer,

Grenoble, le 26.06.80

Le Président de l'Université Scientifique et Médicale

L. Cau

G. CAU

UNIVERSITÉ SCIENTIFIQUE ET MÉDICALE DE GRENOBLE

Dr G. CAU