Étude numérique du procédé de Peaceman-Rachford pour la résolution de problèmes elliptiques

Chaker Joubran

N° d'ordre :

T H E S E

présentée à

LA FACULTE DES SCIENCES DE L'UNIVERSITE DE

GRENOBLE

pour obtenir

LE TITRE DE DOCTEUR DE TROISIEME CYCLE

Mathématiques Appliquées

---

par

Chaker JOUBRAN

ETUDE NUMERIQUE DU PROCEDE DE PEACEMAN-RACHFORD POUR LA RESOLUTION DE PROBLEMES ELLIPTIQUES

Thèse soutenue le 6 avril 1965 devant la Commission d'examen :

Monsieur J. KUNTZMANN Président

Messieurs N. GASTINEL, G. HACQUES Examinateurs

N° d'ordre :

T H E S E

présentée à

LA FACULTE DES SCIENCES DE L'UNIVERSITE DE GRENOBLE

pour obtenir

LE TITRE DE DOCTEUR DE TROISIEME CYCLE

Mathématiques Appliquées

---

par

Chaker JOUBRAN

ETUDE NUMERIQUE DU PROCEDE DE PEACEMAN-RACHFORD POUR LA RESOLUTION DE PROBLEMES ELLIPTIQUES

Thèse soutenue le 6 avril 1965 devant la Commission d'examen :

Monsieur	J. KUNTZMANN	Président
Messieurs	N. GASTINEL	Examinateurs
	G. HACQUES	

FACULTE DES SCIENCES

L I S T E D E S P R O F E S S E U R S

DOYENS HONORAIRES — M. FORTRAT P.

M. MORET L.

DOYEN — M. WEIL L.

PROFESSEURS TITULAIRES

MM.	NEEL L.	MAGNETISME ET PHYSIQUE DU SOLIDE
	DORIER A.	ZOOLOGIE
	HEILMANN R.	CHIMIE ORGANIQUE
	KRAVTCHENKO J.	MECANIQUE RATIONNELLE
	CHABAUTY C.	CALCUL DIFFERENTIEL ET INTEGRAL
	PARDE M.	POTAMOLOGIE
	BENOIT J.	RADIOELECTRICITE
	CHENE M.	CHIMIE PAPETIERE
	BESSON J.	ELECTROCHIMIE
	WEIL L.	THERMODYNAMIQUE
	FELICI N.	ELECTROSTATIQUE
	KUNTZMANN J.	MATHEMATIQUES APPLIQUEES
	BARBIER R.	GEOLOGIE APPLIQUEE
	SANTON L.	MECANIQUE DES FLUIDES
	OZENDA P.	BOTANIQUE
	FALLOT M.	PHYSIQUE INDUSTRIELLE
	GALVANI O.	MATHEMATIQUES
	MOUSSA A.	CHIMIE NUCLEAIRE
	TRAYNARD P.	CHIMIE
	SOUTIF M.	PHYSIQUE
	CRAYA A.	HYDRODYNAMIQUE
	REULOS R.	THEORIE DES CHAMPS
	AYANT Y.	PHYSIQUE APPROFONDIE
	GALISSOT F.	MATHEMATIQUES APPLIQUEES
Melle	LUTZ E.	MATHEMATIQUES
MM.	BLAMBERT M.	MATHEMATIQUES
	BOUCHEZ R.	PHYSIQUE NUCLEAIRE
	ILLIBOUTRY L.	GEOPHYSIQUE
	MICHEL R.	GEOLOGIE ET MINERALOGIE
	BONNIER E.	ELECTROCHIMIE
	DESSAUX	PHYSIQUE ANIMALE
	PILLET E.	ELECTROCHIMIE
	DEBELMAS J.	GEOLOGIE
	GERBER R.	MATHEMATIQUES
	PAUTHENET R.	ELECTROTECHNIQUE
	VAUQUOIS B.	MATHEMATIQUES APPLIQUEES
	SILBER R.	MECANIQUE DES FLUIDES
	MOUSSIEGT J.	ELECTRONIQUE
	BARBIER J.C.	PHYSIQUE
	KPSZUL J.L.	MATHEMATIQUES
	BUYLE-BODIN M.	ELECTRONIQUE

PROFESSEURS SANS CHAIRE

M.	LACASE A.	THERMODYNAMIQUE
Mme	KOFLER L.	BOTANIQUE
MM.	DREYFUS B.	THERMODYNAMIQUE
	VAILLANT F.	ZOOLOGIE ET HYDROBIOLOGIE
	GIRAUD P.	GEOLOGIE
	GIDON P.	GEOLOGIE ET MINERALOGIE
	ARNAUD P.	CHIMIE
	PERRET R.	SERVOMECANISMES
Mme	LUMER L.	MATHEMATIQUES
Mme	BARBIER M.J.	ELECTROCHIMIE
Mme	SOUTIF J.	PHYSIQUE
MM.	BRISSONNEAU P.	PHYSIQUE
	COHEN J.	ELECTROCHIMIE
	DEPASSEL R.	MECANIQUE
	GASTINEL N.	MATHEMATIQUES APPLIQUEES

PROFESSEURS ASSOCIES

MM.	LUMER G.	MATHEMATIQUES
	HIGUCHI	BIOSYNTHESE DE LA CELLULOSE
	WAGNER	BOTANIQUE

MAITRES DE CONFERENCES

MM.	ROBERT A.	CHIMIE PAPETIERE
	ANGLES D'AURIAC	MECANIQUE DES FLUIDES
	BIAREZ J.P.	MECANIQUE PHYSIQUE
	COUMES A.	ELECTRONIQUE
	DODU J.	MECANIQUE DES FLUIDES
	DUCROS PL	MINERALOGIE ET CRISTALLOGRAPHIE
	CLENAT P.	CHIMIE
	HACQUES G.	CALCUL NUMERIQUE
	LANCIA R.	PHYSIQUE AUTOMATIQUE
	PEBAY-PEROULA	PHYSIQUE
	KAHANE	PHYSIQUE GENERALE
	DOLIQUE	ELECTRONIQUE
Mme	KAHANE J.	PHYSIQUE
MM.	DEGRANGE C.	ZOOLOGIE
	GAGNAIRE D.	CHIMIE PAPETIERE
	RASSAT A.	CHIMIE SYSTEMATIQUE
	KLEIN J.	MATHEMATIQUES
	POULOUJADOFF M.	ELECTROTECHNIQUE
	DEPOMMIER P.	PHYSIQUE NUCLEAIRE
	DEPORTES C.	CHIMIE
	BARRA J.	MATHEMATIQUES APPLIQUEES
Mme	BOUCHE L.	MATHEMATIQUES
MM.	PERRIAUX J.	GEOLOGIE
	SARROT-REYNAULD	GEOLOGIE
	CAUQUIS G.	CHIMIE GENERALE
	LABRE A.	BOTANIQUE
	BONNET G.	PHYSIQUE GENERALE
	BARNOUD F.	BIOSYNTHESE DE LA CELLULOSE
Mme	BONNIER M.J.	CHIMIE
	CAUBET	MATHEMATIQUES APPLIQUEES
	BERTRANDIAS	MATHEMATIQUES APPLIQUEES

MAITRES DE CONFERENCES ASSOCIES

MM.	ISHIKAWA Y.	MAGNETISME
	QUATTROPANI	THERMODYNAMIQUE

Le présent travail a été entièrement fait, sous la direction bienveillante de Monsieur le Professeur GASTINEL, au Laboratoire de calcul de Mathématiques Appliquées de l'Université de Grenoble.

Je voudrais remercier tout particulièrement :

Monsieur le Professeur KUNTZMANN pour l'honneur qu'il m'a fait en acceptant de présiderle jury de thèse.

Monsieur le Professeur GASTINEL pour ses conseils judicieux et ses encouragements précieux.

Monsieur HACQUES, maître de Conférences et membre du jury pour l'accueil chaleureux qu'il m'a réservé à mon arrivée à Grenoble.

Messieurs N. SALHAB et H. MOUCHARRAFIEH pour l'aide matérielle qu'ils m'ont fournie en leur qualité de doyen et d'ex-doyen de la faculté des Sciences de l'Université libanaise.

Madame FRAIMBAULT et Mademoiselle MEUNIER pour la réalisation matérielle de cette thèse.

INTRODUCTION

-=-=-

Le présent travail consiste en l'étude numérique d'une méthode de résolution des systèmes linéaires et par suite de la résolution de certains problèmes elliptiques qui se ramènent par discrétisation à des systèmes linéaires : il s'agit de la méthode de Peaceman-Rachford.

Cette méthode nécessite un choix adéquat de paramètres r_i qui se ramène à un problème de min-max pour une fraction rationnelle qui, pour le cas présent, s'avère semblable au problème de Tchebycheff pour les polynômes. Nous avons calculé ces paramètres par deux méthodes : l'une générale (mais approximative) due à Wachpress, l'autre restreinte (mais exacte) au cas où le nombre de paramètres est de la forme 2^k. Nous avons, par ailleurs, étudié l'influence de ces paramètrès sur la rapidité de convergence de la méthode et nous avons comparé le temps de calcul de la présente méthode avec celui de la méthode dite "directe SH". De plus, nous avons déterminé le domaine dans lequel la méthode de Peaceman-Rachfor présente ses avantages. Enfin, nous avons étudié l'erreur de calcul due à la capacité limitée de la machine et donné une idée sur les plages d'erreurs globales.

Il est à notér que tous les résultats numériques figurent à la fin et forment la quatrième partie de cette étude.

I - PARTIE

-=-

METHODE DES DIRECTIONS ALTERNEES

-=-=-=-=-=-=-

I. - METHODE DES DIRECTIONS ALTERNEES

A. - Rappel de quelques définitions et théorèmes.

Soient une matrice carrée A d'ordre n, ayant pour valeurs propres λ_i, $1 \leqslant i \leqslant n$, et A* sa transposée conjuguée, soit X un vecteur appartenant à l'espace vectoriel C^n des colonnes de n nombres complexes dans l'espace à n dimensions avec :

$$X = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{vmatrix} \qquad X^T = \begin{vmatrix} x_1 & x_2 & \cdots & x_n \end{vmatrix}$$

$$X^* = \begin{vmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_n \end{vmatrix}$$

où les x_i sont en général des nombres complexes et les $\bar{x}_i$ sont les conjugués des x_i.

A. 1.

On appelle norme euclidienne de X la quantité

$$\| X \| = (X^*.X)^{1/2} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$

A.2.

On appelle rayon spectral de A la quantité :

$$\rho(A) \equiv \max_{1 \leqslant i \leqslant n} |\lambda_i|$$

A.3.

On appelle norme spectrale de A la quantité :

$$\| A \| = \sup_{X \neq 0} \left(\frac{\| AX \|}{\| X \|} \right)$$

A.4.

On dit que la matrice A est convergente (vers zéro) si la suite des matrices A, A^2, A^3, ..., A^t, ... converge vers la matrice nulle 0. ($t \to +\infty$)

A.5. Théorème

Soient X et Y deux vecteurs appartenant à l'espace C^n et soit α un scalaire quelconque, alors :

$$\| X \| > 0 \qquad \text{pour tout } X \in C^n \text{ sauf pour } X \equiv 0$$
$$\| \alpha X \| = |\alpha| . \| X \|$$
$$\| X+Y \| \leq \| X \| + \| Y \|$$

A.6. Théorème

Soient A et B deux matrices carrées d'ordre n, X un vecteur appartenant à C^n et α un scalaire, alors :

$$\| A \| > 0 \quad \text{sauf si } A \equiv 0$$
$$\| A+B \| \leq \| A \| + \| B \|$$
$$\| A.B \| \leq \| A \| . \| B \|$$
$$\| AX \| \leq \| A \| . \| X \|$$

et il existe un vecteur Y non nul appartenant à C^n et pour lequel on a :

$$\| AY \| = \| A \| . \| Y \|$$

A.6.1. Corollaire

Pour toute matrice A, on a :

$$\rho(A) \leq \| A \|$$

A.7. Théorème

Pour toute matrice A, on a :

$$\| A \| = \left(\rho(A^*A) \right)^{\frac{1}{2}}$$

En effet :

la matrice A* A est une matrice hermitienne, (une matrice B est dite hermitienne si B* = B), définie non-négative :

$$(A^*A)^* = A^* A$$

$$X^*A^*AX = \|AX\|^2 \geq 0 \qquad \forall X$$

Soit donc $\{\alpha_i\}_{i=1}^{n}$ l'ensemble orthonormal des vecteurs propres de la matrice A*A et soit $\{\nu_i\}_{i=1}^{n}$ l'ensemble des valeurs propres correspondant, alors :

$$A^* A\, \alpha_i = \nu_i\, \alpha_i \qquad \text{où } 0 \leq \nu_1 \leq \nu_2 \leq \ldots\ldots \leq \nu_n$$

et
$$\begin{cases} \alpha_i^*\, \alpha_j = 0 & \text{pour tout } i \neq j \\ \alpha_i^*\, \alpha_i = 1 & \text{pour tout } i \text{ tel que } 1 \leq i \leq n \end{cases}$$

Si $X = \sum_{i=1}^{n} C_i\, \alpha_i$, où les C_i sont des constantes, est un vecteur non nul, par un calcul direct nous obtenons :

$$\left(\frac{\|AX\|}{\|X\|}\right)^2 = \frac{X^*A^*AX}{X^*X} = \frac{\sum_{i=1}^{n} |C_i|^2\, \nu_i}{\sum_{j=1}^{n} |C_j|^2}$$

Alors, nous pourrons écrire :

$$0 \leq \nu_1 \leq \left(\frac{\|AX\|}{\|X\|}\right)^2 \leq \nu_n$$

De plus, en choisissant $X = \alpha_n$, on voit que l'égalité à droite est atteinte, alors nous aurons :

$$\|A\|^2 = \sup_{X \neq 0} \left(\frac{\|AX\|}{\|X\|}\right)^2 = \nu_n = \rho\,(A^*A)$$

et :

$$\|A\| = \left(\rho(A^*A)\right)^{1/2} \qquad \text{c.q.f.d.}$$

A.7.1. Corollaire

Si A est hermitienne alors :

$$\rho(A) = \|A\|$$

De plus si $P_m(z)$ est un polynôme réel de degré m en z alors :

$$\|P_m(A)\| = \rho(P_m(A))$$

En effet : si A est hermitienne on a $A^* = A$ et d'après le théorème (A.7.) on peut écrire :

$$\|A\|^2 = \rho(A^*A)$$
$$= \rho(A^2) = (\rho(A))^2$$

d'où :

$$\rho(A) = \|A\| .$$

le second point s'en déduit facilement.

A.8. Théorème.

La matrice A est convergente si et seulement si $\rho(A) < 1$.

En effet, il existe une matrice non singulière S telle que :

$$SAS^{-1} = \tilde{A}$$

$\tilde{A}$ étant de la forme de Jordan suivante

$$\tilde{A} = \begin{pmatrix} J_1 & & & & & 0 \\ & J_2 & & & & \\ & & J_3 & & & \\ & & & J_4 & & \\ & & & & J_5 & \\ & & & & & \ddots \\ 0 & & & & & & J_r \end{pmatrix}$$

où J_K est une matrice carrée d'ordre n_K ayant la forme suivante :

$$J_K = \begin{vmatrix} \lambda_K & 1 & & & \\ & \lambda_K & 1 & & 0 \\ & & \lambda_K & 1 & \\ & & & \ddots & \ddots \\ 0 & & & & \lambda_K \quad 1 \end{vmatrix}$$

où λ_K est une valeur propre de A et $n_1 + n_2 + \ldots + n_r = n$; n étant l'ordre de la matrice carrée A.

De l'écriture $\tilde{A} = SAS^{-1}$ on tire : $(\tilde{A})^m = SA^m S^{-1}$ d'où :

$$A^m = S^{-1} \tilde{A}^m S$$

de sorte qu'il est équivalent de chercher la condition nécéssaire et suffisante pour que $(\tilde{A})^m \longrightarrow 0$

mais $(\tilde{A})^m = \begin{vmatrix} J_1^m & & & \\ & J_2^m & & 0 \\ & & \ddots & \\ 0 & & & J_r^m \end{vmatrix}$

et par conséquent $A^m \xrightarrow[m\to\infty]{} o$ si et seulement si $J_K^m \xrightarrow[m\to\infty]{} o$ pour tout K.

Considérons donc J_K et calculons J_K^2 :

$$J_K^2 = J_K . J_K = \begin{vmatrix} \lambda_K & 1 & & & \\ & \lambda_K & 1 & & 0 \\ & & \ddots & \ddots & \\ 0 & & & \ddots & 1 \\ & & & & \lambda_K \end{vmatrix} \times \begin{vmatrix} \lambda_K & 1 & & & \\ & \lambda_K & 1 & & 0 \\ & & \ddots & \ddots & \\ 0 & & & \ddots & 1 \\ & & & & \lambda_K \end{vmatrix}$$

$$J^2_K = \begin{vmatrix} \lambda^2_K & 2\lambda_K & 1 & 0 & & \\ & \lambda^2_K & 2\lambda_K & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & 1 \\ 0 & & & & \ddots & 2\lambda_K \\ & & & & & \lambda^2_K \end{vmatrix}$$

et par récurrence, on démontre que $J^{(m)}_K$ a pour éléments les quantités $d^{(m)}_K(i,j)$ où

$$d^{(m)}_K(i,j) = \begin{cases} 0 & \text{si } j < i \\ C^{j-1}_m \lambda^{m-j+i}_K & \text{si } i \leqslant j \leqslant \max(n_K, m+i) \\ 0 & \text{si } m+i \leqslant j < n_K \end{cases}$$

Par conséquent pour que J^m_K converge vers zéro, il est nécessaire que $|\lambda_K| < 1$, on constate que cette condition est suffisante, donc $J^m_K \longrightarrow 0$ si et seulement si $|\lambda_K| < 1$. Donc $A^m \longrightarrow 0$ si et seulement si $|\lambda_K| < 1$ pour tout K. Donc finalement $A^m \longrightarrow 0$ si et seulement si $\rho(A) < 1$.

c.q.f.d.

A.9. Théorème d'Hadamard

Soit A une matrice carrée d'ordre n ayant pour éléments a_{ij} et soit D_i le disque du plan complexe :

$$D_i = \left\{ z;\ |z-a_{ii}| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\}$$

Alors les valeurs propres de A sont contenues dans $\bigcup_{i=1}^{n} D_i$.

En effet soit λ une valeur propre de A et soit X un vecteur propre correspondant, c'est-à-dire :

$$AX = \lambda X$$

X ayant pour composantes $x_1, \ldots, x_n$. Soit x_r la plus grande composante en valeur absolue, avec $x_r \neq 0$, alors :

$$(\lambda - a_{rr})\, x_r = \sum_{j \neq r} a_{rj}\, x_j$$

et comme $x_r \neq 0$ alors :

$$|\lambda - a_{rr}| \cdot |x_r| < \sum_{j \neq r} |a_{rj}| \cdot |x_j|$$

$$|\lambda - a_{rr}| \leq \sum_{j \neq r} |a_{rj}|$$

c.q.f.d.

A.9.1. Corollaire

$$\rho(A) \leq \max_i \left(\sum_{j=1}^{n} |a_{ij}| \right)$$

et

$$\rho(A) \leq \max_j \left(\sum_{i=1}^{n} |a_{ij}| \right)$$

A.9.2. Corollaire

Les modules des valeurs propres d'une matrice A se trouvent entre la plus grande des quantités $\sum_{j=1}^{n} |a_{ij}|$ pour tout i allant de 1 à n d'une part, et la plus petite des quantités $\left(|a_{ii}| - \sum_{\substack{j \neq i \\ j=1}}^{n} |a_{ij}| \right)$ pour tout i allant de 1 à n d'autre part.

B. - Principe de la méthode.

Considérons le problème de Dirichlet dans un carré R de côté π, ce problème est connu sous le nom de "problème modèle" ** : essayons de trouver une approximation de la fonction U (x,y) définie à l'intérieur de ce carré et satisfaisant à l'équation de Laplace suivante :

$$(1)\quad \Delta U = \frac{\partial^2}{\partial x^2} U(x,y) + \frac{\partial^2}{\partial y^2} U(x,y) = 0 \qquad 0 < x,y < \pi$$

** le "problème modèle" est connu aussi sous le nom de "problème type".

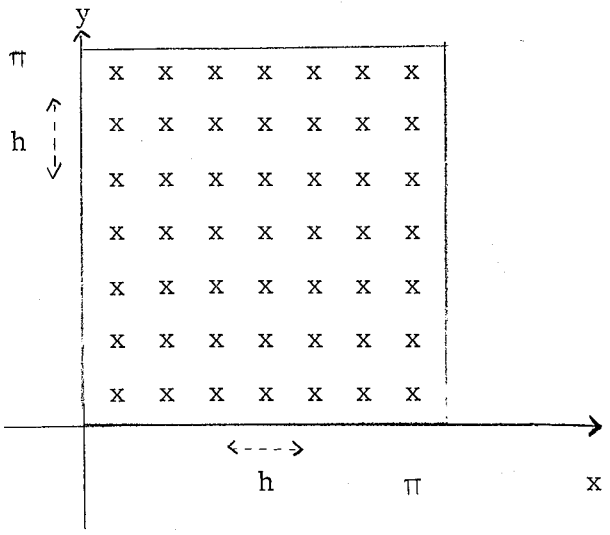
et telle que :

$$(2) \qquad U(x,y) = g(x,y) \text{ pour tout } (x,y) \in \Gamma$$

où Γ est la frontière du carré R et g (x,y) est une fonction connue, définie sur Γ .

Nous pratiquons dans R un quadrillage uniforme de pas h sur ox et de même sur oy, nous aurons ainsi n points suivant ox et n points suivant oy tels que

$$\pi = (n+1)\, h$$

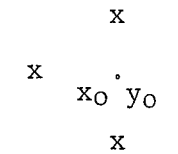


Alors en supposant h suffisamment petit pour pouvoir considérer h^p, pour $p > 3$, comme négligeable et en appliquant le développement de Taylor sur la fonction U (x,y) supposée suffisamment différentiable, nous aurons :

$$(3)\quad \begin{cases} U(x_o+h,y_o) = U(x_o,y_o) + h\frac{\partial}{\partial x}U(x_o,y_o) + \frac{h^2}{2}\frac{\partial^2}{\partial x^2}U(x_o,y_o) + \frac{h^3}{3!}\frac{\partial^3 U}{\partial x^3}(x_o,y_o) + \theta \\ U(x_o-h,y_o) = U(x_o,y_o) - h\frac{\partial}{\partial x}U(x_o,y_o) + \frac{h^2}{2}\frac{\partial^2}{\partial x^2}U(x_o,y_o) - \frac{h^3}{3!}\frac{\partial^3 U}{\partial x^3}(x_o,y_o) + \theta \\ U(x_o,y_o+h) = U(x_o,y_o) + h\frac{\partial}{\partial y}U(x_o,y_o) + \frac{h^2}{2}\frac{\partial^2}{\partial y^2}U(x_o,y_o) + \frac{h^3}{3!}\frac{\partial^3}{\partial x^3}(x_o,y_o) + \theta \\ U(x_o,y_o-h) = U(x_o,y_o) - h\frac{\partial}{\partial y}U(x_o,y_o) + \frac{h^2}{2}\frac{\partial^2}{\partial y^2}U(x_o,y_o) - \frac{h^3}{3!}\frac{\partial^3 U}{\partial x^3}(x_o,y_o) + \theta \end{cases}$$

où θ est une quantité en h^4 donc négligeable

où le point (x_o,y_o) appartient à l'ensemble des points intérieurs au carré et où les quatre points $(x_o \pm h, y_o)$ et $(x_o, y_o \pm h)$ peuvent appartenir ou bien à l'interieur de R ou bien à la fois à l'intérieur de R et à sa frontière Γ selon le point autour duquel on fait le développement.



En additionnant les quatre relations de (3) nous aurons :

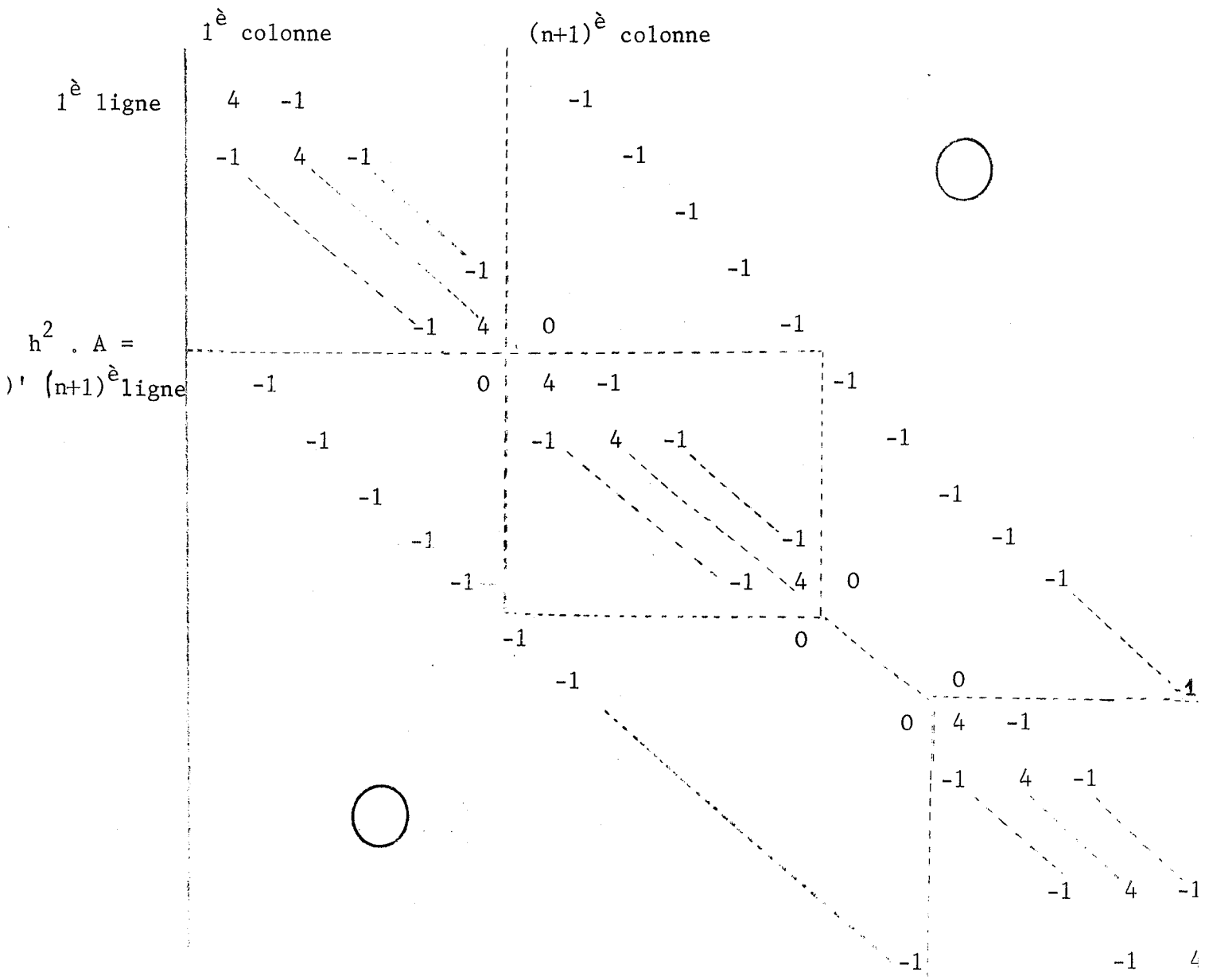
$$\frac{1}{h^2}\left\{ U(x_o+h,y_o) - 2U(x_o,y_o) + U(x_o-h,y_o) \right\}$$
$$+ \frac{1}{h^2}\left\{ U(x_o,y_o+h) - 2U(x_o,y_o) + U(x_o,y_o-h) \right\} =$$
$$\frac{\partial^2}{\partial x^2} U(x_o,y_o) + \frac{\partial^2}{\partial y^2} U(x_o,y_o) = 0 \qquad (4)$$

et en écrivant la même chose pour tous les points du quadrillage et en faisant passer les quantités $\frac{1}{h^2} U(x_o \pm h, y_o)$ et $\frac{1}{h^2} U(x_o, y_o \pm h)$ dans le second membre, chaque fois que les points $(x_o \pm h, y_o)$ et $(x_o, y_o \pm h)$ sont des points de Γ, donc la U correspondant connue par l'intermédiaire de $g(x,y)$, nous obtenons un système linéaire de la forme :

$$A X = K \qquad (5)$$

où A est une matrice carrée d'ordre N ($N = n^2$) et où X est un vecteur ayant N éléments qui ne sont autres que les quantités $U(ih, jh)$ i et j allant de 1 à n, K étant le vecteur second membre.

La matrice A du système (5) à la forme (5') suivante :



Une matrice creuse ayant une diagonale principale possédant des termes tous égaux à 4 ; les deux diagonales adjacentes à la diagonale principale ayant tous leurs éléments égaux à - 1 sauf les $(j.n)^{ième}$ éléments qui sont nuls pour j = 1, ..., n ; les deux diagonales qui restent ayant chacune n (n-1) éléments tous égaux à - 1.

La résolution du système (5), A ayant la forme (5'), par une méthode classique de résolution des systèmes linéaires (méthode de Gauss, ou variantes) nécessite la réservation d'un très grand nombre de mémoires, et comme le nombre des mémoires dans la machine est limitée, cette résolution est presque impossible.

Il est donc avantageux d'extraire de la matrice A deux autres matrices H et V telles

$$(6) \qquad A = H + V$$

H ayant la forme (22) et V la forme (23) données toutes les deux plus loin. En examinant les formes (22) et (23) nous nous apercevons que le système linéaire ayant H ou V comme matrice est très pratique pour être résolue vu que sa résolution n'exige pas plus que 2N mémoires, N étant l'ordre de la matrice A, car ce sont des matrices ayant 3 [illegible] à la diagonale seulement.

Ainsi le système (5) s'écrit alors :

$$(7) \qquad (H + V)\, X = K$$

Considérons un scalaire $r > 0$ et soit I la matrice unité ayant même ordre que A, alors (7) peut être remplacée par :

$$(8) \qquad (H + V + rI - rI)\, X = K \qquad\qquad r > 0$$

La relation (8) peut être remplacée par le couple suivant :

$$(9) \quad \begin{cases} (H + rI)\, X = (rI - V)\, X + K \\ (V + rI)\, X = (rI - H)\, X + K \end{cases}$$

chacune des équations (9) étant identique à (8)

Soit $X^{(o)}$ un vecteur approchant la solution du système (5), alors je calcule un premier vecteur itéré $X^{(1)}$ par l'intermédiaire d'un autre vecteur $X^{(1/2)}$ de la façon suivante :

$$(10) \quad \begin{cases} (H + rI)\, X^{(1/2)} = (rI - V)\, X^{(o)} + K \\ (V + rI)\, X^{(1)} = (rI - H)\, X^{(1/2)} + K \end{cases}$$

et ayant calculé un vecteur itéré $X^{(m)}$ le suivant $X^{(m+1)}$ sera calculé par :

$$(11) \quad \begin{cases} (H + rI)\, X^{(m+1/2)} = (rI - V)\, X^{(m)} + K \\ (V + rI)\, X^{(m+1)} = (rI - H)\, X^{(m+1/2)} + K \end{cases} \qquad \forall\ m \geq 0$$

En combinant les deux relations (11) il vient :

$$(12) \quad X^{(m+1)} = Tr\, X^{(m)} + Gr\,(K)$$

ou

$$(13) \quad Tr = (rI + V)^{-1}\,(rI - H)\,(rI + H)^{-1}\,(rI - V)$$

et

$$(14) \quad Gr = (rI + V)^{-1}\left\{(rI - H)\,(rI + H)^{-1} + I\right\} . K$$

Remarques et définitons

1. La méthode ainsi définie est dite méthode de directions alternées, car si on explicite H et V dans la résolution du "problème modèle" nous aurons :

$$(15)^{**}\begin{cases}\left[HU\right](x_o,y_o) = \dfrac{1}{h^2}\left\{U(x_o - h, y_o) - 2U(x_o,y_o) + U(x_o+h,y_o)\right\}\\ \left[VU\right](x_o,y_o) = \dfrac{1}{h^2}\left\{U(x_o, y_o - h) - 2U(x_o,y_o) + U(x_o,y_o+h)\right\}\end{cases}$$

l'écriture $\left[HU\right](x_o,y_o)$ désigne la composante relative au point (x_o,y_o) du vecteur HU

On constate que la première de (11) est résolue suivant les lignes horizontales du quadrillage, tandis que la deuxième, elle est résolue suivant les lignes verticales.

2. La matrice Tr est dite matrice de la méthode.
3. Le paramètre r est dit paramètre accélérateur.
4. Dans le calcul de $X^{(m+1)}$ à partir de $X^{(m)}$ le vecteur $X^{(m+1/2)}$ ne joue qu'un rôle auxiliaire.
5. Le passage de $X^{(m)}$ à $X^{(m+1/2)}$ nous lui donnons le nom de sous-itération, de même pour celui de $X^{(m+1/2)}$ à $X^{(m+1)}$, tandis que le passage de $X^{(m)}$ à $X^{(m+1)}$ nous lui donnons le nom d'itération
6. Pour que le calcul de $X^{(m+1)}$ à partir de $X^{(m)}$ soit rapide et facile à faire, il nous faut au moins réserver 2N mémoires plus 4 n mémoires pour les valeurs de la fonction sur la frontière et 2n mémoires pour stocker des valeurs intervenant dans la résolution des systèmes linéaires

** H et V ayant ici une signification symbolique.

C. - Convergence de la méthode.

C.1.

Soit X* le vecteur solution exacte du système linéaire AX = K, et soit $E^{(m}$ le vecteur erreur associé au vecteur itéré $X^{(m)}$ tel que :

$$(16) \quad E^{(m)} = X^{(m)} - X^*$$

compte tenu de (12), il vient :

$$(17) \quad E^{(m+1)} = X^{(m+1)} - X^*$$

$$E^{(m+1)} = Tr\ X^{(m)} - X^* + Gr\ (K)$$
$$= Tr\ X^{(m)} - (X^* - Gr\ (K))$$
$$= Tr\ X^{(m)} - Tr\ X^* = Tr\ (X^{(m)} - X^*)$$
$$E^{(m+1)} = Tr\ E^{(m)}$$

d'où

$$(18)^{**} \quad E^{(m+1)} = (Tr)^{m+1}\ E^{(o)}$$

$E^{(o)}$ étant le vecteur erreur associé au vecteur initial $X^{(o)}$.

Donc la méthode converge si $E^{(m+1)} = X^{(m+1)} - X^*$ tend vers zéro quand m augmente indéfiniment, c'est-à-dire si :

$$X^{(m+1)} \xrightarrow[m \longrightarrow \infty]{} X^*$$

et en examinant (18) et en tenant compte du théorème (A.8.), on voit que $E^{(m+1)} \xrightarrow[m \longrightarrow \infty]{}$ o si le rayon spectral de la matrice Tr est < 1.

C.2.

Calcul de ρ (Tr)

Considérons la matrice $\tilde{Tr}$ donnée par :

$$(19) \quad \tilde{Tr} = (rI + V)\ Tr\ (rI + V)^{-1}$$

les deux matrices Tr et $\tilde{Tr}$ ont les mêmes valeurs propres puisqu'elles sont semblables ; compte tenu de ce qui a été fait précédemment (relation (13)), il vient :

** le m entre parenthèses : (m) a une signification symbolique tandis que le m+1 de $(Tr)^{m+1}$ est une puissance.

$$(20) \quad \tilde{Tr} = \left\{ (rI-H)\ (rI+H)^{-1} \right\} \left\{ (rI-V)\ (rI+V)^{-1} \right\}$$

et en tenant compte du théorème (A.6.) et de son corollaire nous aurons :

$$(21) \quad \rho(Tr) = \rho\ (\tilde{Tr}) \leqslant \left\| Tr \right\| \leqslant \left\| (rI-H)\ (rI+H)^{-1} \right\| \cdot \left\| (rI-V)\ (rI+V)^{-1} \right\|$$

Soient λ_j les valeurs propres de H $\qquad 1 \leqslant j \leqslant N$

ν_j les valeurs propres de V $\qquad 1 \leqslant j \leqslant N$

Si nous examinons bien les relation (15) nous constatons que H a la forme suivante :

$$(22) \quad H = \frac{1}{h^2} \begin{vmatrix} H_1 & & & \\ & H_2 & & \\ & & H_3 & \\ & & & \ddots \\ & & & & H_n \end{vmatrix}$$

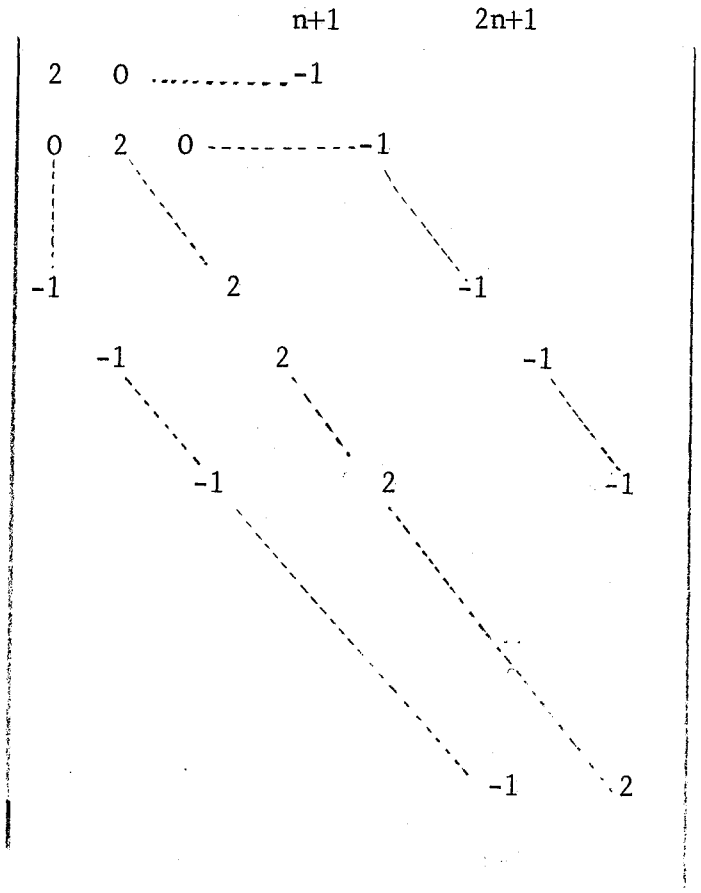
H d'ordre N

toutes les H_i dans le problème type sont des matrices carrés identiques d'ordre n avec :

$$H_i = \begin{vmatrix} 2 & -1 & & & \\ -1 & 2 & 1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{vmatrix}$$

et V la forme suivante :

$$(23)\quad V = \frac{1}{h^2}$$



la diagonale principale contient N = n x n éléments tous égaux à 2, les deux autres contiennent chacune N - n éléments tous égaux à - 1.

Remarque :

H et V n'ont les formes (22) et (23) que pour un vecteur inconnu X convenablement ordonné ; si u_{ij}, $1 \leqslant i,j \leqslant n$, sont les inconnues qu'on désigne par le vecteur X_ℓ, $1 \leqslant \ell \leqslant N$, dans ce cas X_ℓ est ordonné de la façon suivante :

$$(23)^{**}\quad X_\ell = u_{ij} \text{ avec } \ell = n\,(i-1) + j$$

i désignant la ligne et j la colonne.

** u_{ij} désigne la numérotation des points du carré.

On constate que pour le problème type H est hermitienne définie positive, alors $(rI - H)(rI + H)^{-1}$ est aussi hermitienne. Vu sa forme la matrice H possède n valeurs propres $\lambda_j \geq o$, $1 \leq j \leq n$ qui sont multiples d'ordre de multiplicité n, alors $(rI - H)(rI + H)^{-1}$ est hermitienne et possède n valeurs propres de la forme :

$$\mu_j = \left(\frac{r - \lambda_j}{r + \lambda_j} \right) \qquad 1 \leq j \leq n$$

les μ_j ayant le même ordre de multiplicité que λ_j

d'où :

$$(24) \quad \left\| (rI-H).(rI+H)^{-1} \right\| = \max_{1 \leq j \leq n} \left| \frac{r-\lambda_j}{r+\lambda_j} \right| < 1 \qquad \underline{\text{car } r > 0}$$

De même, on démontre que :

$$(25) \quad \left\| (rI-V).(rI+V)^{-1} \right\| < 1$$

Il s'en suit que :

$$\boxed{\rho\,(Tr) < 1} \qquad \forall r > 0$$

d'où le théorème suivant :

C.2.1.

Théorème :

Si les deux matrices H et V, dont la somme est identique à la matrice A, sont hermitiennes définies non-négatives, où au moins l'une d'elles est définie positive, alors pour tout scalaire $r > 0$, la matrice Tr et par suite la méthode des directions alternées est convergente.

Maintenant, nous allons démontrer un théorème connu sous le nom du théorème de Frobenius et qui est d'une grande importance dans tout ce qui va suivre.

C.2.2.

Théorème de Frobenius.

Soient H et V deux matrices hermitiennes d'ordre N, alors il existe une base orthonormale de vecteurs propres $\left\{\alpha_i\right\}_{i=1}^{N}$ commune avec $H\alpha_i = \lambda_i \alpha_i$ et $V\alpha_i = \nu_i \alpha_i$, si et seulement si HV = VH.

En effet :

- si une base orthonormale commune de vecteurs propres $\left\{\alpha_i\right\}_{i=1}^{N}$

existe avec $H\,\alpha_i = \lambda_i\,\alpha_i$ et $V\,\alpha_i = \nu_i\,\alpha_i$, alors : $HV\alpha_i = \lambda_i\,\nu_i\,\alpha_i = VH\alpha_i$

pour tout i ; $1 \leqslant i \leqslant N$.

Soit maintenant un vecteur $X \in C^n$; comme $\left\{\alpha_i\right\}_{i=1}^{N}$ est une base de C^n alors il existe des constantes c_i telles que :

$$X = \sum_{i=1}^{N} c_i\,\alpha_i$$

en tenant compte de ce qui précède, il s'en suit que :

$$HVX = VHX$$

et comme ceci est vrai pour tout vecteur $X \in C^n$ nous en concluons que :

$$\underline{HV = VH.}$$

- Réciproquement, supposons HV = VH. Comme H est hermitienne soit U une matrice unitaire d'ordre N qui diagonalise H de la façon suivante :

$$UHU^{*} = \begin{vmatrix} \lambda_1 I_1 & & & \\ & \lambda_2 I_2 & & \\ & & \ddots & \\ & & & \lambda_r I_r \end{vmatrix} = \tilde{\tilde{H}}$$

où les $\lambda_1 < \lambda_2 \ldots < \lambda_r$ sont les différentes valeurs propres de H et où les I_j sont des matrices unité d'ordre n_j avec :

$$N = \sum_{1}^{r} n_j \qquad \text{et } \lambda_j \neq 0$$

Formons maintenant $\tilde{V} = U V U^*$, nous aurons :

$$\tilde{V} = \begin{vmatrix} V_{11} & V_{12} & \cdots\cdots & V_{1r} \\ V_{21} & V_{22} & \cdots\cdots & V_{2r} \\ \vdots & & & \vdots \\ V_{r1} & V_{r2} & & V_{rr} \end{vmatrix}$$

mais il est clair que $HV = VH$ implique $\tilde{V}\tilde{H} = \tilde{V}H$ en effet :

$$\tilde{V}\,\tilde{H} = UV\,U^*\,U\,H\,U^* = UV\,H\,U^*$$

et

$$\tilde{H}\,\tilde{V} = UH\,U^*\,U\,V\,U^* = UH\,V\,U^*$$

et comme $HV = VH$ alors :

$$UV\,H\,U^* = UH\,V\,U^*$$

ce qui implique $\tilde{V}\,\tilde{H} = \tilde{H}\,\tilde{V}$

Alors en effectuant la multiplication $\tilde{V}\,\tilde{H}$ et $\tilde{H}\,\tilde{V}$ et en égalant terme à terme, nous trouvons que $V_{jk} \equiv 0$ pour tout $j \neq k$.

Ainsi la matrice hermitienne $\tilde{V}$ se trouve réduite à une matrice formée par des blocs de sous-matrices V_{jj} $1 \leqslant j \leqslant r$, chaque V_{jj} hermitienne possède n_j vecteurs propres orthonormaux qui sont en même temps vecteurs propres de la sous-matrice diagonale correspondante $\lambda_j\, I_j$ de $\tilde{H}$; et la totalité de ces vecteurs propres engendre N vecteurs orthonormaux α_i avec :

$$H\,\alpha_i = \lambda_i\,\alpha_i$$

$$V\,\alpha_i = \nu_i\,\alpha_i$$

c.q.f.d.

C.2.3.

Expression de $\rho(Tr)$

Alors si on admet que H et V ont même base $\{\alpha_i\}$ de vecteurs propres ** on peut écrire :

$$(26) \quad Tr\ \alpha_i = \left(\frac{r-\lambda_i}{r+\lambda_i}\right)\left(\frac{r-\nu_i}{r+\nu_i}\right)\alpha_i \qquad 1 \leq i \leq N$$

Soient μ_i les valeurs propres de Tr, de (26), nous tirons :

$$(27) \quad \mu_i = \left(\frac{r-\lambda_i}{r+\lambda_i}\right)\left(\frac{r-\nu_i}{r+\nu_i}\right) \qquad 1 \leq i \leq N$$

et

$$(28) \quad \rho\ (Tr) = \max_i \left|\left(\frac{r-\lambda_i}{r+\lambda_i}\right)\left(\frac{r-\nu_i}{r+\nu_i}\right)\right| = \max_i \left|\left(\frac{\lambda_i-r}{\lambda_i+r}\right)\left(\frac{\nu_i-r}{\nu_i+r}\right)\right|$$

$$\rho\ (Tr) = \max_i \left| \phi\ (\lambda_i).\ \phi\ (\nu_i) \right|$$

avec :

$$\phi\ (x_i) = \frac{x_i-r}{x_i+r}$$

Donc si je pose $\phi\ (x) = \frac{x-r}{x+r}$ et si je suppose que :

$$a \leq \nu_i,\ \lambda_i \leq b$$

alors :

$$(29) \quad \rho\ (Tr) \leq \left[\max_{x\ \epsilon\ [a,b]} \left|\phi\ (x)\right| \right]^2$$

et en général :

$$\left|\mu\right| = \left|\phi\ (\nu_i).\ \phi\ (\lambda_i)\right| \leq \rho\ (Tr)$$

** démonstration plus loin, en C.3.

C.2.4.

Minimisation de ρ (Tr).

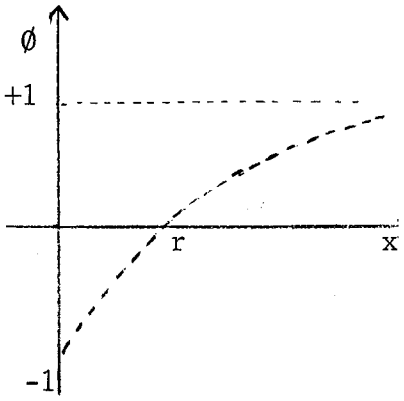
Il est bien évident que la méthode de directions alternées est d'autant plus convergente que le rayon spectral de la matrice Tr est plus petit, ce rayon ρ (Tr) dépendant d'un paramètre r, on peut choisir ce paramètre de telle sorte que ρ (Tr) soit le plus petit possible.

Mais minimiser ρ (Tr) revient à minimiser $\max_x \left| \phi(x) \right|$. Cherchons d'abord

$$\max_{a \leq x \leq b} \left| \phi(x) \right| = \max_{a \leq x \leq b} \left| \frac{x-r}{x+r} \right| = \max_{a \leq x \leq b} \left| \frac{r-x}{r+x} \right|$$

Considérons la fonction :

$$\phi(x) = \frac{x-r}{x+r}$$



alors :

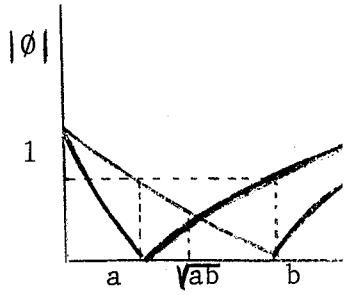
$$\frac{d\,\phi(x)}{dx} = \frac{2r}{(x+r)^2} > 0 \qquad \text{car } r > 0$$

donc la fonction ϕ (x) est croissante et son maximum absolu dans l'intervalle $[a,b]$ est atteint soit en a, soit en b

$$(30) \quad \max_{a \leq x \leq b} \left| \frac{r-x}{r+x} \right| = \max \left(\left| \frac{r-a}{r+a} \right| , \left| \frac{r-b}{r+b} \right| \right)$$

d'où :

$$(31) \quad \max_{a \leq x \leq b} \left| \frac{r-x}{r+x} \right| = \begin{cases} \dfrac{b-r}{b+r} & \text{si } 0 < r \leq \sqrt{ab} \\ \dfrac{r-a}{r+a} & \text{si } r \geq \sqrt{ab} \end{cases}$$



Je dis que, d'après (31) le $\min_{r} \max_{a \leq x \leq b} \left|\frac{r-x}{r+x}\right|$ est

donné pour x = b (ou x = a) quand $r = \sqrt{ab}$; en effet :

ou bien le maximum minimisé de $\left|\frac{r-x}{r+x}\right|$ est atteint deux fois dans $[a,b]$ une fois en $\underline{a}$ et une fois en b ou bien il est atteint une seule fois en $\underline{b}$ par exemple (ou en $\underline{a}$).

Ⓐ Supposons que le minimax est atteint deux fois, ce qui implique :

$$(31)' \qquad \frac{\bar{r}-a}{\bar{r}+a} = \frac{b-\bar{r}}{b+\bar{r}}$$

$\bar{r}$ étant le paramètre répondant au problème, donc

$$(\bar{r} - a)(b + \bar{r}) = (b - \bar{r})(\bar{r} + a)$$

et en effectuant le calcul, on trouve $\bar{r} = \pm \sqrt{ab}$ et comme $r > 0$ alors :

$$\underline{\bar{r} = \sqrt{ab}}$$

Ⓑ Supposons que :

$$(31)'' \qquad \left|\frac{r-a}{r+a}\right| < \left|\frac{b-r}{b+r}\right| \qquad \text{par exemple}$$

donc, il s'agit de minimiser la quantité $\left|\frac{b-r}{b+r}\right|$

Si r diminue la quantité $\left|\frac{b-r}{b+r}\right|$ augmente, par contre si r augmente la même quantité diminue ; mais r ne peut pas augmenter indéfiniment, car l'inégalité (31)" n'est vraie que si $r \leq \sqrt{ab}$ (voir (31)) donc le minimum est atteint en b pour $\bar{r} = \sqrt{ab}$ et il vaut :

$\frac{b-\sqrt{ab}}{b+\sqrt{ab}} = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}$; mais si $\bar{r} = \sqrt{ab}$ alors la valeur de la fonction $\frac{\bar{r}-a}{\bar{r}+a}$ serait $\frac{\sqrt{ab}-a}{\sqrt{ab}+a} = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}$. Donc le minimax de $\left|\frac{x-r}{x+r}\right|$ est atteint deux fois : une fois en $\underline{a}$ et une autre en $\underline{b}$ pour $\bar{r} = \sqrt{ab}$.

Donc le $\min_r \rho(Tr)$ est atteint pour $\bar{r} = \sqrt{ab}$, et

$$(32) \quad \min_r \rho(Tr) = \left(\frac{\sqrt{ab} - a}{\sqrt{ab} + a}\right)^2 = \left(\frac{b - \sqrt{ab}}{b + \sqrt{ab}}\right)^2 = \left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}\right)^2$$

ou

$$(32)' \quad \min_r \rho(Tr) = \left(\frac{1 - \sqrt{a/b}}{1 + \sqrt{a/b}}\right)^2 = \rho(T\bar{r})$$

et nous sommes alors sûrs que la matrice $T\bar{r}$ et par suite la méthode de directions alternées, convergerait plus vite que n'importe quelle autre matrice Tr où $r \neq \bar{r} = \sqrt{ab}$

Résultats numériques :

Pour un carré de côté 1 avec un pas h = 0,1 donc avec 9 points suivant o x et 9 points suivant o y, nous avons pris :

$$a = 9,\ 7887 \qquad b = 390,\ 21$$

a et b étant donnés par les relations (46)' et (46)"

$$\begin{cases} \text{avec } r = \bar{r} = \sqrt{ab} = 61,\ 803 & 17 \text{ itérations} = 2^{sec} \\ \text{avec } r = 390 & 96 \text{ itérations} = 10^{sec} \end{cases}$$

Nous avons choisi comme test d'arrêt la valeur du résidu relatif, c'est-à-dire :

$$\text{Résidu} = \frac{\max_{x_i} |AX - K|}{\max_{x_i} |x_i|} \qquad x_i \text{ composantes de } X$$

et nous avons pris le test suivant Résidu $< 10^{-5}$ avec $r = \bar{r}$ le programme écrit en Fortran s'arrête après 2^{sec} de calcul, et avec $r = 390 \neq \bar{r}$ après 10^{sec}. et passe sur 7044 (a 2 µs de cycle de base)

C.3.

Calcul des valeurs et vecteurs propres de H et de V.

On a supposé pour faire le calcul précédent que HV = VH, ou bien d'après le théorème de Frobenius que H et V ont même base orthonormale de vecteurs propres, dans ce qui va suivre, nous allons justifier cette supposition.

Pour le problème type H et V ayant les forme (22) et (23), soient λ_j les valeurs propres de H et ν_j celles de V. Si X^j est un vecteur propre de H correspondant à λ_j alors :

$$(33)^{**} \quad H\, X^j = \lambda_j\, X^j \qquad 1 \leq j \leq n$$

qui se traduit par :

$$(34) \quad -X^j{}_{i-1} + 2\, X^j{}_i - X^j{}_{i+1} = \lambda_j\, X^j{}_i \qquad \forall_i = 1, \ldots, n$$

avec :

$$(35) \quad \begin{cases} X^j{}_o = 0 \\ X^j{}_{n+1} = 0 \end{cases}$$

ou plus généralement

$$(36) \quad -X_{i-1} + 2\, X_i - X_{i+1} = \lambda\, X_i \qquad \forall_i = 1, \ldots, n$$

en posant $X_i = \rho^i$, (ρ^i = ρ puissance i), il vient :

$$(37) \quad \rho^2 - (2-\lambda)\,\rho + 1 = 0$$

$$\rho^2 - \alpha\,\rho + 1 = 0 \qquad \text{avec} \quad \alpha = 2 - \lambda$$

d'où l'on tire :

$$\rho = \frac{1}{2}\,(\alpha \pm \sqrt{\alpha^2 - 4}) \qquad \text{d'où } \rho_1 \text{ et } \rho_2$$

et la solution générale est :

$$(39) \quad X_i = C_1\, \rho_1{}^i + C_2\, \rho^i{}_2$$

en tenant compte des relations (35) nous devons avoir :

$$(40) \quad \begin{cases} X_o = C_1 + C_2 = 0 \\ X_{n+1} = C_1\, \rho^{n+1}{}_1 + C_2\, \rho^{n+1}{}_2 = 0 \end{cases}$$

** l'exposant j est symbolique.

** H étant une matrice par blocs de la forme (22) elle possède n^2 valeurs propres chacune est n - uple, les valeurs propres qu'on cherche sont celles d'un bloc.

la solution non nulle du système (40) est $C_1 = -C_2 = -C_2$ qui implique :

$\rho_1^{n+1} = \rho_2^{n+1}$, c'est-à-dire :

$$\left(\alpha + \sqrt{\alpha^2 - 4}\right)^{n+1} = \left(\alpha - \sqrt{\alpha^2 - 4}\right)^{n+1}$$

ou encore :

$$(41) \quad \left(\alpha + \sqrt{\alpha^2 - 4}\right)^{n+1} = e^{2iK\pi} \left(\alpha - \sqrt{\alpha^2 - 4}\right)^{n+1}$$

d'où :

$$\left(\alpha + \sqrt{\alpha^2 - 4}\right) = e^{\frac{2iK\pi}{n+1}} \left(\alpha - \sqrt{\alpha^2 - 4}\right)$$

$$\left(\alpha + \sqrt{\alpha^2 - 4}\right)^2 = 4 . e^{\frac{2iK\pi}{n+1}}$$

$$\alpha + \sqrt{\alpha^2 - 4} = \pm 2 e^{\frac{iK\pi}{n+1}}$$

$$-\alpha \pm 2 e^{\frac{iK\pi}{n+1}} = (\alpha^2 - 4)^{1/2}$$

et en élevant au carré les deux membres :

$$(42) \quad \alpha^2 + 4 e^{\frac{2iK\pi}{n+1}} \mp 4 \alpha e^{\frac{iK\pi}{n+1}} = \alpha^2 - 4$$

on en tire :

$$\pm \alpha = e^{\frac{iK\pi}{n+1}} + e^{-\frac{iK\pi}{n+1}} = 2 \cos \frac{K\pi}{n+1}$$

et :

$$\alpha = \pm 2 \cos \frac{K\pi}{n+1} = 2 - \lambda$$

d'où :

$$\lambda = 2 \pm 2 \cos \frac{K\pi}{n+1} = 2 \left(1 \pm \cos \frac{K\pi}{n+1}\right) , \quad 1 \leq K \leq n$$

$$\lambda = 4 \sin^2 \frac{K\pi}{2(n+1)} \qquad K = 1, \ldots, n$$

car :

$$\cos \frac{\pi}{n+1} = -\cos \frac{n\pi}{n+1}$$

$$\vdots$$

$$\cos \frac{j\pi}{n+1} = -\cos \frac{n+1-j}{n+1} \pi$$

et en tenant compte du facteur $\frac{1}{h^2}$ qui multiplie H (voir (22)) il vient :

$$(43) \quad \lambda = \frac{4}{h^2} \sin \frac{K\pi}{2(n+1)}$$

mais comme $\pi = (n+1)\,h$, alors (43) devient :

$$(44) \quad \boxed{\lambda = \frac{4}{h^2} \sin^2 \frac{Kh}{2}}$$

pour K allant de 1 à n

$\lambda K = \lambda$ pour un K donné

Considérons le vecteur $\in R^{n^2}$

$$(45) \quad X_{KL}(x,y) = \sin Kx \sin Ly$$

$K \; ; \; 1 \leqslant K \leqslant n$

$L \; ; \; 1 \leqslant L \leqslant n$

$x \; ; \; 1 \leqslant x \leqslant n$

$y \; ; \; 1 \leqslant y \leqslant n$

On vérifie bien que :

$$(45)' \quad HX_{KL}(x,y) = \lambda_K X_{KL}(x,y) \qquad \forall K = 1, \ldots, n$$

L quelconque.

En effet, en combinant les relations (15) et (45), nous aurons :

$$HX_{KL}(x,y) = \frac{1}{h^2} \left\{ - X_{KL}(x-h, y) + 2 X_{KL}(x,y) - X_{KL}(x+h,y) \right\}$$

$$= \frac{1}{h^2} \left\{ - \sin K(x-h) + 2 \sin Kx - \sin K(x+h) \right\} \sin Ly$$

$$= \frac{1}{h^2} \left\{ - 2 \sin Kx \cos Kh + 2 \sin Kx \right\} \sin Ly$$

$$= \frac{1}{h^2} \left\{ 2 (1 - \cos Kh) \right\} \sin Kx \sin Ly$$

$$= \frac{4}{h^2} \sin^2 \frac{Kh}{2} + \sin Kx \sin Ly$$

$$= \lambda_K X_{KL}(x,y) \qquad \text{c.q.f.d.}$$

Appliquons V à $X_{KL}(x,y)$.

$$VX_{KL}(x,y) = \frac{1}{h^2} \sin Kx \left\{ -\sin L(y-h) + 2 \sin Ly - \sin L(y+h) \right\}$$

$$= \frac{1}{h^2} (2 \sin Ly - 2 \sin Ly \cos Lh) . \sin Kx$$

$$= \frac{2}{h^2} (1 - \cos Lh) \sin Kx \sin Ly$$

$$= \frac{4}{h^2} \sin^2 \frac{Lh}{2} . \sin Kx \sin Ly$$

$$= \frac{4}{h^2} \sin^2 \frac{Lh}{2} \; X_{KL}(x,y)$$

d'où :

$$(46) \quad \boxed{\nu = \frac{4}{h^2} \sin^2 \frac{Lh}{2}} \qquad \text{pour } L = 1, \ldots, n$$

$$\nu_L = \nu \quad \text{pour } L \text{ donné}$$

et les deux matrices H et V ont la même base de vecteurs propres qui sont les $X_{KL}(x,y)$ donnés par (45) , donc l'étude faite précédemment pour le cas d'un carré est valable.

et les bornes strictes sont bien déterminées à savoir :

$$(46)' \quad a = \frac{4}{h^2} \sin^2 \frac{h}{2}$$

$$(46)'' \quad b = \frac{4}{h^2} \sin^2 \frac{nh}{2} = \frac{4}{h^2} \cos^2 \frac{h}{2} \qquad \text{car } (n+1)h = \pi$$

alors en tenant compte de (32)', il vient :

$$\min_r \rho(Tr) = \rho(T\bar{r}) = \left(\frac{1 - \sqrt{a/b}}{1 + a/b} \right)^2$$

$$= \left(\frac{1 - \operatorname{tg} \frac{h}{2}}{1 + \operatorname{tg} \frac{h}{2}} \right)^2$$

$$(47) \qquad = \left(\frac{1 + \cos h - \sin h}{1 + \cos h + \sin h} \right)^2$$

Mais si nous calculons le rayon spectral de la matrice $\mathcal{L}$ de la méthode de relaxation successive pour le même problème et avec le paramètre ω^* minisant ce rayon spectral, on trouve :

$$(48)^{**} \quad \rho\ (\mathcal{L}_{\omega^*}) = \frac{1 - \sin h}{1 + \sin h}$$

Il est facile alors de vérifier que :

$$(49) \quad \min_r \rho\ (Tr) = \rho\ (\mathcal{L}_{\omega^*})$$

En effet :

$$(1 + \cos h - \sin h)^2 = 1 + \cos^2 h + \sin^2 h + 2\cos h - 2 \sin h - 2\cos h \sin h$$

$$= 2\ (1 - \sin h + \cos h - \cos h \sin h)$$

$$= 2\ (1 - \sin h)\ (1 + \cos h)$$

$$(1 + \cos h + \sin h)^2 = 1 + \cos^2 h + \sin^2 h + 2\cos h + 2 \sin h + 2 \cos h \sin h$$

$$= 2\ (1 + \sin h + \cos h + \cos h \sin h)$$

$$= 2\ (1 + \sin h)\ (1 + \cos h)$$

d'où :

$$\min_r \rho\ (Tr) = \left(\frac{1+\cos h - \sin h}{1+\cos h + \sin h}\right)^2 = \frac{1 - \sin h}{1 + \sin h} = \rho\ (\mathcal{L}_{\omega^*})$$

Conclusion

Comme le nombre d'opérations arithmétiques est plus grand dans la méthode présente que dans celle de relaxation successive, cette méthode ne sera éventuellement intéressante que pour un choix non constant du paramètre r.

** Voir Varga page 110 et 216.

II - PARTIE

-=-=-

METHODE DE PEACEMAN - RACHFORD

-=-=-=-=-=-

II

METHODE DE PEACEMAN - RACHFORD

A. - Principe

Considérons les équations (11) i.e.

$$(50)\quad \begin{cases} (H + rI)\, X^{(m+1/2)} = (rI - V)\, X^{(m)} + K \\ (V + rI)\, X^{(m+1)} = (rI - H)\, X^{(m+1/2)} + K \end{cases}$$

L'idée principale de Peaceman-Rachford est de faire varier r d'une itération à une autre, c'est-à-dire de sélectionner un ensemble de t paramètres $\{r_1, \ldots\ldots, r_t\}$ et de faire la première itération avec r_1, la deuxième avec r_2, etc .... la $t^{ième}$ itération avec r_t et ensuite grouper ces t itérations en un seul "pas", i.e.

$$(51)\quad \begin{cases} (H + r_1 I)\, X^{(m+1/2)} = (r_1 I - V)\, X^{(m)} + K \\ (V + r_1 I)\, X^{(m+1)} = (r_1 I - H)\, X^{(m+1/2)} + K \\ (H + r_2 I)\, X^{(m+3/2)} = (r_2 I - V)\, X^{(m+1)} + K \\ (V + r_2 I)\, X^{(m+2)} = (r_2 I - H)\, X^{(m+3/2)} + K \\ \vdots \\ (H + r_t I)\, X^{(m+t-1/2)} = (r_t I - V)\, X^{(m+t-1)} + K \\ (V + r_t I)\, X^{(m+t+1)} = (r_t I - H)\, X^{(m+t-1/2)} + K \end{cases}$$

On en tire :

$$(52)\quad \begin{cases} X^{(m+1)} = Tr_1\, X^{(m)} + Gr_1\,(K) \\ X^{(m+2)} = Tr_2\, X^{(m+1)} + Gr_2\,(K) \\ \vdots \\ X^{(m+t)} = Tr_t\, X^{(m+t-1)} + Gr_t\,(K) \end{cases}$$

et en posant :

$$(53)\quad \mathcal{C} = \prod_{j=1}^{t} Tr_j = Tr_t \cdot Tr_{t-1} \cdot \ldots\ldots\ldots Tr_2 \cdot Tr_1$$

il vient :

$$(54)\quad X^{(m+t)} = \mathcal{C}\, X^{(m)} + G$$

Donc ayant calculé $X^{(m+t)}$ par (54), alors je recommence le calcul en remplaçant dans la relation (54) $X^{(m)}$ par $X^{(m+t)}$ et je calcule ainsi $X^{(m+2t)}$,. donc en posant :

$$(55)\quad \begin{cases} X^{(m)} = X_o \\ X^{(m+t)} = X_1 \\ \vdots \\ X^{(m+jt)} = X_j \end{cases}$$

il vient :

$$(56)\quad \begin{cases} X_1 = \mathcal{C}\, X_o + G \\ X_j = \mathcal{C}\, X_{j-1} + G \end{cases}$$

Si j'associe au vecteur $X^{(m)}$ un vecteur erreur $\epsilon^{(m)} = \eta_o$, alors au vecteur $X^{(m+t)}$ serait associé un vecteur erreur

$$\epsilon^{(m+t)} = \eta_1$$

avec :

$$\epsilon^{(m+t)} = \left(\prod_{j=1}^{t} Tr_j\right)\ \epsilon^{(m)}$$

c'est-à-dire :

$$(57)\quad \begin{cases} \epsilon^{(m+t)} = \mathcal{C}\, \epsilon^{(m)} \\ \eta_1 = \mathcal{C}\, \eta_o \end{cases}$$

La convergence de cette méthode est d'autant plus rapide que le rayon spectral de la matrice $\mathcal{C}$ est plus faible.

B. - Calcul de $\rho(\mathcal{C})$

B.1. Supposons H V = V H, ce qui est vrai pour le "problème modèle ", alors en tenant compte des théorèmes qui précèdent et en particulier celui de Frobénius, on peut facilement démontrer que :

$$(58) \quad \mathcal{C}\,\alpha_i = \left\{ \prod_{j=1}^{t} \left(\frac{r_j - \lambda_i}{r_j + \lambda_i} \cdot \frac{r_j - \nu_i}{r_j + \nu_i} \right) \right\} \alpha_i \qquad 1 \leqslant i \leqslant n$$

$\{\alpha_i\}$ étant un vecteur propre (commun à H et V)

Alors :

$$(59) \quad \rho(\mathcal{C}) = \left\| \prod_{j=1}^{t} \right\| = \max_{1 \leqslant i \leqslant n} \prod_{j=1}^{t} \left(\left| \frac{r_j - \lambda_i}{r_j + \lambda_i} \right| \cdot \left| \frac{r_j - \nu_i}{r_j + \nu_i} \right| \right)$$

donc $\rho(\mathcal{C}) < 1$ car les r_j, λ_i, ν_i sont tous > 0

Soient a et b les bornes de l'intervalle qui contient les λ_i et ν_i, i. e.

$$0 < a < \lambda_i, \nu_i < b$$

Alors :

$$\max_i \prod_{j=1}^{t} \left(\left| \frac{r_j - \lambda_i}{r_j + \lambda_i} \right| \cdot \left| \frac{r_j - \nu_i}{r_j + \nu_i} \right| \right) \leqslant \left\{ \max_i \prod_{j=1}^{t} \left| \frac{r_j - \lambda_i}{r_j + \lambda_i} \right| \right\} \left\{ \max_i \prod_{j=1}^{t} \left| \frac{r_j - \nu_i}{r_j + \nu_i} \right| \right\}$$

$$(60) \qquad \leqslant \left\{ \max_{a<x<b} \prod_{j=1}^{t} \left| \frac{r_j - x}{r_j + x} \right| \right\}^2$$

Posons $\Phi_t(x ; r_j) \equiv \prod_{j=1}^{t} \left(\frac{r_j - x}{r_j + x} \right)$

alors :

$$(61) \quad \rho(\mathcal{C}) = \left\| \prod_{j=1}^{t} \mathrm{Tr}_j \right\| \leqslant \left\{ \max_{a<x<b} \left| \Phi_t(x ; r_j) \right| \right\}^2$$

Soit S_t l'ensemble des fonctions $\Phi_t (x ; r_j)$, posons alors :

$$(62) \quad m M_t (a,b) = \min_{\Phi_t \in S_t} \left\{ \max_{a<x<b} \left| \Phi_t (x ; r_j) \right| \right\}$$

Comme il a été dit précédemment, la méthode est d'autant plus convergente que le rayon spectral de la matrice $\mathcal{C}$ est plus petit ; donc ayant fixé le nombre t des paramètres, minimiser le rayon spectral de $\mathcal{C}$ revient tout simplement à la recherche de la quantité $m M_t$.

B.2. Théorème

"Il existe une fonction unique $\bar{\Phi}_t (x ; \bar{r}_j)$ appartenant à S_t et pour laquelle :

$$m M_t (a,b) = \max_{a \leq x \leq b} \left| \bar{\Phi}_t (x ; \bar{r}_j) \right| .$$

En effet :

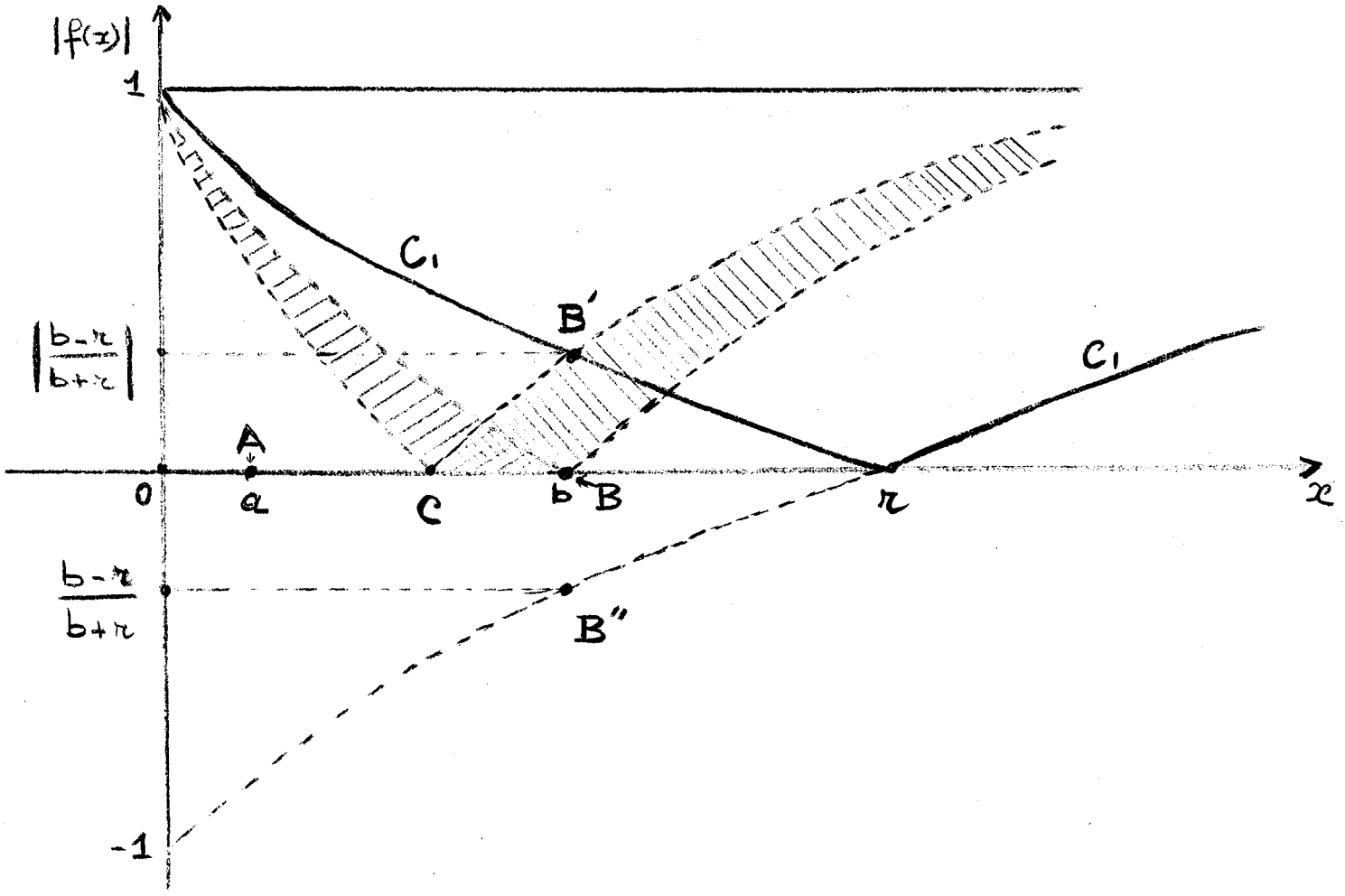
B.2.1. Lemme

Soient $[a,b]$ un intervalle donné et r un scalaire n'appartenant pas à cet intervalle, alors il existe un nombre $\sigma \in [a,b]$ tel que :

$$\left| \frac{x - \sigma}{x + \sigma} \right| < \left| \frac{x - r}{x + r} \right| \quad \text{pour tout } x \in [a,b] .$$

en effet, supposons $r > b$ et soit C_m le faiseau d'hyperboles équilatères d'équations :

$$(63) \quad y \equiv f_m (x) = \frac{x-m}{x+m}$$



Pour un $r > b$ le graphe de la fonction $x \longrightarrow \left|f(x)\right|$ a la forme ci-dessus : C_1 .

Le point B" ayant pour abscisse b a pour ordonnée $\frac{b-r}{b+r}$.

Si je considère la fonction f(x) passant par B' symétrique de B" par rapport à l'axe des x, elle aura pour expression :

$$y = f(x) = \frac{x - \tilde{r}}{x + \tilde{r}}$$

avec :

$$\left|\frac{b - r}{b + r}\right| = \frac{b - \tilde{r}}{b + \tilde{r}} = \frac{r - b}{r + b}$$

d'où l'on tire :

$$\tilde{r} = \frac{b^2}{r}$$

Soit **C** le point de l'axe des x ayant pour abscisse $\frac{b^2}{r}$; Ainsi tout σ tel que $\frac{b^2}{r} \leqslant \sigma \leqslant b$ vérifie le lemme en question. On procèdera de la même façon si $r < a$.

B.2.2. Lemme

"Si $\Phi(x) = \prod_{i=1}^{t} \left(\frac{x-r_i}{x+r_i} \right)$ est telle que tous les r_i ne soient pas contenus dans $[a,b]$, il existe une autre $\Phi'(x)$ avec des r'_i tous contenus dans $[a,b]$ telle que $M' \leqslant M$ avec :

$$M' = \max_x \left| \Phi'(x) \right|$$

$$M = \max_x \left| \Phi(x) \right| .$$

En effet :

Soit $\Phi(x)$ avec, par exemple, $r_1 \leqslant r_2 \leqslant \ldots \leqslant r_p \notin [a,b]$ et soit I l'ensemble des indices tels que $r_i \notin [a,b]$ et posons :

$$\varphi(x) = \prod_{i \notin I} \frac{x - r_i}{x + r_i}$$

Or d'après le lemme (B.2.1.) à tout r_i ($i \in I$) il correspond un $\mu_i \in [a,b]$ qui est tel :

$$(64) \quad \left| \frac{x - \mu_i}{x + \mu_i} \right| < \left| \frac{x - r_i}{x + r_i} \right| \qquad \text{pour tout } x \in [a,b]$$

$\mu_i \in [a,b]$ $\quad i \in I$ $\qquad$ $r_i \notin [a,b]$ $\quad i \in I$

Posons alors :

$$(65) \quad \Phi'(x) = \varphi(x) \cdot \prod_{i \in I} \left(\frac{x - \mu_i}{x + \mu_i} \right)$$

d'où :

$$\Phi'(x) = \Phi(x) \cdot \prod_{i \in I} \left(\frac{x+r_i}{x-r_i} \cdot \frac{x-\mu_i}{x+\mu_i} \right)$$

d'où, en tenant compte de l'inégalité (64), il vient :

$$M' \leq M \qquad \text{pour tout } x \in [a,b]$$

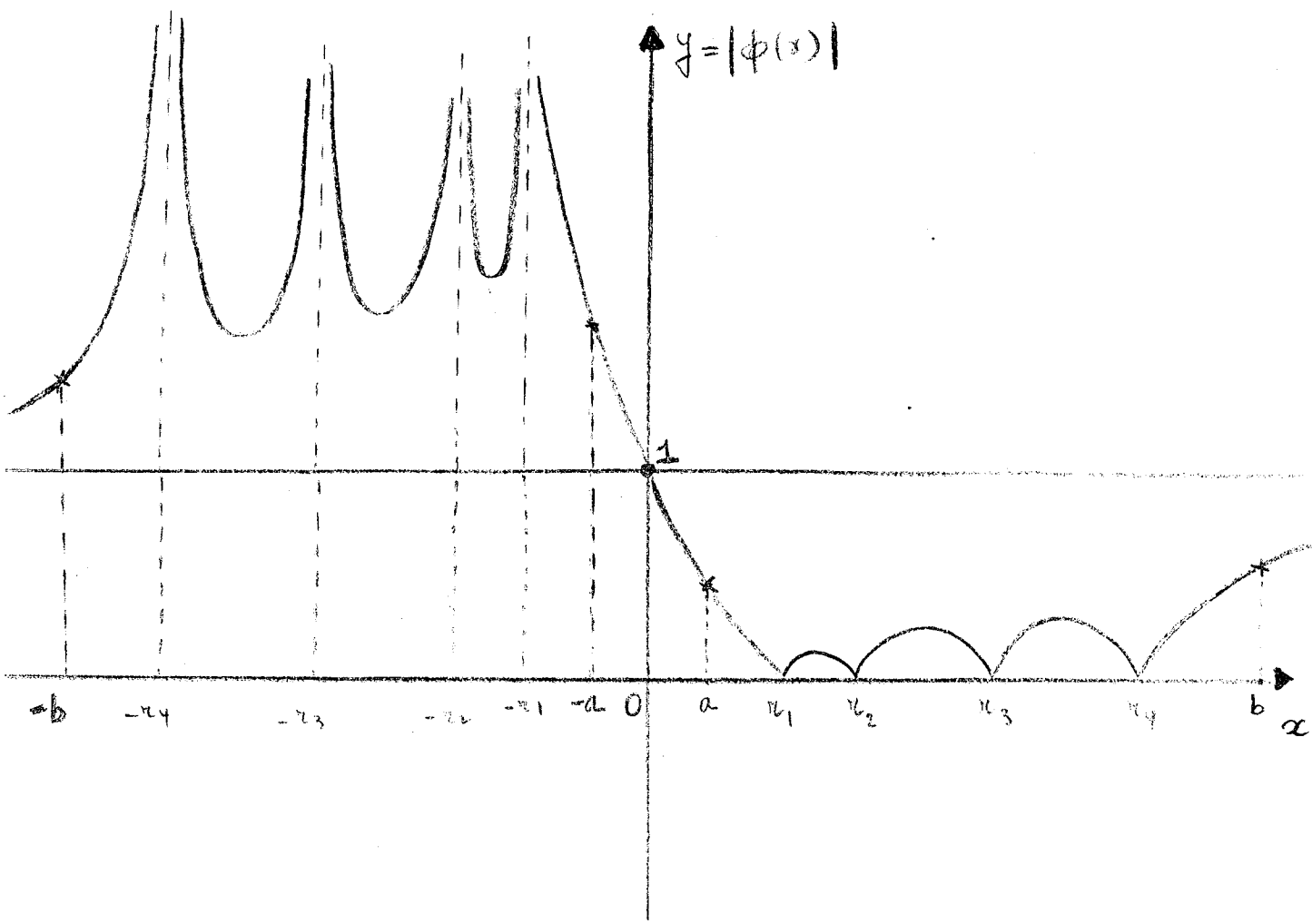
c.q.f.d.

B.2.3. Remarque fondamentale

Le lemme qu'on vient de démontrer nous permet, sans équivoque, en cherchant le minimum en question, donc en calculant les paramètres accélérateurs qui conviennent le plus, de supposer que tous ces paramètres sont contenus dans [a,b] .

B.2.4.

Allure générale de la courbe $x \longrightarrow |\Phi(x)|$



Pour des r_i différents l'allure générale de la courbe de la fonction $x \longrightarrow |\bar{\Phi}(x)|$ est celle indiquée ci - dessus.

Pour des valeurs de x opposées en signe, les deux valeurs correspondantes de la fonction sont positives et inverses l'une de l'autre.

Il est clair que, pour x positive, $|\bar{\Phi}(x)| < 1$. On constate que $|\bar{\Phi}(x)|$ admet au plus $(t+1)$ maxima relatif dans $[a,b]$: $(t-1)$ pour les zéros des dérivées séparés par les r_i et deux autres en a et b (t étant le nombre de paramètres).

B.2.5. Théorème

Supposons qu'il existe un choix des paramètres $r_i \in [a,b]$, distincts et tels que la fonction $|\bar{\Phi}(x)|$, qui correspond à ce choix, atteigne sa valeur maximale pour a,b et (t-1) valeurs distinctes séparées par les t valeurs des r_i. Soient x_i, i = 1, ..., (t-1), ces valeurs ; soit $M = |\bar{\Phi}(x)|$, alors si une autre fonction $|\bar{\Phi}'(x)|$ est telle que son maximum M' est $M' \leqslant M$ on a forcément $\bar{\Phi}'(x) \equiv \bar{\Phi}(x)$.

En effet, considérons la différence :

$$D(x) = \bar{\Phi}(x) - \bar{\Phi}'(x)$$

elle est définie et continue dans l'intervalle $[a,b]$.

Pour un x_i sa valeur est :

$$(66) \quad D(x_i) = \bar{\Phi}(x_i) - \bar{\Phi}'(x_i)$$

Je suppose que :

$$\bar{\Phi}(x_i) = +M$$

$$\bar{\Phi}(x_{i+1}) = -M$$

alors (66) devient :

$$D(x_i) = +M - \bar{\Phi}'(x_i) \geq 0$$

$$D(x_{i+1}) = -M - \bar{\Phi}'(x_{i+1}) \leqslant 0$$

car on a par hypothèse $M' \leqslant M$

Donc D (x) change de signe (t+1) fois dans $[a,b]$: il y a donc au moins t zéros de cette fonction dans $[a,b]$.

Posons alors :

$$(67)\quad \begin{cases} f(x) = (x-r_1)(x-r_2) \ldots\ldots (x-r_t) \\ f'(x) = (x-r'_1)(x-r'_2)\ldots\ldots(x-r'_t) \end{cases}$$

Nous pouvons alors écrire :

$$(68)\quad \begin{cases} \Phi(x) = \dfrac{f(x)}{(-1)^t f(-x)} \\ \Phi'(x) = \dfrac{f'(x)}{(-1)^t f'(-x)} \end{cases}$$

et :

$$D(x) = \frac{f(x).f'(-x) - f'(x).f(-x)}{(-1)^t\ f\ (-x).\ f'(-x)}$$

les t zéros de D(x) qui se trouvent dans $[a,b]$ sont ceux du polynôme

$$(69)\quad P\ (x) = f(x).\ f'(-x) - f'(x).\ f(-x)$$

On constate bien que :

- P(x) est de degré(2t-1) puisque le terme de degré 2t est : $x^t.(-1)^t\ x^t - x^t(-1)^t.\ x^t = 0$
- son terme constant est nul
- $P(x) = -\ P\ (-x)$

Donc P est impair.

D'où, en supposant F(x) un polynôme de degré t-1, il vient :

$$(70)\quad P(x) = x\ [F(x^2)]$$

Mais P(x) admet t zéros distincts dans $[a,b]$, donc F(x) admet t zéros aussi, donc $F(x) \equiv 0$ et par suite :

$$(71)\quad P\ (x) \equiv 0$$

c.q.f.d.

Mais $f(x).\ f'(-x) = 0$ pour les $x = r_i$; en tenant compte de (71) et (69), on peut écrire :

$$f'(x)\ \ f(-x) = 0 \qquad \text{pour les } x = r_i$$

donc :

$$\begin{cases} f(r_i).\ f'(-r_i) = 0 \\ f'(r_i).f(-r_i) = 0 \end{cases} \qquad i = 1, \ldots\ldots, t$$

mais $f(-r_i) \neq 0$

donc $f'(r_i) = 0 \qquad i = 1, \ldots\ldots, t$

Mais on a aussi :

$$f'(r'_i) = 0 \qquad i = 1, \ldots., t$$

D'où : $r_i = r'_i$ pour tout $i = 1, \ldots\ldots, t$

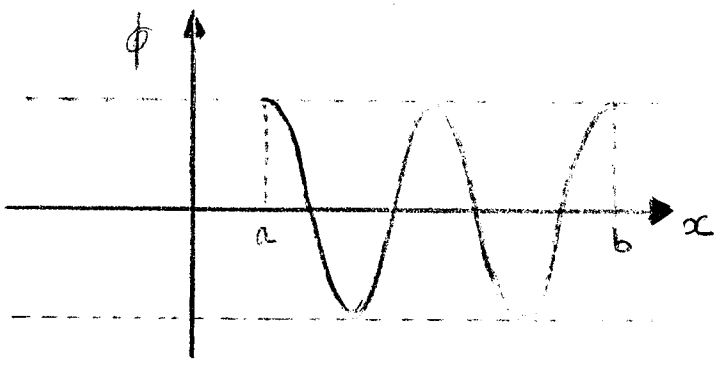
et :

$$\Phi(x) \equiv \Phi'(x)$$

ce qui démontre l'unicité de la solution et par conséquent le théorème (B.2)

Le théorème (B.2) est d'une très grande importance car il prouve qu'un principe analogue à celui de Tchebyschef s'applique ici, ce qui n'est pas, à priori, évident. M. GASTINEL ** qui, le premier, a étudié l'unicité de la solution, a démontré que la fraction, répondant au problème du minimax a la forme suivante :

D'autre part on trouve (même référence) qu'il existe une fonction de la famille S_t qui a les propriétés d'alternance, donc cela prouve l'existence et l'unicité de la solution du problème.



** Voir Chiffres 1962. [3]

B.2.6. Corollaire

Soit $\left\{\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_t\right\}$ l'ensemble des paramètres répondant à la solution du problème de min- max. Si $\bar{r}_j$ est un paramètre de cet ensemble, alors $ab/\bar{r}_j$ en est un aussi, de plus nous aurons :

$$\left|\Phi_t\ (x,\ \bar{r}_j)\right| = \left|\Phi_t\ \left(\frac{ab}{x}\ ,\ \bar{r}_j\ \right)\right| .$$

En effet :

Soit $y = ab/x$

alors :

$$\Phi_t(x,\ \bar{r}_j) = (-1)^t\ \Phi_t\ (y,\ \frac{ab}{\bar{r}_j}\ )$$

Comme $0 < a \leqslant x \leqslant b$ alors $0 < a \leqslant y \leqslant b$

et :

$$\max_{a\leqslant x\leqslant b}\ \left|\Phi_t\ (x,\bar{r}_j)\right| = \max_{a\leqslant y\leqslant b}\ \left|\Phi_t(\ y,\ \frac{ab}{\bar{r}_j}\ )\right|$$

et en tenant compte de l'unicité de la solution , la démonstration s'en suit.

* Remarque :

Pour le cas $t = 1$

il découle du corollaire précédent que :

$$\bar{r}_1 = \frac{ab}{\bar{r}_1}$$

donc :

$$\bar{r}^2_1 = ab \qquad \text{et } r_1 = \sqrt{ab} \qquad \text{car : } r_1 > 0$$

ce qui a été justement démontré auparavant.

C. - Calcul des paramètres accélérateurs

C.1. Nous résolvons maintenant le problème du min - max pour un ensemble de 2t paramètres :

$$r_1, r_2, \ldots\ldots, r_t, \ldots., r_{2t} \qquad \text{avec } 0 < a \leqslant x \leqslant b,$$

d'où :

$$(72) \qquad m\,M_{2t}(a,b) = \max_{a \leqslant x \leqslant b} \left| \prod_{j=1}^{2t} \left(\frac{\bar{r}_j - x}{\bar{r}_j + x} \right) \right|$$

ou bien, en tenant compte du corollaire précédent :

$$(73) \qquad m\,M_{2t}(a,b) = \max_{a \leqslant x \leqslant b} \left| \prod_{j=1}^{t} \left(\frac{\bar{r}_j - x}{r_j + x} \right) \left(\frac{ab/\bar{r}_j - x}{ab/\bar{r}_j + x} \right) \right|$$

Les 2t paramètres de (72) sont tels que $a < \bar{r}_j < b$, les t paramètres de (73) sont, à priori, aussi tels que $a < \bar{r}_j < b$. Mais comme $\bar{r}^2_j = ab$, alors les 2t paramètres de (72) sont situés de part et d'autre de $\sqrt{ab}$: t paramètres entre a et $\sqrt{ab}$ et t autres paramètres entre $\sqrt{ab}$ et b.

Ainsi, tous les paramètres de (73), au nombre de t, je les prends compris entre a et $\sqrt{ab}$, alors les $ab/\bar{r}_j$ seront compris entre $\sqrt{ab}$ et b, ce qui ne change rien à l'étude du problème du min-max.

Par ailleurs, je constate que le produit de deux facteurs de (73) ne change pas de valeur si on remplace x par ab/x : quand x varie entre a et $\sqrt{ab}$ alors ab/x varie entre $\sqrt{ab}$ et b.

D'où (73) devient :

$$(74) \qquad m\,M_{2t}(a,b) = \max_{a \leqslant x \leqslant \sqrt{ab}} \left| \prod_{j=1}^{t} \left(\frac{\bar{r}_j - x}{\bar{r}_j + x} \right) \cdot \left(\frac{ab/\bar{r}_j - x}{ab/\bar{r}_j + x} \right) \right|$$

$$= \max_{a \leqslant x \leqslant \sqrt{ab}} \left| \prod_{j=1}^{t} \left(\frac{ab - x\,\bar{r}_j - x.\,ab/\bar{r}_j + x^2}{ab + x.\,\bar{r}_j + x\,.\,ab/\bar{r}_j + x^2} \right) \right|$$

en divisant haut et bas par 2x il vient :

$$m\,M_{2t}(a,b) = \max_{a\leq x\leq \sqrt{ab}} \left| \prod_{j=1}^{t} \left(\frac{\frac{ab}{2x} - \frac{\bar{r}_j}{2} - ab/2\bar{r}_j + x/2}{\frac{ab}{2x} + \frac{\bar{r}_j}{2} + ab/2\bar{r}_j + x/2} \right) \right|$$

$$(75) \qquad m\,M_{2t}(a,b) = \max_{a\leq x\leq \sqrt{ab}} \left| \prod_{j=1}^{t} \frac{\frac{\sqrt{ab}}{2}\left(\frac{\sqrt{ab}}{x} + \frac{x}{\sqrt{ab}}\right) - \frac{1}{2}\left(\bar{r}_j + \frac{ab}{\bar{r}_j}\right)}{\frac{\sqrt{ab}}{2}\left(\frac{\sqrt{ab}}{x} + \frac{x}{\sqrt{ab}}\right) + \frac{1}{2}\left(\bar{r}_j + \frac{ab}{\bar{r}_j}\right)} \right|$$

en posant :

$$(76) \begin{cases} y \equiv \frac{\sqrt{ab}}{2}\left(\frac{\sqrt{ab}}{x} + \frac{x}{\sqrt{ab}}\right) \\ \bar{\omega}_j \equiv \frac{1}{2}\left(\bar{r}_j + \frac{ab}{\bar{r}_j}\right) \end{cases}$$

il vient :

$$(77) \qquad m\,M_{2t}(a,b) = \max_{\sqrt{ab}\leq y\leq \frac{a+b}{2}} \left| \prod_{j=1}^{t} \frac{y - \bar{\omega}_j}{y + \bar{\omega}_j} \right|$$

d'où :

$$(77)' \qquad m\,M_{2t}(a,b) = m\,M_t\left(\sqrt{ab}, \frac{a+b}{2}\right)$$

Ainsi, ayant les $\bar{\omega}_j$, $j = 1, \ldots, t$, il est facile de trouver les $\bar{r}_j$, $j = 1, \ldots, 2t$ par la formule (76).

Théorème :

On en conclut que le problème du min-max à 2t paramètres se ramène à un autre problème du min - max à t paramètres mais avec des bornes différentes

C.2. Cas où $t = 2^k$

Soit $o < a \leq x \leq b$

Posons alors :

$$\alpha_o = a$$
$$\beta_o = b$$

et formons les deux suites suivantes, connues sous le nom de suites de Fibonaci :

$$(78) \quad \begin{cases} \alpha_{i+1} = \sqrt{\alpha_i \beta_i} \\ \beta_{i+1} = \dfrac{\alpha_i + \beta_i}{2} \end{cases} \qquad i \geqslant 0$$

D'après (77)' on peut écrire :

$$(79) \quad m M_{2^k} \left[\alpha_o, \beta_o\right] = m M_{2^{k-1}} \left[\alpha_1, \beta_1\right] = \ldots\ldots = m M_{2^{\circ}=1} \left[\alpha_k, \beta_k\right]$$

Ainsi le problème se ramène à la recherche d'un seul paramètre dans l'intervalle $\alpha_k \leqslant x \leqslant \beta_k$ dont la solution est bien connue : $\bar{r} = \sqrt{\alpha_k \beta_k}$.

On a ainsi :

$$(80) \quad m M_{2^k} = m M_1 \left[\alpha_k, \beta_k\right] = \Phi_1 \left(\alpha_k, \sqrt{\alpha_k \beta_k}\right)$$

d'où :

$$(81) \quad m M_{2^k} \left[\alpha_o, \beta_o\right] = \frac{\sqrt{\beta_k} - \sqrt{\alpha_k}}{\sqrt{\beta_k} + \sqrt{\alpha_k}} \qquad k \geqslant 0$$

et en tenant compte de (61) il vient :

$$(82) \quad \rho(\mathcal{C}) = \left\| \prod_{j=1}^{t} Tr_j \right\| \leqslant \left(\left| m M_t (a,b) \right| \right)^2$$

D. - <u>Théorème important</u> :

On a :

$$"m M_{2^k} < \left(m M_{2^{k-1}}\right)^2 \ . \ "$$

En effet :

On a :

$$m M_{2^k} = \frac{\sqrt{\beta_k} - \sqrt{\alpha_k}}{\sqrt{\beta_k} + \sqrt{\alpha_k}}$$

$$m M_{2^{k-1}} = \frac{\sqrt{\beta_{k-1}} - \sqrt{\alpha_{k-1}}}{\sqrt{\beta_{k-1}} + \sqrt{\alpha_{k-1}}}$$

avec :

$$(83)\begin{cases} \beta_k = \dfrac{\alpha_{k-1} + \beta_{k-1}}{2} \\ \alpha_k = \sqrt{\alpha_{k-1} \cdot \beta_{k-1}} \\ \sqrt{\alpha_i \beta_i} < \dfrac{\alpha_i + \beta_i}{2} \qquad i > 0 \end{cases}$$

Posons :

$$A = m M_{2^k}$$

$$B = (m M_{2^{k-1}})^2$$

alors :

$$A - B = \frac{\sqrt{2}/2 (\alpha_{k-1} + \beta_{k-1})^{1/2} - (\alpha_{k-1}\beta_{k-1})^{1/4}}{\sqrt{2}/2 (\alpha_{k-1} + \beta_{k-1})^{1/2} + (\alpha_{k-1}\beta_{k-1})^{1/4}} - \frac{\alpha_{k-1} + \beta_{k-1} - 2(\alpha_{k-1}\beta_{k-1})^{1/2}}{\alpha_{k-1} + \beta_{k-1} + 2(\alpha_{k-1}\beta_{k-1})^{1/2}}$$

$$\Omega^* (A-B) = \frac{\sqrt{2}}{2} (\alpha_{k-1}+\beta_{k-1})^{3/2} - (\alpha_{k-1}\beta_{k-1})^{1/4}(\alpha_{k-1}+\beta_{k-1}) + \sqrt{2}(\alpha_{k-1}+\beta_{k-1})^{1/2} \cdot \cdot \cdot \cdot \cdot (\alpha_{k-1}\beta_{k-1})^{1/2}$$

$$- 2 (\alpha_{k-1}\beta_{k-1})^{3/4} - \frac{\sqrt{2}}{2} (\alpha_{k-1}\beta_{k-1})^{3/2} - (\alpha_{k-1}+\beta_{k-1})(\alpha_{k-1}\beta_{k-1})^{1/4}$$

$$+ \sqrt{2} (\alpha_{k-1}+\beta_{k-1})^{1/2} (\alpha_{k-1}\beta_{k-1})^{1/2} + 2(\alpha_{k-1}\beta_{k-1})^{3/4}$$

avec Ω = produit des deux dénominateurs ; Ω est positive car les α_{k-1} et β_{k-1} sont positives ; d'où il vient :

$$\Omega^* (A-B) = 2\left[\sqrt{2}(\alpha_{k-1}+\beta_{k-1})^{1/2}\,(\alpha_{k-1}\cdot\beta_{k-1})^{1/2} - (\alpha_{k-1}+\beta_{k-1})(\alpha_{k-1}\cdot\beta_{k-1})^{1/4}\right]$$

$$= 2\,(\alpha_{k-1}\beta_{k-1})^{1/4}\,(\alpha_{k-1}+\beta_{k-1})^{1/2}\left[\sqrt{2}\,(\alpha_{k-1}\cdot\beta_{k-1})^{1/4} - (\alpha_{k-1}+\beta_{k-1})^{1/2}\right]$$

posons :

$$\Sigma = \Omega^* \quad \frac{1}{2\sqrt{2}\,(\alpha_{k-1}\cdot\beta_{k-1})^{1/4}(\alpha_{k-1}+\beta_{k-1})} = \text{quantité positive}$$

il vient :

$$\Sigma^* (A-B) = (\alpha_{k-1}\cdot\beta_{k-1})^{1/4} - \left(\frac{\alpha_{k-1}+\beta_{k-1}}{2}\right)^{1/2}$$

en tenant compte du fait que :

$$\left(\alpha_{k-1}\cdot\beta_{k-1}\right)^{1/2} < \frac{\alpha_{k-1}+\beta_{k-1}}{2}$$

nous en concluons que $A - B < 0$

c.q.f.d.

Conclusion

Il est avantageux à priori de prendre un nombre de paramètres très grand ; en effet le théorème précédent le démontre facilement : l'inégalité $m\,M_{2^k} < (m\,M_{2^{k-1}})^2$ prouve que pour un même nombre d'opérations la valeur du rayon spectral est plus petite pour $t = 2^k$ que pour $t = 2^{k-1}$

On a considéré le cas où $a = \alpha_o = 10$, $b = \beta_o = 400$, ce qui correspond à peu près au problème modèle avec 9 points suivant o x et 9 points suivant oy, pour les valeurs de t : 1, 2, 4, 8 on a obtenu les résultats suivants :

N^{bre} des paramètres	N^{bre} de "pas"	N^{bre} d'itérations	Valeur de m M_t
t	NP	NI	μ_t
1	8	8	$60818162.\ 10^{-10}$
2	4	8	$445251.\ 10^{-10}$
4	2	8	$28015.\ 10^{-10}$
8	1	8	$7004.\ 10^{-10}$

le tableau suivant donne les valeurs des paramètres correspondants

t	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
1	63,245							
2	19,18	208,55						
4	12,079	34,222	116,88	331,13				
8	10,507	14,876	25,421	46,444	86,124	157,34	268,87	380,69

* Cas où le nombre t est quelconque.

Pour le cas où t n'est plus de la forme $t = 2^k$ ou plus généralement pour t quelconque Wachpress** en utilisant les fonctions elliptiques a pu démontrer que les r_j solutions du problème du min- max sont donnés par la relations suivante :

$$(84) \qquad r_j \simeq 2b \; \frac{(\frac{k'}{4})^{s_j}\left[1 + (\frac{k'}{4})^{2(1-s_j)}\right]}{1 + (\frac{k'}{4})^{2s_j}} \qquad \text{(une approximation)}$$

où $k' = a/b$

et $s_j = (2j-1)/2t \qquad j = 1, \ldots\ldots, t$

Les r_j par la formule (84) ne sont pas très différentes des r_j calculés par la méthode des suites de Fibonaci dans le cas de $t = 2^k$; nous avons calculé les r_j par les deux méthodes pour les cas t = 1,2,4,8 et nous avons constaté qu'on a au moins cinq chiffres décimaux exacts (voir tableau plus loin) ; Par ailleurs aucune modification n'intervient sur la convergence du

** Voir papiers de Wachpress 1963 [6]

fait qu'on utilise les r_j calculés par l'une ou l'autre des deux méthodes. On constate de plus que même dans le cas où t n'est pas de la forme $t = 2^k$, les paramètres r_j sont disposés équitablement de part et d'autre de $\sqrt{ab}$.

E. - Cas général.

E.1. Tous les calculs que nous avons faits jusqu'à présent ne sont valables que si les matrices H et V ont même base de vecteurs propres, ce qui se traduit d'après le théorème de Frobénius par la relation suivante :

$$HV = VH$$

Or il n'est pas facile de vérifier chaque fois si H et V possèdent cette propriété en essayant de calculer leurs valeurs et vecteurs propres. Nous allons voir dans quel cas nous avons HV = VH et par suite H et V ont même base de vecteurs propres, et par conséquent la méthode de Peaceman - Rachford est applicable avec tous les avantages qu'elle présente, et ceci en raisonnant directement sur l'équation différentielle en question et sur le domaine dans lequel nous essayons de la résoudre.

Considérons alors l'équation elliptique aux dérivées partielles de la forme générale, i.e.

$$(1) \quad -\frac{\partial}{\partial x}\left(P(x,y)\frac{\partial}{\partial x}U(x,y)\right) - \frac{\partial}{\partial y}\left(\varphi(x,y)\frac{\partial}{\partial y}U(x,y)\right) + \sigma(x,y) \cdot U(x,y) = S(x,y)$$

Soit R le domaine dans lequel nous cherchons une solution de l'équation (1) avec les conditions aux limites suivantes :

$$(2) \quad U(x,y) = \nu(x,y) \quad \text{pour tout } (x,y) \in \Gamma$$

Γ étant la frontière du domaine R.

Nous supposons d'autre part que $P(x,y)$, $\varphi(x,y)$ et $\sigma(x,y)$ continues à l'intérieur de R et sont telles :

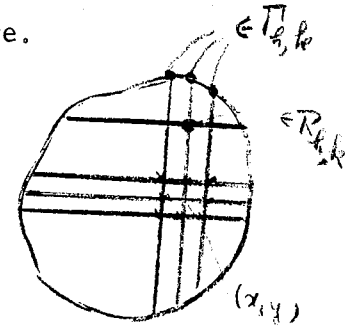
$$(3) \quad \left\{ \begin{array}{l} P(x,y) > 0 \\ \varphi(x,y) > 0 \\ \sigma(x,y) \geqslant 0 \end{array} \right\} \quad \text{dans R}$$

Nous pratiquons dans R un quadrillage de pas h suivant ox et de pas k suivant oy et nous désignons par $R_{h,k}$ l'ensemble des points du quadrillage qui sont à l'intérieur de R et par $\Gamma_{h,k}$ l'ensemble des points d'intersection de Γ avec les droites support du quadrillage (figure ci-dessous).

En approchant l'équation différentielle (1) par les différences centrales, nous aurons le système linéaire suivant :

$$AX = K \qquad (4)$$

X étant le vecteur des inconnues et K le vecteur second membre.



Si nous posons $A \equiv H_1 + V_1 + \Sigma$ il vient en explicitant H_1, V_1 et Σ et en désignant par $\left[H_1 U\right](x,y)$ la composante relative au point (x,y), du vecteur $H_1.U$:

$$h^2\left[H_1.U\right](x,y) = -P(x+\frac{h}{2},y)U(x+h,y) - P(x-\frac{h}{2},y)\,U(x-h,y) + \left[P(x+\frac{h}{2},y)+P(x-\frac{h}{2},y)\right] U(x,y) \qquad (5)$$

pour tout $(x,y) \in R_{h,k}$

$$k^2\left[V_1.U\right](x,y) = -\varphi(x,y+\frac{k}{2})U(x,y+k) - \varphi(x,y-\frac{k}{2})\,U(x,y-k) + \left[\varphi(x,y+\frac{k}{2})+\varphi(x,y-\frac{k}{2})\right] U(x,y) \qquad (6)$$

pour tout $(x,y) \in R_{h,k}$

$$\left[\Sigma.\,U\right](x,y) = \sigma(x,y).\,U(x,y) \qquad (7)$$

pour tout $(x,y) \in R_{h,k}$

En examinant(5) et (6), nous constatons que P (x,y) est évaluée en $x - \frac{h}{2}$ et $x + \frac{h}{2}$ et que φ (x,y) est évaluée en $y - \frac{k}{2}$ et $y + \frac{k}{2}$ tandis que U (x,y) est évaluée en $(x \pm h, y)$ et en $(x, y \pm k)$, ceci provient du fait suivant :

Considérons la quantité :

$$- \frac{\partial}{\partial x} \left[P(x,y) \quad \frac{\partial}{\partial x} \quad U(x,y) \right]$$

qui figure dans le membre de gauche de l'égalité (1) et cherchons une approximation de $\frac{\partial}{\partial x}$ U (x,y) au voisinage de (x,y), il vient :

$$\frac{\partial}{\partial x} U(x,y) \simeq \frac{U(x+\frac{h}{2}, y) - U(x - \frac{h}{2}, y)}{h}$$

et par conséquent :

$$P(x,y) \frac{\partial}{\partial x} U(x,y) \simeq P(x,y) \frac{U(x + \frac{h}{2}, y) - U(x - \frac{h}{2}, y)}{h}$$

et en cherchant une approximation de $\frac{\partial}{\partial x} \left[P(x,y) \frac{\partial}{\partial x} U(x,y) \right]$

il vient :

$$\frac{\partial}{\partial x} \left[P(x,y) \frac{\partial}{\partial x} U(x,y) \right] \simeq \frac{1}{h^2} \left\{ P(x + \frac{h}{2}, y) \left[U(x+h,y) - U(x,y) \right] - P(x - \frac{h}{2}, y) \left[U(x,y) - U(x-h,y) \right] \right\}$$

qui n'est autre que le second membre de (5) (à un facteur multiplicatif près) de la même façon, on calcule le second membre de (6).

Nous constatons aussi d'après ce qui précède que la composante de $\left[H_1 U \right]$ relative qu point (x,y) fait intervenir en plus du point (x,y) lui-même, les deux points situés horizontalement de part et d'autre de (x,y), c'est-à-dire (x+h,y) et (x-h,y) et que la composante de $\left[V_1 U \right]$ relative au point (x,y) fait intervenir en plus de ce dernier, les deux points situés verticalement de part et d'autre de (x,y), c'est-à-dire (x,y+h) et (x,y-k). Les points $(x \pm \frac{h}{2}, y)$, $(x, y \pm \frac{k}{2})$ n'interviennent que dans les fonctions P (x,y) et φ (x,y) (voir (5) et (6)) qui sont bien connues et, par exemple, la quantité $- P(x + \frac{h}{2}, y)$ serait le coefficient de l'inconnue U (x + h, y) dans le système linéaire.

Posons alors :

$$(8) \qquad \begin{aligned} H &= H_1 + \frac{1}{2}\Sigma \\ V &= V_1 + \frac{1}{2}\Sigma \end{aligned}$$

la relation $A \equiv H_1 + V_1 + \Sigma$ devient :

$$(9) \qquad A = H + V$$

où :

$$(10) \qquad \begin{aligned} h^2\left[HU\right](x,y) = &- P(x + \tfrac{h}{2},y)\, U(x+h,y) - P(x - \tfrac{h}{2},y)\, U(x-h,y) \\ &+\left[P(x+\tfrac{h}{2},y) + P(x-\tfrac{h}{2},y) + \tfrac{h^2}{2}\sigma(x,y)\right] U(x,y) \\ &\text{pour tout } (x,y) \in R_{h,k} \end{aligned}$$

et :

$$(11) \qquad \begin{aligned} k^2\left[VU\right](x,y) = &- \varphi(x,y+\tfrac{k}{2})\, U(x,y+k) - \varphi(x,y-\tfrac{k}{2})\, U(x,y-k) \\ &+\left[\varphi(x,y+\tfrac{k}{2}) + \varphi(x,y-\tfrac{k}{2}) + \tfrac{k^2}{2}\sigma(x,y)\right] U(x,y) \\ &\text{pour tout } (x,y) \in R_{h,k} \end{aligned}$$

E.2. Calcul direct de HV et de VH

Nous venons de constater que H appliquée à U (x,y) fait intervenir, en plus de ce dernier U (x-h,y) et U (x+h,y) et que V appliquée à U (x,y) fait intervenir, en plus de ce dernier U (x,y-k) et U (x,y+k), donc HV et VH appliquées à U (x,y) font intervenir, chacune à part, les neufs points suivants :

$$(12) \qquad \begin{array}{ccc} (x-h,y+k) & (x,y+k) & (x+h,y+k) \\ * & * & * \\ & (x,y) & \\ * & * & * \\ (x-h,y) & & (x+h,y) \\ (x-h,y-k) & (x,y-k) & (x+h,y-k) \\ * & * & * \end{array}$$

En calculant directement $\left[VH\right] U(x,y)$, nous obtenons une somme des 9 valeurs de (12) avec les coefficients suivants :

$$(12)_1 \quad \text{Coef. } U(x-h,y+k) = \frac{1}{k^2}\cdot\frac{1}{h^2} P\left(x-\frac{h}{2},y\right) \varphi\left(x-h,y+\frac{k}{2}\right)$$

$$(12)_2 \quad \text{Coef. } U(x,y+k) = -\frac{1}{k^2}\cdot\frac{1}{h^2} \varphi\left(x,y+\frac{k}{2}\right)\left[P\left(x+\frac{h}{2},y\right) + P\left(x-\frac{h}{2},y\right) + \frac{h^2}{2}\sigma(x,y)\right]$$

$$(12)_3 \quad \text{Coef. } U(x+h,y+k) = \frac{1}{k^2}\cdot\frac{1}{h^2} P\left(x+\frac{h}{2},y\right) \varphi\left(x,y+\frac{k}{2}\right)$$

$$(12)_4 \quad \text{Coef. } U(x-h,y) = -\frac{1}{k^2}\cdot\frac{1}{h^2} P\left(x-\frac{h}{2},y\right)\left[\varphi\left(x-h,y+\frac{k}{2}\right)+\varphi\left(x-h,y-\frac{k}{2}\right) + \frac{k^2}{2}\sigma(x-h,y)\right]$$

$$(12)_5 \quad \text{Coef. } U(x,y) = \frac{1}{k^2}\frac{1}{h^2}\left[P\left(x+\frac{h}{2},y\right)+P\left(x-\frac{h}{2},y\right) + \frac{h^2}{2}\sigma(x,y)\right] * \left[\varphi\left(x,y+\frac{h}{2}\right)+\varphi\left(x,y-\frac{k}{2}\right) + \frac{k^2}{2}\sigma(x,y)\right]$$

$$(12)_6 \quad \text{Coef. } U(x+h,y) = -\frac{1}{k^2}\frac{1}{h^2} P\left(x+\frac{h}{2},y\right)\left[\varphi\left(x+h,y+\frac{k}{2}\right)+\varphi\left(x+\frac{h}{2},y-\frac{k}{2}\right) + \frac{k^2}{2}\sigma(x+h,y)\right]$$

$$(12)_7 \quad \text{Coef. } U(x-h,y-k) = -\frac{1}{k^2}\frac{1}{h^2} P\left(x-\frac{h}{2},y\right) \varphi\left(x-h,y-\frac{k}{2}\right)$$

$$(12)_8 \quad \text{Coef. } U(x,y-k) = -\frac{1}{k^2}\frac{1}{h^2} \varphi\left(x,y-\frac{k}{2}\right)\left[P\left(x+\frac{h}{2},y\right) + P\left(x-\frac{h}{2},y\right) + \frac{h^2}{2}\sigma(x,y)\right]$$

$$(12)_9 \quad \text{Coef. } U(x+h,y-k) = \frac{1}{k^2}\cdot\frac{1}{h^2} P\left(x+\frac{h}{2},y\right) \varphi\left(x+h,y-\frac{k}{2}\right)$$

Et en calculant directement $\left[HV\right]\left[U\ (x,y)\right]$ nous obtenons une autre somme des 9 valeurs de (12) avec les coefficients suivants :

$$(13)_1 \quad \text{Coef. } U\ (x-h,y+k) = \frac{1}{k^2} \cdot \frac{1}{h^2}\ \varphi\ (x,y+\frac{k}{2})\ P\ (x-\frac{h}{2},\ y+k)$$

$$(13)_2 \quad \text{Coef. } U\ (x,y+k) = -\frac{1}{k^2}\ \frac{1}{h^2}\ \varphi(x,y+\frac{k}{2})\Big[P\ (x+\frac{h}{2},y+k)+P(x-\frac{h}{2},y+k) + \frac{h^2}{2}\ \sigma\ (x,y+k)\Big]$$

$$(13)_3 \quad \text{Coef. } U\ (x+h,y+k) = \frac{1}{k^2} \cdot \frac{1}{h^2}\ \varphi(x,y+\frac{k}{2})\ P\ (x+\frac{h}{2},\ y+k)$$

$$(13)_4 \quad \text{Coef. } U\ (x-h,y) = -\frac{1}{k^2}\ \frac{1}{h^2}\ P\ (x-\frac{h}{2},y)\Big[\varphi(x,y+\frac{k}{2})+\varphi\ (x,y-\frac{k}{2}) + \frac{k^2}{2}\ \sigma\ (x,y)\Big]$$

$$(13)_5 \quad \text{Coef. } U\ (x,y) = -\frac{1}{k^2} \cdot \frac{1}{h^2}\Big[\varphi(x,y+\frac{k}{2}) + \varphi(x,y-\frac{k}{2})+\frac{k^2}{2}\sigma(x,y)\Big] \cdot \Big[P(x+\frac{h}{2},y)+P(x-\frac{h}{2},y)+\frac{h^2}{2}\sigma(x,y)\Big]$$

$$(13)_6 \quad \text{Coef. } U\ (x+h,y) = -\frac{1}{k^2}\ \frac{1}{h^2}\ P(x+\frac{h}{2},y)\Big[\varphi(x,y+\frac{k}{2})+\varphi(x,y-\frac{k}{2}) + \frac{k^2}{2}\ \sigma\ (x,y)\Big]$$

$$(13)_7 \quad \text{Coef. } U(x-h,y-k) = \frac{1}{k^2}\ \frac{1}{h^2}\ \varphi(x,y-\frac{k}{2})\ P(x-\frac{h}{2},y-k)$$

$$(13)_8 \quad \text{Coef. } U(x,y-k) = -\frac{1}{k^2}\ \frac{1}{h^2}\ \varphi(x,y-\frac{k}{2})\Big[P(x+\frac{h}{2},\ y-k)+P(x-\frac{h}{2},y-k) + \frac{h^2}{2}\ \sigma\ (x,y-k)\Big]$$

$$(13)_9 \quad \text{Coef. } U\ (x+h,y-k) = \frac{1}{k^2}\ \frac{1}{h^2}\ P\ (x+\frac{h}{2},y-h)\ \varphi\ (x,y-\frac{k}{2})$$

E.3. Lemme :

"Si H et V commutent alors la fonction P (x,y) est une fonction de x seule et la fonction φ (x,y) est une fonction de y seule".

En effet : si HV = VH les termes correspondants dans l'expression de $\left[HV\right]$ U (x,y) et dans celle de $\left[VH\right]\left[U\ (x,y)\right]$ sont égaux ; écrivons alors que les coefficients de U (x + h, y + k) dans l'expression de $\left[VH\right]\left[U\ (x,y)\right]$ et dans celle de $\left[H\ V\right]\ \left[U\ (x,y)\right]$ sont égaux, nous aurons :

$$(14)\quad P\ (x+\frac{h}{2},y)\ \varphi(x+h,y+\frac{k}{2}) = P\ (x+\frac{h}{2},\ y+k)\ \varphi\ (x,y+\frac{k}{2}) \qquad \forall (x,y) \in R_{h,k}$$

Egalons de même les coefficients de U (x-h,y+k) :

$$(15)\quad P\ (x-\frac{h}{2},y)\ \varphi\ (x-h,y+\frac{k}{2}) = P\ (x-\frac{h}{2},y+k)\ \varphi(x,y+\frac{k}{2}) \qquad \forall (x,y) \in R_{h,k}$$

Remplaçons dans (15) x par (x + h), il vient :

$$(15)'\quad P(x+\frac{h}{2},y)\ \varphi(x,y+\frac{k}{2}) = P\ (x+\frac{h}{2},y+k)\ \varphi\ (x+h,y+\frac{k}{2}) \qquad \forall (x,y) \in R_{h,k}$$

en combinant (14) et (15)' il vient

$$(16)\quad \left[\varphi(x+h,y+\frac{k}{2})\right]^2 = \left[\varphi\ (x,y+\frac{k}{2})\right]^2 \quad \forall (x,y) \in R_{h,k}$$

et comme $\varphi\ (x,y) > 0$ alors :

$$(16)'\quad \varphi(x+h,y+\frac{k}{2}) = \varphi\ (x,y+\frac{k}{2}) \qquad \forall (x,y) \in R_{h,k}$$

la relation (16)' montre que $\varphi\ (x,y)$ est indépendante de x, donc $\underline{\varphi(x,y)=g(y)}$

En combinant (14) et (16)' nous aurons :

$$(17)\quad P(x+\frac{h}{2},y) = P\ (x+\frac{h}{2},\ y+k) \quad \forall\ (x,y) \in R_{h,k}$$

Cette égalité montre que P (x,y) est indépendante de la variable y, donc :

$$\underline{P\ (x,y) = f\ (x)}$$

c.q.f.d.

Remarque :

Nous constatons qu'avec le lemme E.3. vérifié, les coefficients de U (x-h, y+k), de U (x+h,y+k), de U (x-h, y-k) et de U(x+h,y-k) dans l'expression de VH sont égaux à leurs homologues dans l'expression de HV. Quant au coefficient de U (x,y) il est le même dans les deux expressions.

E.4. Lemme :

Si H et V commutent, alors la matrice Σ est une matrice multiple de l'unité : $\Sigma = cI$ où c est un scalaire non-négatif.

En effet : Ecrivons que $(12)_2 = (13)_2$, et en tenant compte du lemme E.3. il vient :

(18) $\sigma (x,y) = \sigma (x,y + k)$

Donc $\sigma (x,y)$ est indépendante de y, écrivons maintenant que : $(12)_6 = (13)_6$, et en tenant compte du lemme E.3. nous aurons :

(19) $\sigma (x,y) = \sigma (x + h,y)$

donc : $\sigma (x,y)$ est indépendante de x.

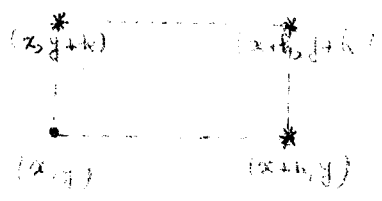
En combinant (19) et (18), nous en tirons que $\sigma (x,y)$ est une constante, et en tenant compte des conditions (3) nous aurons :

$$\sigma = c^{te} \geqslant 0 \qquad \text{c.q.f.d.}$$

E.5. Lemme :

"Si H et V commutent et si trois points du quadrillage situés sur les sommets d'un rectangle élémentaire (h,k) appartiennent à $R_{h,k}$ alors le quatrième sommet appartient aussi à R_{hk}.

En effet : Supposons que les points (x, y+k), (x+h,y+k) et (x+h,y) appartiennent à $R_{h,k}$ et que (x,y) n'appartient pas à R_{hk} mais à Γ_{hk}.



Ici, aussi, nous allons procéder à un raisonnement direct en appliquant simultanément les matrices HV et VH au point (x,y + k) :

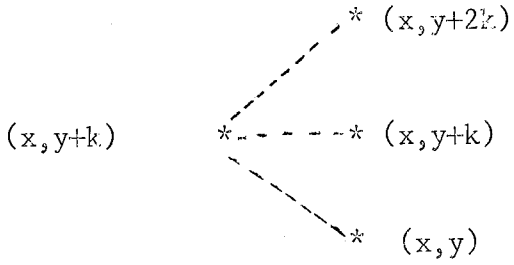
a) $\left[VU\right]$ (x,y+k)

contient des termes

en (x,y + 2k) $\in R_{hk}$,

en (x,y+k) $\in R_{hk}$, et

en (x,y) $\in \Gamma_{hk}$ (par hypothèse)



La quantité U (x,y) = ν (x,y) est bien connue et passe au second membre. Appliquons maintenant la matrice H à (VU) (x,y+k) c'est-à-dire H aux points U (x,y + 2k) et U(x,y+k) seulement :

$\left[HV\right]\left[U\ (x,y+k)\right]$ =

Contient des termes

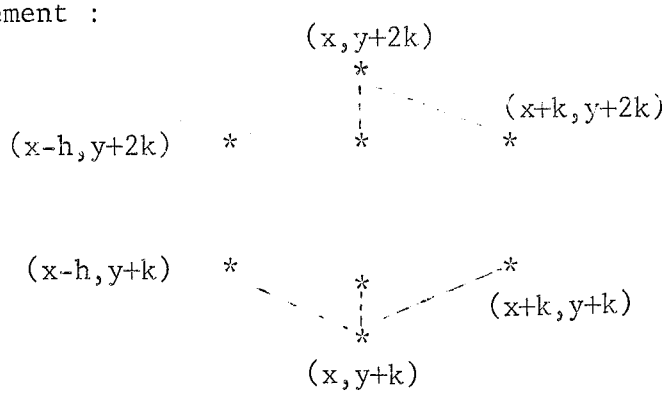
en (x-h,y+2k)

(x,y+2k)

(x+h,y+2k)

(x-h,y+k)

(x,y+k) et (x+h,y+2k) seulement

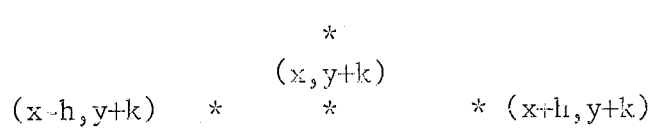


b) $\left[HU\right]$ (x,y+k) =

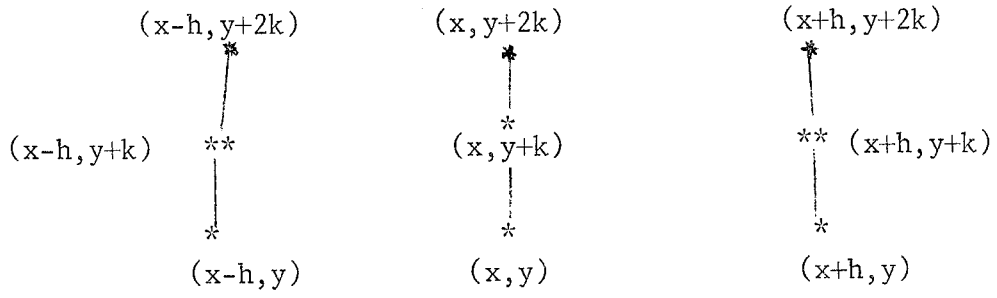
contient des termes

en (x,y+k) $\in R_{hk}$

en (x-h,y+k) $\in R_{hk}$ et en (x + h,y+k) $\in R_{hk}$



Donc pour appliquer V à $\left[HU\right]$ (x,y+k) il suffit d'appliquer V aux points (x - h, y+k), (x,y + k) et (x + h, y + k)



$\left[VHU\right]$ (x,y+k) contient des termes en (x-h,y+2k), (x,y+2k), (x+h,y+2k), (x-h,y+k), (x,y+k), (x+h,y+k), (x-h,y), (x,y) et (x+h,y).

Donc $\left[VHU\right]$(x,y+h) contient un terme en U (x+h,y) dont le coefficient est :

$$P(x + \frac{h}{2},\ y)\ \varphi\ (x, y + \frac{h}{2}) \neq 0$$

Mais dans l'expression de $\left[HVU\right]$ x,y+k) le coefficient de U(x+h,y) est nul, et alors les deux matrices H et V ne peuvent pas commuter, ce qui est une contradiction.

c.q.f.d.

E.5.1. Corollaire

Si H et V commutent alors la clôture convexe des points du quadrillage appartenant à R_{hk} est un rectangle dont les côtés sont parallèles aux axes des coordonnées.

N.B.

On appelle clôture convexe, le plus petit polynôme régulier qui contient ces points.

E.6. Théorème

"Les matrices H et V définies plus haut commutent si et seulement si :

- la clôture convexe des points du quadrillage R_{hk} est un rectangle dont les côtés sont parallèles aux axes des coordonnées.

- $\Sigma = cI$ où c est un scalaire non-négatif

- $P(x,y) = f(x)$

- $\varphi\ (x,y) = g\ (y)$."

F. - Cas d'un rectangle

F.1. Nous nous proposons de chercher une solution de l'équation de Laplace dans un rectangle R de côtés S et T, c'est-à-dire :

$$(19) \quad \Delta U = \frac{\partial^2}{\partial x^2} U(x,y) + \frac{\partial^2}{\partial y^2} U(x,y) = 0$$

avec :

$$(20) \quad U(x,y) = \nu(x,y) \text{ pour tout } (x,y) \in \Gamma$$

Γ étant la frontière de R.

Nous pratiquons dans R un quadrillage de pas h sur x et de pas k sur y : nous aurons ainsi m points suivant o y et n points suivant ox tels que :

$$S = (n+1)\,h$$
$$T = (m+1)\,k$$

Alors l'équation (20) peut-être approximée pour des h et k suffisamment petits par :

$$(21) \quad \frac{1}{h^2}\left\{U(x-h,y) - 2U(x,y) + U(x+h,y)\right\} + \frac{1}{k^2}\left\{U(x,y-h) - 2U(x,y) + U(x,y-h)\right\} = 0$$
$$\forall (x,y) \in R$$

Les mêmes remarques faites antérieurement concernant le "problème modèle" sont toujours valables pour le cas présent. Ainsi en écrivant la relation (21) pour tous les points du quadrillage nous obtenons un système linéaire de la forme :

$$(22) \quad A\,X = K$$

et en écrivant $A = H + V$ il est facile de vérifier que :

$$(23) \quad \left[H\,U\right](x,y) = \frac{1}{h^2}\left\{U(x-h,y) - 2U(x,y) + U(x+h,y)\right\}$$

$$(23)' \quad \left[V\,U\right](x,y) = \frac{1}{h^2}\left\{U(x,y-h) - 2U(x,y) + U(x,y+h)\right\}$$

$$\forall (x \pm h,\ y \pm h) \in R.$$

Tandis que dans le cas du "problème modèle" les deux matrices H et V avaient les mêmes valeurs propres et les mêmes vecteurs propres, dans le cas présent bien qu'elles aient les mêmes vecteurs propres elles ont des valeurs propres différentes.

Si nous ordonnons le vecteur inconnu X par rapport aux lignes du quadrillage et puis par rapport aux colonnes, c'est-à-dire si :

$$X_{\ell} = U_{ij}$$

i indice des lignes

j celui des colonnes

avec $\ell = (n-1)\, i+j$

les matrices H et V auront les formes suivantes :

$$(24)\quad H = \frac{1}{h^2} \begin{vmatrix} H_1 & & & & \\ & H_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & H_m \end{vmatrix}$$

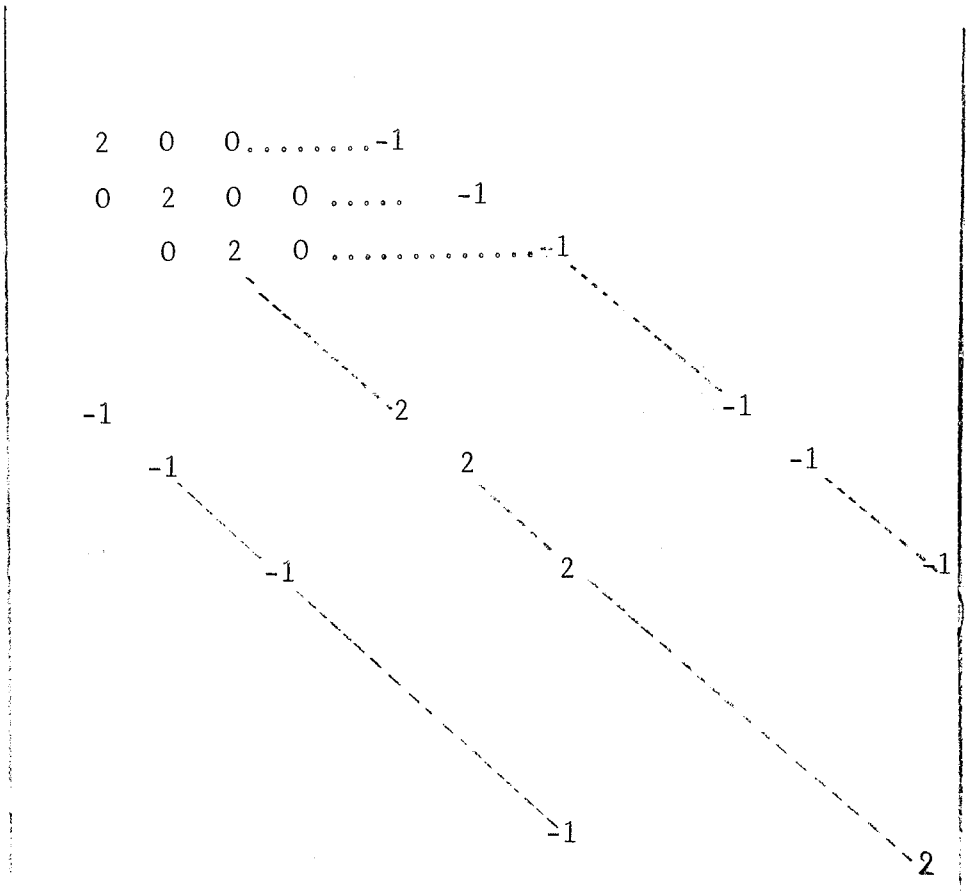
tous les H_i sont (n,n) et identiques

$$(24)'\quad H_i = \begin{vmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{vmatrix}$$

$$V = \frac{1}{k^2}$$

$$n + 1$$

(25)



Il est facile de vérifier, en procédant de la même sorte que pour le "problème modèle" qui si λ_j est une valeur propre de H et ν_j est une valeur propre de V,

$$\lambda_j = \frac{4}{h^2} \sin^2 \left(\frac{j\pi}{2(n+1)}\right)$$

et :

$$\nu_j = \frac{4}{k^2} \sin^2 \left(\frac{j\pi}{2(m+1)}\right)$$

Mais comme :

$$S = (n + 1)\, h$$

$$T = (m + 1)\, k$$

Alors :

(26)

$$\lambda_j = \frac{4}{h^2} \sin^2 \left(\frac{j\,\pi\,h}{2\,S}\right)$$

$$\nu_j = \frac{4}{k^2} \sin^2 \left(\frac{j\,\pi\,k}{2\,T}\right)$$

On vérifie aisément que les vecteurs propres de H et V sont de la forme :

$$(27) \quad X_{pq}(x,y) = \sin \frac{p \pi x}{S} \sin \frac{q \pi y}{T}$$

Remarque :

Généralement une matrice de la forme $\begin{vmatrix} C & -1 & & \\ -1 & C & -1 & \\ & \ddots & \ddots & \ddots \end{vmatrix}$ d'ordre (n,n)

a pour valeurs propres les quantités :

$$C - 2 \cos \left(\frac{j \pi}{n+1}\right)$$

Si C = 2, on retrouve bien les résultats antérieurs. Du fait que les matrices H et V n'aient plus les mêmes valeurs propres, on peut écrire :

$$a \leqslant \lambda_i \leqslant b$$

$$c \leqslant \nu_i \leqslant d$$

a, b, c, d, sont positifs et généralement différents les uns des autres.

F.2. Choix d'un seul ensemble de paramètres

Soit :

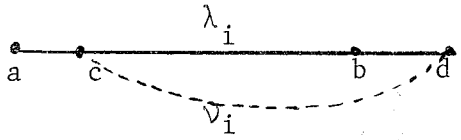
$$\alpha = \min (a,c)$$

et soit : $\beta = \max (b,d)$

alors : $\alpha \leqslant \lambda_i, \nu_i \leqslant \beta$

et ainsi pour le calcul de l'ensemble de t paramètres minimisant le maximum des valeurs propres de la matrice du Peaceman-Racheford, on procède de la même façon que pour le "modèle problème" en supposant que les bornes de λ et ν sont α et β.

On constate ici qu'on a élargi l'intervalle dans lequel varie λ et aus-



si celui dans lequel varie ν, il en résulte de ce fait que le rayon spectral de la matrice de P.R augmente et la méthode devient moins convergente, en effet, considérons le "problème modèle" et appliquons la méthode de Peaceman et Racheford avec un seul paramètre, alors, si a et b sont les bornes de λ et ν, le paramètre optimum est $\bar{r} = \sqrt{ab}$ et la borne de la valeur propre du rayon spectral est donnée par :

$$(28) \quad \rho(\mathcal{C}) \leq \left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} \right)^2$$

On vérifie facilement que si $\underline{b}$ augmente la quantité de droite de l'inégalité (28) augmente aussi, et si $\underline{a}$ diminue, la même quantité augmente aussi.

Remarque :

Il peut arriver que, $\underline{b}$ augmentant et $\underline{a}$ diminuant, le $\bar{r}$ optimum ne change pas de valeur, c'est-à-dire $\underline{ab}$ reste constant, mais ceci n'empêche pas la borne du rayon spectral de varier, car cette dernière dépend de $\left(\frac{a}{b}\right)$

F.3. Choix de deux ensembles de paramètres.

F.3.1. Pour faire varier λ et ν dans leurs intervalles respectifs, c'est-à-dire λ dans $[a,b]$ et ν dans $[c,d]$ nous allons utiliser deux séries de paramètres :

$$\left\{ \begin{bmatrix} R \\ H \end{bmatrix}^k \right\} \text{et} \left\{ \begin{bmatrix} R \\ V \end{bmatrix}^k \right\} \qquad k = 1, \ldots\ldots t$$

considérons les relations suivantes :

$$(29) \begin{cases} (r_k I + H) X^{(m+\frac{1}{2})} = (r_k I - V) X^{(m)} + K \\ (r_k I + V) X^{(m+1)} = (r_k I - H) X^{(m+\frac{1}{2})} + K \end{cases}$$

et au lieu de faire les deux sous-itérations de (29) avec le même paramètre r_k, nous ferons la première avec R_{H^k} qui sera pris entre a et b et la deuxième avec R_{V^k} qui sera pris entre c et d, et (29) s'écrira :

$$(30)\quad \begin{cases} (R_{H^k} I + H)\, X^{(m+\frac{1}{2})} = (R_{H^k} I - V)\, X^{(m)} + K \\ (R_{V^k} I + V)\, X^{(m+1)} = (R_{V^k} I - V)\, X^{(m+\frac{1}{2})} + K \\ \vdots \end{cases}$$

Et si k varie de 1 à t la nouvelle méthode de P.R. serait l'ensemble des relations (31) suivantes :

$$(31)\quad \begin{cases} (R_{H^1} I + H)\, X^{(m+\frac{1}{2})} = (R_{H^1} I - V)\, X^{(m)} + K \\ (R_{V^1} I + V)\, X^{(m+1)} = (R_{V^1} I - H)\, X^{(m+\frac{1}{2})} + K \\ \vdots \\ (R_{H^t} I + H)\, X^{(m+t-\frac{1}{2})} = (R_{H^t} I - V)\, X^{(m+t-1)} + K \\ (R_{V^t} I + V)\, X^{(m+t)} = (R_{V^t} I - H)\, X^{(m+t-\frac{1}{2})} + K \end{cases}$$

C'est-à-dire :

$$(32)\quad X^{(m+t)} = \mathcal{C}\, X^{(m)} + G\,(R_H, R_V, K)$$

avec :

$$(33)\quad \mathcal{C} = \prod_{k=1}^{t} \left[(R_{V^k} I + V)^{-1} (R_{V^k} I - H)\right] \cdot \left[(R_{H^k} I + H)^{-1} \cdot (R_{H^k} I - V)\right]$$

De sorte que, vu la forme de H et V, le rayon spectral de la matrice $\mathcal{C}$ est égal à μ, i. e.

$$(34)\quad \mu = \max_{i,j} \left| \prod_{k=1}^{t} \left(\frac{\lambda_i - R_V^k}{\lambda_i + R_H^k} \right) \left(\frac{\nu_j - R_H^k}{\nu_j + R_V^k} \right) \right|$$

$$\lambda_i \in [a,b] \quad \text{et} \quad \nu_j \in [c,d]$$

Si, au lieu de considérer que λ_n et ν_n prennent des valeurs en nombre fini respectivement dans $[a,b]$ et dans $[c,d]$, nous supposons λ_n (que nous désignons dorénavant par λ) et ν_n (que nous désignons dorénavant par ν) varient continûment l'une dans $[a,b]$ l'autre dans $[c,d]$ alors nous obtenons une borne du rayon spectral, cette borne étant le maximum de la fonction suivante :

$$(35)\quad M = \max_{\substack{\lambda \in [a,b] \\ \nu \in [c,d]}} \left| \prod_{k=1}^{t} \left(\frac{\lambda - R_V^k}{\lambda + R_H^k} \right) \left(\frac{\nu - R_H^k}{\nu + R_V^k} \right) \right|$$

et bien sûr il s'agit de minimiser M !

F.3.2. Choix des meilleurs paramètres $\left\{ R_V^k \right\}$ et $\left\{ R_H^k \right\}$?

a) Cas où $t = 2^n$?

a.1. Désignons par $M_t(\lambda, \nu, R)$ la quantité à minimiser son maximum, i. e.

$$(36)\quad M_t(\lambda, \nu, R) = \prod_{k=1}^{t} \left(\frac{\lambda - R_V^k}{\lambda + R_H^k} \right) \left(\frac{\nu - R_H^k}{\nu + R_V^k} \right)$$

avec $\lambda \in [a,b] \equiv [a_o, b_o]$ et $\nu \in [c,d] \equiv [c_o, d_o]$ et $a_o + c_o > 0$

Posons :

$$\lambda' = \lambda - \theta_o \qquad \lambda' \in [a'_o, b'_o]$$

$$\nu' = \nu + \theta_o \qquad \nu' \in [c'_o, d'_o]$$

Nous supposons en outre que :

$$(37) \quad a'_o \, b'_o = c'_o \, d'_o$$

d'où il vient :

$$(a_o - \theta_o)(b_o - \theta_o) = (c_o + \theta_o)(d_o + \theta_o)$$

d'où :

$$(38) \quad \theta_o = \frac{a_o \, b_o - c_o \, d_o}{a_o + b_o + c_o + d_o}$$

et :

$$a'_o = a_o - \theta_o$$

$$a'_o = (a_o^2 + a_o \, b_o + a_o \, c_o + a_o \, d_o - a_o \, b_o + c_o \, d_o) / (a_o + b_o + c_o + d_o)$$

$$a'_o = \left[a_o (a_o + c_o) + d_o (a_o + c_o)\right] / (a_o + b_o + c_o + d_o)$$

$$a'_o = (a_o + c_o)(a_o + d_o) / (a_o + b_o + c_o + d_o)$$

mais si $a_o + c_o > 0$

alors :

$$a_o + d_o > 0 \qquad \text{car } d_o > c_o$$

et :

$$a_o + b_o + c_o + d_o > 0 \quad \text{car } d_o > c_o \text{ et } b_o > a_o$$

donc finalement :

si $a_o + c_o > 0$ alors $\boxed{a'_o > 0}$

De même :

$$c'_o = (a_o \, c_o + b_o \, c_o + c_o^2 + c_o \, d_o + a_o \, b_o - c_o \, d_o) / (a_o + b_o + c_o + d_o)$$

$$c'_o = (a_o + c_o)(b_o + c_o) / (a_o + b_o + c_o + d_o)$$

si $a_o + c_o > 0$

alors $b_o + c_o > 0$ car, $b_o > a_o$

et $a_o + b_o + c_o + d_o > 0$ car $b_o > c_o$ et $d_o > c_o$

donc :

$$\underline{a_o + c_o > 0} \qquad \text{alors} \qquad \boxed{c'_o > 0}$$

Ceci dit, on définit deux nombres R'_{H^k} et R'_{V^k} par :

$$R'_{H^k} = R_{H^k} + \theta_o$$

$$R'_{V^k} = R_{V^k} - \theta_o$$

Alors la relation (36) devient :

$$(39) \quad M_t(\lambda', \nu', R') = \prod_{k=1}^{t} \left(\frac{\lambda' - R'_{V^k}}{\lambda' + R'_{H^k}} \right) \left(\frac{\nu' - R_{H^k}}{\nu' + R_{V^k}} \right)$$

$$\lambda' \in \left[a'_o \; b'_o \right]$$

$$\nu' \in \left[c'_o \; d'_o \right] \qquad a'_o \, b'_o = c'_o \, d'_o$$

a.2. Lemme :

Si $R'_{H^k} \in \left\{ R'_H \right\}$ et $R'_{V^k} \in \left\{ R'_V \right\}$ alors :

$$a'_o b'_o / R'_{H^k} \in \left\{ R'_H \right\} \text{ et } c'_o d'_o / R'_{V^k} \in \left\{ R'_V \right\}$$

En effet, soient :

$$y = a'_o \, b'_o / \lambda' \quad \text{et}$$

$$z = c'_o \, d'_o / \nu'$$

Multiplions le second membre de (39) haut et bas et pour chaque k par $\frac{y}{R'_{V^k}}$ $\frac{z}{R'_{H^k}}$, alors (39) devient :

$$(40)\quad M_t(\lambda',\nu',R') = \prod_{k=1}^{t}\left[\left(\frac{\lambda'-R'_{V^k}}{\lambda'+R'_{H^k}}\right)\cdot\left(\frac{\nu'-R'_{H^k}}{\nu'+R'_{V^k}}\right)\cdot\left(\frac{\frac{y}{R'_{V^k}} \quad \frac{z}{R'_{H^k}}}{\frac{y}{R'_{V^k}} \quad \frac{z}{R'_{H^k}}}\right)\right]$$

$$(41)\quad M_t(\lambda',\nu',R') = \prod_{k=1}^{t}\left(\frac{\frac{a'_o b'_o}{R'_{V^k}} - y}{\frac{a'_o b'_o}{R'_{H^k}} + y}\right)\left(\frac{\frac{c'_o d'_o}{R'_{H^k}} - z}{\frac{c'_o d'_o}{R'_{V^k}} + z}\right)$$

Puisque $a'_o\, b'_o = c'_o\, d'_o$ nous définissons alors :

$$R''_{V^k} = c'_o\, d'_o / R'_{V^k} \quad \text{et} \quad R''_{H^k} = a'_o\, b'_o / R'_{H^k}$$

d'où :

$$(42)\quad M_t(y,z,R'') = \prod_{k=1}^{t}\left(\frac{y - R''_{V^k}}{y + R''_{H^k}}\right)\left(\frac{z - R''_{H^k}}{z + R''_{V^k}}\right)$$

Nous constatons par ailleurs que quand :

λ' varie de a'_o à b'_o alors y variera de b'_o à a'_o

ν' varie de c'_o à d'_o alors z variera de d'_o à c'_o

Donc les intervalles de y et λ' étant confondus et ceux de z et ν' étant aussi confondus alors :

$$(43)\quad M_t(\lambda',\nu',R') = M_t(y,z,R'')$$

qui est une conséquence de l'unicité des paramètres $\{R'\} = \{R''\}$.

c.q.f.d.

a.3. Soient les trois variables :

$$\lambda^{(1)},\ \nu^{(1)} \text{ et } R^{(1)}$$

et posons :

$$(44)\quad M_t(\lambda', \nu', R') = M_{t/2}(\lambda^{(1)}, \nu^{(1)}, R^{(1)})$$

Comme R'_{V^k} et $c'_o d'_o / R'_{V^k}$ sont tous les deux dans $\left\langle \left[R'_V \right] \right\rangle$ alors (39)

devient :

$$(45)\quad M_t(\lambda',\nu',R') = \prod_{k=1}^{k/2} \frac{\left(\lambda'-R'_{V^k}\right)\left(\lambda'-c'_o d'_o/R'_{V^k}\right)}{\left(\lambda'+R'_{H^k}\right)\left(\lambda'+a'_o b'_o/R'_{H^k}\right)} * \frac{\left(\nu'-R'_{H^k}\right)\left(\nu'-a'_o b'_o/R'_{H^k}\right)}{\left(\nu'+R'_{V^k}\right)\left(\nu'+c'_o d'_o/R'_{V^k}\right)}$$

Mais :

$$(\lambda'-R'_{V^k})\left(\lambda'-\frac{c'_o d'_o}{R'_{V^k}}\right) = \lambda'^2 + c'_o d'_o - \lambda'\left(R'_{V^k} + \frac{c'_o d'_o}{R'_{V^k}}\right)$$

$$= 2\lambda'\left[\frac{1}{2}\left(\lambda' + \frac{a'_o b'_o}{\lambda'}\right) - \frac{1}{2}\left(R'_{V^k} + \frac{c'_o d'_o}{R'_{V^k}}\right)\right.$$

Posons :

$$(46)\quad \lambda^{(1)} \equiv \left(\lambda' + \frac{a'_o b'_o}{\lambda'}\right),\quad \nu^{(1)} \equiv \frac{1}{2}\left(\nu' + \frac{c'_o d'_o}{\nu'}\right)$$

$$R^{(1)}_{V^k} \equiv \frac{1}{2}\left(R'_{V^k} + \frac{c'_o d'_o}{R'_{V^k}}\right) \text{ et } R^{(1)}_{H^k} \equiv \frac{1}{2}\left(R'_{H^k} + \frac{a'_o b'_o}{R'_{H^k}}\right)$$

en combinant (39) et (46), on obtient (44) ; par ailleurs :

$$\lambda^{(1)} \in \left[\sqrt{a'_o b'_o}\ ,\quad \frac{a'_o+b'_o}{2}\right] \equiv \left[a_1,\ b_1\right]$$

$$\nu^{(1)} \in \left[\sqrt{c'_o d'_o}\ ,\quad \frac{c'_o+d'_o}{2}\right] \equiv \left[c_1,\ d_1\right]$$

$$R'_{V^{k1}} = R^{(1)}_{V^k} + \sqrt{\left[R^{(1)}_{V^k}\right]^2 - a'_o b'_o}$$

$$R'_{V^{k2}} = R^{(1)}_{V^k} - \sqrt{\left[R^{(1)}_{V^k}\right]^2 - a'_o b'_o} = a'_o\ b'_o / R^{(1)}_{V^{k1}}$$

$$R'_{H^{k1}} = R^{(1)}_{H^k} + \sqrt{\left[R^{(1)}_{H^k}\right]^2 - a'_o\ b'_o}$$

$$R'_{H^{k2}} = R_{H^k} - \sqrt{\left[R^{(1)}_{H^k}\right]^2 - a'_o\ b'_o} = a'_o\ b'_o / R^{(1)}_{H^{k1}}$$

et si $t = 2^n \Longrightarrow$

$$(47)\quad M_t\ (\lambda,\ \nu,\ R) = M_1\ (\lambda'^{(n)},\ \nu'^{(n)},\ R'^{(n)})$$

Donc partant de :

$$\lambda \in \left[a_o,\ b_o\right]$$

$$\nu \in \left[c_o,\ d_o\right]$$

Nous aurons :

$$\lambda' \in \left[a'_o,\ b'_o\right] = \left[a_o - \theta_o,\ b_o - \theta_o\right]$$

$$\nu' \in \left[c'_o,\ d'_o\right] = \left[c_o + \theta_o,\ d_o + \theta_o\right]$$

où :

$$\theta_o = \frac{a_o\ b_o - c_o\ d_o}{a_o+b_o + c_o+d_o}$$

Alors :

$$\lambda^{(1)} \in \left[a_1,\ b_1\right] = \left[\sqrt{a'_o\ b'_o}\ ,\ \frac{a'_o + b'_o}{2}\right]$$

$$\lambda'^{(1)} \in \left[a'_1, b'_1\right] = \left[a_1 - \theta_1,\ b_1 - \theta_1\right]$$

$$\nu^{(1)} \in \left[c_1, d_1\right] = \left[\sqrt{c'_o\ d'_o}\ ,\ \frac{c'_o + d'_o}{2}\right]$$

$$\nu'^{(1)} \in \left[c'_1, d'_1\right] = \left[c_1 + \theta_1,\ d_1 + \theta_1\right]$$

où :

$$\theta_1 = a_1\ (b_1 - d_1)\ /\ (2\ a_1 + b_1 + d_1)$$

et ainsi de suite :

$$\lambda^{(n)} \in \left[a_n,\ b_n\right] = \left[\sqrt{a'_{n-1}\ b'_{n-1}},\ \frac{a'_{n-1} + b'_{n-1}}{2}\right]$$

$$\lambda'^{(n)} \in \left[a'_n,\ b'_n\right] = \left[a_n - \theta_n\ ,\ b_n - \theta_n\right] \qquad (48)$$

$$\nu^{(n)} \in \left[c_n,\ d_n\right] = \left[\sqrt{c'_{n-1}\ d'_{n-1}},\ \frac{c'_{n-1} + d'_{n-1}}{2}\right]$$

$$\nu'^{(n)} \in \left[c'_n, d'_n\right] = \left[c_n + \theta_n,\ d_n + \theta_n\right]$$

où :

$$\theta_n = \frac{a_n\ (b_n - d_n)}{2\ a_n + b_n + c_n}$$

Notons que $a'_n\ b'_n = c'_n\ d'_n$ pour tout n

et que $a_n = c_n$ pour tout $n > 0$

De (47) Nous tirons :

$$R'^{(n)} = \sqrt{c'_n \, d'_n} = \sqrt{a'_n \, b'_n} = a_{n+1}$$

Ayant trouvé $R'^{(n)}$ l'algorithme à suivre pour le calcul des R_{H^k} et R_{V^k} est le suivant :

$$R'^{(n)} = a_{n+1}$$

$$R^{(n)}_H = R'^{(n)} - \theta_n = a_{n+1} - \theta_n$$

$$R^{(n)}_V = R'^{(n)} + \theta_n = a_{n+1} + \theta_n$$

$$R'^{(n-1)}_{H^1} = R^{(n)}_H + \sqrt{\left[R^{(n)}_H\right]^2 - a^2_n}$$

$$R'^{(n-1)}_{H^2} = a^2_n \; / \; R'^{(n-1)}_{H^1}$$

$$R'^{(n-1)}_{V^1} = R^{(n)}_V + \sqrt{\left[R^{(n)}_V\right]^2 - a^2_n}$$

$$R'^{(n-1)}_{V^2} = a^2_n \; / \; R'^{(n-1)}_{V^1}$$

$$R^{(n-1)}_{H^1} = R'^{(n-1)}_{H^1} - \theta_{n-1}$$

$$R^{(n-1)}_{H^2} = R'^{(n-1)}_{H^2} - \theta_{n-1}$$

$$R^{(n-1)}_{V^1} = R'^{(n-1)}_{V^1} + \theta_{n-1}$$

$$R^{(n-1)}_{V^2} = R'^{(n-1)}_{V^2} + \theta_{n-1}$$

$$R'^{(j-1)}_{H^{k1}} = R^{(j)}_{H^k} + \sqrt{\left[R^{(j)}_{H^k}\right]^2 - a^2_j}$$

$$R'^{(j-1)}_{H^{k2}} = a^2_j \,/\, R'^{(j-1)}_{H^{k1}}$$

$$R'^{(j-1)}_{V^{k1}} = R^{(j)}_{V^k} + \sqrt{\left[R^{(j)}_{V^k}\right]^2 - a^2_j}$$

$$R'^{(j-1)}_{V^{k2}} = a^2_j \,/\, R'^{(j-1)}_{V^{k1}}$$

$$R^{(j-1)}_{H^{k1}} = R'^{(j-1)}_{H^{k1}} - \theta_{j-1} \qquad (49)$$

$$R^{(j-1)}_{H^{k2}} = R'^{(j-1)}_{H^{k2}} - \theta_{j-1}$$

$$R^{(j-1)}_{V^{k1}} = R'^{(j-1)}_{V^{k1}} + \theta_{j-1}$$

$$R^{(j-1)}_{V^{k2}} = R'^{(j-1)}_{V^{k2}} + \theta_{j-1}$$

$$j = n,\ n-1,\ n-2,\ \ldots\ldots,\ 1$$

et les paramètres $R^{(o)}_{H^k}$ et $R^{(o)}_{V^k}$ ne sont autres que les paramètres recherchés et à utiliser dans les équation matricielles (30).

Alors le maximum de la fonction $M_t(\lambda, \nu, R)$ est donné par :

$$\mu_t = \max_{\substack{\lambda \in [a_o b_o] \\ \nu \in [c_o d_o]}} \left| M_t(\lambda, \nu, R^{(o)}) \right|$$

$$\mu_t = \max_{\substack{\lambda'^{(n)} \in [a'_n,\ b'_n] \\ \nu'^{(n)} \in [c'_n,\ d'_n]}} \left| M_1 \left(\lambda'^{(n)},\ \nu'^{(n)},\ R'^{(n)} \right) \right|$$

qui n'est autre que :

$$(50) \qquad \mu_t = \left(\frac{a_{n+1} - a'_n}{a_{n+1} + a'_n} \right) \left(\frac{a'_{n+1} - c'_n}{a'_{n+1} + c'_n} \right)$$

b. Cas où t est quelconque.

Dans l'équation (36), nous procédons au changement de variables suivant :

$$(51) \qquad \lambda = \frac{u - \alpha}{\beta - \delta u} \qquad \nu = \frac{v + \alpha}{\beta + \delta v}$$

où α, β, δ sont les constantes à déterminer ; ainsi l'équation (36) devient :

$$(52) \qquad M_t = \prod_{k=1}^{t} \left(\frac{u - R'_{V^k}}{u + R'_{H^k}} \right) \left(\frac{v - R'_{H^k}}{v + R'_{V^k}} \right)$$

où :

$$(52)' \begin{cases} R'_{H^k} = \dfrac{\beta R_{H^k} - \alpha}{1 - \delta R_{H^k}} \\ R'_{V^k} = \dfrac{\beta R_{V^k} + \alpha}{1 + \delta R_{V^k}} \end{cases}$$

Nous choisissons α, β, δ de telle façon que :

quand $\lambda = a$ alors $u = k'$

quand $\lambda = b$ alors $u = 1$

quand $\nu = c$ alors $v = k'$

quand $\nu = d$ alors $v = 1$

ceci est possible si k' est racine de l'équation :

$$(53) \quad k'^2 - 2(1 + m)k' + 1 = 0$$

où

$$(54) \quad m = 2(b-a)(d-c) / (a+c)(b+d)$$

et puisque $b > a$, $d > c$ et $a + c > 0$ alors : $m > 0$

Alors les racines de l'équation (53) sont réelles et positives et inverses l'un de l'autre ; nous considérons k' comme étant le plus petit de ces racines :

$$k' = \frac{1}{1+m+\sqrt{m(m+2)}}$$

$$0 < k' < 1$$

Ainsi, nous avons :

$$(56) \begin{cases} \delta = 2 \dfrac{k'(b+d) - (a+c)}{(a+c)(b-d)+k'(b+d)(c-a)} \\ \beta = \dfrac{2 + \delta(b-d)}{b+d} \\ \alpha = \dfrac{k'(c-a+2ac\,\delta)}{a+c} \end{cases}$$

les intervalles de u et v sont depuis identiques, c'est-à-dire u et v varient toutes les deux dans $[k',1]$. Dès lors si $R_{V^k} \in [R'_V]$ et $R'_{H^k} \in [R'_H]$ où $[R'_V]$ et $[R'_H]$ sont deux ensembles de paramètres résolvant le problème du minimax, nous pouvons conclure que $[R'_V] = [R'_H]$ du fait de l'unicité de la solution :

d'où : $R'_{H^k} = R'_{V^k} = R_k$

et :

$$M = P(u)\, P(v)$$

où :

$$P(u) = \prod_{k=1}^{t} \frac{u - R_j}{u + R_j}$$

Wachpress * a démontré que les R_j répondant au problème du minimax s'expriment au moyen des fonctions elliptiques :

$$(57) \qquad R_j = dn \frac{(2j - 1)\, K}{2t}$$

D'où en utilisant les relations (52), il vient :

$$(58) \begin{cases} R_{H^k} = \dfrac{R_k - \alpha}{\beta + \delta R_k} \\ R_{V^k} = \dfrac{R_k + \alpha}{\beta + \delta R_k} \end{cases}$$

et ainsi le problème du minimax est complètement résolu.

La relation (57) donne R_j; une approximation de R_j est donnée par :

$$(59) \qquad R_j \approx 2 \left(\frac{k'}{4}\right)^{r_j} \frac{\left[1 + \left(\frac{k'}{4}\right)^{2(1-r_j)}\right]}{1 + \left(\frac{k'}{4}\right)^{2 r_j}}$$

avec : $r_j = \dfrac{2j-1}{2t}$ $\qquad j = 1, \ldots\ldots, t$

les R_j sont situés de part et d'autre de $\sqrt{k'.1} = \sqrt{k'}$, donc ils sont tels que :

$$R_1 \cdot R_t = k'$$

ou encore :

$$R_j \cdot R_{t-j+1} = k'$$

les R_j étant ordonnés de la façon suivante :

$$R_1 < R_2 < \ldots\ldots R_j < \ldots\ldots < R_t$$

et le maximum absolu de la fonction étant :

$$(60) \quad \mu_t = 4 \; e^{\pi^2 \, t/\log (k'/4)}$$

en posant :

$$c = \frac{\pi^2 \, t}{\log (\frac{k'}{4})}$$

$$(60)' \quad \boxed{\mu_t = 4 \; e^c}$$

Une comparaison entre les paramètres calculés par la méthode exposée au paragraphe a) et ceux calculés par la méthode exposée au paragraphe b) montre que ces paramètres ont au moins cinq chiffres significatifs exacts.

III

ETUDE THEORIQUE *

DES ERREURS DANS LA METHODE

DE

PEACEMAN-RACHFORD

-=-=-=-=-=-=-=-=-=-=-

* Principales références : - Thèse de Monsieur GASTINEL (Bibliographie n°2)

- Thèse de Monsieur LIOT (Bibliographie n°4)

III

ETUDE THEORIQUE*

DES ERREURS DANS LA METHODE

DE

PEACEMAN-RACHFORD

I. - Notions préliminaires

a) Nous savons que cette méthode diffère des autres méthodes itératives par :

1) Le fait que le vecteur $X^{(m+1)}$ est obtenu à partir de $X^{(m)}$ par l'intermédiaire d'un autre vecteur $X^{(m+1/2)}$ de la façon suivante :

$$(1) \quad (rI + H)\, X^{(m+\frac{1}{2})} = (rI - V)\, X^{(m)} + S$$

$$(2) \quad (rI + V)\, X^{(m+1)} = (rI - H)\, X^{(m+\frac{1}{2})} + S$$

r étant un paramètre et en combinant (1) et (2), il vient :

$$(3) \quad X^{(m+1)} = (rI + V)^{-1}\,(rI-H)\,(rI+H)^{-1}\,(rI-V)\,X^{(m)} + (rI + V)^{-1}\left[(rI-H) + I\right] S$$

ou :

$$(3)' \quad X^{(m+1)} = M\,(r)\,X^{(m)} + N\,(r)$$

où M (r) est une matrice représentant le coefficient de $X^{(m)}$ et N (r) une matrice représentant la partie restante à droite de l'égalité (3).

---

* Principales références : - Thèse de Monsieur GASTINEL (Bibliographie n°2)
- Thèse de Monsieur LIOT (bibliographie n°4)

2) Le fait que le vecteur $X^{(m+2)}$ est obtenu à partir de $X^{(m+1)}$ de la même manière que précédemment au seul changement de prendre un autre paramètre r ; donc si pour calculer $X^{(m+1)}$ on prend $r = r_1$ alors pour calculer $X^{(m+2)}$ on prend $r = r_2$, et ainsi de suite, de sorte que nous aurons les relations suivantes :

$$(4)\quad \begin{aligned} X^{(m+1)} &= M(r_1)\, X^{(m)} + N(r_1)\, S \\ X^{(m+2)} &= M(r_2)\, X^{(m-1)} + N(r_2)\, S \\ &\vdots \\ X^{(m+k)} &= M(r_k)\, X^{(m+k-1)} + N(r_k)\, S \end{aligned}$$

3) Le fait que ayant calculer $X^{(m+k)}$, de grouper les k équations de (4) en une seule et de recommencer le calcul en prenant $X^{(m)} = X^{(m+k)}$, etc ....

d'où :

$$(5)\quad X^{(m+k)} = \mathcal{M}\, X^{(m)} + \mathcal{N}$$

en posant :

$$Y_o = X^{(m)}$$

et

$$Y_1 = X^{(m+k)}$$

il vient :

$$(5)'\quad Y_1 = \mathcal{M}\, Y_o + \mathcal{N}$$

la méthode n'est alors qu'une méthode linéaire obéissant à la relation suivante :

$$(6)\quad Y_{i+1} = \mathcal{M}\, Y_i + \mathcal{N}$$

d'où :

$$(7)^*\begin{cases} Y_i = \mathcal{M}^i\, Y_o + \sum_{j=o}^{i-1} \mathcal{M}^j\, \mathcal{N} \\ Y_i = \mathcal{M}^{i-\alpha}\, Y_\alpha + \sum_{j=o}^{i-\alpha-1} \mathcal{M}^j\, \mathcal{N} \end{cases}$$

* L'écriture $\mathcal{M}^i$ veut dire $\mathcal{M}$ puissance i.

b) Considérons d'abord une des relations (4), i.e.

$$X^{(m+i)} = M(r_i)\, X^{(m+i-1)} + N(r_i)\, S$$

théoriquement le calcul des vecteurs itérés X n'est convergent que si le rayon spectral de la matrice $M(r_i)$ est inférieur à 1 ; il en est de même pour le calcul des Y_i qui n'est convergent que si le rayon spectral de la matrice $\mathcal{M}$ est inférieur à 1.

$$\rho(M(r_i)) < 1$$

$$\rho(\mathcal{M}) < 1$$

Donc la stabilité de la méthode, d'ailleurs ceci a été dit antérieurement, dépend de $\rho(\mathcal{M})$ et par conséquent des $\rho(M(r_i))$; cette stabilité dite théorique, il en a été fait pour qu'elle soit assurée.

Mais dans la pratique, les choses se passent un peu différemment et la méthode pourrait diverger même si $\rho(\mathcal{M}) < 1$, ceci étant dû à la propagation des erreurs de calcul. En effet, toute machine a une capacité limitée, ceci veut dire qu'à chaque opération élémentaire (addition, soustraction, multiplication et division) on commet une erreur de calcul (dite d'arrondi) c'est-à-dire si la capacité de la machine est de 8 décimales et si on a un résultat de la forme $1210043789 \cdot 10^{\alpha}$ par exemple, la mémoire dans laquelle se trouve ce nombre n'en conserve que $12100437 \cdot 10^{\alpha+2}$ et le calcul se fait de nouveau avec ce nombre arrondi, et ainsi de suite de sorte qu'en partant d'un vecteur $X^{(m)}$ j'aboutis à un vecteur $\tilde{X}^{(m+1/2)}$ et puis à un autre vecteur $\tilde{X}^{(m+1)}$ au lieu de $X^{(m+1/2)}$ et $X^{(m+1)}$.

Si $X^{(m+1)}$ vérifie la relation matricielle

$$X^{(m+1)} = A\, X^{(m)} + B$$

alors $\tilde{X}^{(m+1)}$ vérifiera la relation matricielle

$$\tilde{X}^{(m+1)} = A'\, X^{(m)} + B'$$

Théoriquement, par un choix convenable des paramètres r_i, nous avons fait de sorte que la matrice A ait un rayon spectral inférieur à l'unité : $\rho(A) < 1$; mais, à priori, rien ne permet de dire que $\rho(A) < 1$. Au contraire, si $\rho(A) < 1$ tout en étant très voisin de 1 ($\rho(A) \neq 1$), il ne faut pas s'étonner si la méthode diverge, divergence due, bien sûre, à la propagation des erreurs de calcul ; ceci correspondrait symboliquement à ce que $\rho(A) > 1$.

Dans tout ce qui va suivre nous allons étudier la propagation de ces erreurs et leur influence sur les matrices $M(r_i)$ et par conséquent sur la matrice $\mathcal{M}$ et trouver en définitive la relation matricielle à laquelle répond le vecteur Y trouvé. Il est bien évident que la propagation des erreurs dépend en premier lieu de la méthode choisie pour la résolution du système linéaire (nous nous bornerons à la méthode de Gauss) et aussi du fait qu'on travaille en virgule flottante ou en virgule fixe (nous considérons seulement le calcul en flottant)

II. - Calcul en flottant.

a) Résolution du système AX = Y

où :

$$(8) \qquad A = \begin{vmatrix} \alpha & 1 & & & \\ 1 & \alpha & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & \alpha & 1 \\ & & & 1 & \alpha \end{vmatrix} \qquad \text{et} \quad Y = \begin{vmatrix} Y_1 \\ \vdots \\ Y_n \end{vmatrix}$$

La méthode de Gauss* consiste en le passage de $A^{(1)} \equiv A$ à $A^{(2)}$, puis de $A^{(2)}$ à $A^{(3)}$, ... puis de $A^{(k)}$ à $A^{(k+1)}$, ... et finalement de $A^{(n-1)}$ à $A^{(n)}$, qui est une matrice triangulaire supérieur, par l'algorithme général suivant :

---

* Cours de Monsieur GASTINEL (Bibliographie n° 7)

$$a^{(K+1)}_{ij} = a^{(K)}_{ij} - \frac{a^{(K)}_{iK}}{a^{(K)}_{KK}} \; a^{(K)}_{Kj}$$

$$K = 1, \ldots\ldots, n-1$$
$$i = K+1, \ldots, n$$
$$j = K+1, \ldots, n+1$$

le terme $a^{(K+1)}_{ij}$ désignant l'élément (i,j) de la matrice $A^{(K+1)}$.

Dans le cas présent, où A est donnée par (8), pour passer de $A^{(K)}$ à $A^{(K+1)}$ il suffit de transformer dans la $(K+1)^{ème}$ ligne de $A^{(K)}$ les éléments suivants :

$(K+1, K)$ qui devient nul.

$(K+1, K+1)$ qui varie

$(K+1, K+2)$ qui reste égal à 1

$(K+1, n+1)$ qui varie

donc seuls sont transformés les éléments (K+1, K+1) et (K+1, n+1) (second membre) ; et à la fin de la trinagulation nous aurons une équation matricielle de la forme :

$$(9) \qquad A_T \, X = y$$

où :

$$(9)' \qquad A_T = A^{(n)} = \begin{bmatrix} d_1 & 1 & 0 & & \\ & d_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \\ 0 & & & & d_n \end{bmatrix} \qquad \text{et } y = \begin{vmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{vmatrix}$$

ceci fait, le calcul des composantes du vecteur X se fait par simple élimination en commençant par :

$$x_n = \frac{y_n}{d_n}$$

et

$$x_{i-1} = \frac{y_{i-1} - x_i}{d_{i-1}} \qquad \text{pour i allant de n à 2}$$

Reprenons le système

(10) $AX = Y$ où A est une matrice quelconque et soient :

$\varepsilon^{(K)}_{K+p}$ l'erreur relative dans le calcul de

$$\bar{\rho}^{(K)}_{K+p} = \frac{\bar{a}^{(K)}_{K+p,K}}{\bar{a}^{(K)}_{K,K}}$$

$\eta^{(K)}_{i,j}$ l'erreur relative dans le calcul de

$$\bar{a}^{(K+1)}_{i,j} = \bar{a}^{(K)}_{i,j} - \bar{\rho}^{(K)}_{i} \cdot \bar{a}^{(K)}_{K,j}$$

$\rho^{(K)}_{i}$ l'erreur relative dans le calcul de

$$\bar{y}^{(K+1)}_{i} = \bar{y}^{(K)}_{i} - \bar{\rho}^{(K)}_{i} \cdot \bar{y}^{(K)}_{(K)}$$

$\mu_{i,K}$ l'erreur relative dans le calcul de

$$\bar{S}_{i,K} = \bar{S}_{i,K+1} + \bar{a}^{(i)}_{i,K} \cdot \bar{\xi}_{K}$$

γ_i l'erreur relative dans le calcul de

$$\bar{\xi}_i = \frac{\bar{S}_{i,i+1}}{\bar{a}^{(i)}_{i,i}}$$

où $\bar{S}_{i,K}$ représente la somme partielle dans le retour arrière avec $S_{i,n+1} = 0$ et où l'écriture $\bar{a}$, par exemple, désigne la valeur calculée de a et non sa vraie valeur.

Le vecteur $\bar{\xi}$ $\begin{vmatrix} \bar{\xi}_1 \\ \vdots \\ \bar{\xi}_n \end{vmatrix}$ réellement obtenu peut-être considéré comme

la solution exacte du système

$$(11)^* \quad (A + \delta A)\, X = Y + \delta Y$$

avec :

$$(12)^* \begin{cases} \delta A = \sum_{K=1}^{n-1} \left(-C^{(K)} + \delta_1 A^{(K)}\right) \\ \delta Y = \left(\sum_{K=1}^{n-1} \delta_1 Y^{(K)}\right) + \delta_2 Y + \delta_3 Y \end{cases}$$

où :

$$(13)^* \quad C^{(K)} = \begin{bmatrix} & & & K \downarrow & & \\ & & & 0 & & \\ & & 0 & \vdots & 0 & \\ 0 & \cdots & 0 & \cdots & 0 & \\ & & \epsilon_{K+1}^{(K)} \cdot \bar{a}_{K+1,K}^{(K)} & & & \\ & & \epsilon_{K+2}^{(K)} \cdot \bar{a}_{K+2,K}^{(K)} & & & \\ & & \vdots & & & \\ 0 & & \vdots & & 0 & \\ & & \epsilon_n^{(K)} \cdot \bar{a}_{n,K}^{(K)} & & & \end{bmatrix} \quad ; \quad \delta_1 A^{(K)} = \begin{bmatrix} & & 0 & & \\ & & \vdots & & \\ 0 & \cdots & 0 & \cdots & 0 \\ & & \vdots & & \\ & & \eta_{i,j}^{(K)} \cdot \bar{a}_{i,j}^{(K+1)} & & \\ & & \vdots & & \\ & & 0 & & \end{bmatrix}$$

* Thèse de Monsieur GASTINEL 1960 [2]

$$(14)^* \qquad \delta_1 Y^{(K)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ p_{K+1}^{(K)} \cdot \bar{y}_{K+1}^{(K+1)} \\ p_{K+2}^{(K)} \cdot \bar{y}_{K+2}^{(K+1)} \\ \vdots \\ p_{n}^{(K)} \cdot \bar{y}_{n}^{(K+1)} \end{bmatrix} \qquad \delta_2 Y = \begin{bmatrix} \gamma_1 \cdot a_{1,1}^{(1)} \cdot \bar{\xi}_1 \\ \gamma_2 \cdot \bar{a}_{2,2}^{(2)} \cdot \bar{\xi}_2 \\ \vdots \\ \gamma_n \cdot \bar{a}_{n,n}^{(n)} \cdot \bar{\xi}_n \end{bmatrix}$$

$$(15)^* \qquad \delta_3 Y = \begin{vmatrix} \mu_{1,2} \cdot \bar{S}_{1,2} + \mu_{1,3} \cdot \bar{S}_{1,3} + \ldots\ldots + \mu_{1,n} \cdot \bar{S}_{1,n} \\ \mu_{2,1} \cdot \bar{S}_{2,1} + \mu_{2,3} \cdot \bar{S}_{2,3} + \ldots\ldots + \mu_{2,n} \cdot \bar{S}_{2,n} \\ \vdots \\ \mu_{n-1,1} \cdot \bar{S}_{n-1,1} + \ldots\ldots + \mu_{n-1,n-2} \cdot \bar{S}_{n-1,n-2} + \mu_{n-1,n} \cdot \bar{S}_{n-1,n} \end{vmatrix}$$

où $(\delta_3 Y)_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \mu_{i,j} \cdot \bar{S}_{i,j}$ pour i allant de 1 à n -1

En tenant compte du fait qu'à chaque pas en K on ne transforme dans la $(K+1)^{ème}$ ligne que les deux éléments (K +1, K+1) et (K+1,n+1) et en remarquant de plus qu'on ne calcule pas successivement :

$$\bar{\rho}_{K+1}^{(K)} = \frac{1}{\bar{a}_{K,K}^{(K)}} \qquad \text{puis } \bar{a}_{K+1,j}^{(K+1)} = \bar{a}_{K+1,j}^{(K+1)} = \bar{a}_{K+1,j}^{(K)} - \bar{\rho}_{K+1}^{(K)} \cdot \bar{a}_{K,j}^{(K)}$$

mais directement*

$$\bar{a}^{(K+1)}_{K+1,j} = \bar{a}^{(K)}_{K+1,j} - \frac{\bar{a}^{(K)}_{K,j}}{\bar{a}^{(K)}_{K,K}} + \text{erreur}$$

On en conclue que les seules erreurs non nulles sont :

$$\rho^{(K)}_{K+1} \; , \quad \eta^{(K)}_{K+1,K+1} \quad \text{et} \quad \gamma_i$$

en posant $\bar{\lambda}_K = \bar{a}^{(K)}_{K,K}$ et en reportant dans (13), (14) et (15) il vient :

$$C^{(K)} = 0$$

$$\delta_1 A^{(K)} = \eta^{(K)}_{K+1,K+1} \cdot \bar{a}^{(K+1)}_{K+1,K+1} \cdot E_{K+1,K+1}$$

$$\delta_1 Y^{(K)} = \rho^{(K)}_{K+1} \cdot \bar{y}^{(K+1)}_{K+1} \cdot E_{K+1,K+1}$$

$$\delta_3 Y = 0$$

$$\delta_2 Y = \begin{vmatrix} \gamma_1 \cdot \bar{\lambda}_1 \cdot \bar{\xi}_1 \\ \gamma_2 \cdot \bar{\lambda}_2 \cdot \bar{\xi}_2 \\ \vdots \\ \gamma_n \cdot \bar{\lambda}_n \cdot \bar{\xi}_n \end{vmatrix}$$

où $E_{i,j}$ est une matrice de base, c'est-à-dire une matrice ayant tous ses éléments nuls sauf l'élément (i,j) qui est égal à 1.

D'où finalement, si on pose $\eta_i = \eta^{(i)}_{i+1,i+1}$, il vient :

$$\delta A = \begin{vmatrix} 0 & & & \\ & \eta_2 \cdot \bar{\lambda}_2 & & \\ & & \eta_3 \cdot \bar{\lambda}_3 & \\ & & & \ddots \\ & & & \eta_n \cdot \bar{\lambda}_n \end{vmatrix} \qquad (16)$$

*Thèse de Mr LIOTARD Grenoble 64

$$(17)\quad \delta Y = \begin{vmatrix} 0 & & \\ p_2^{(1)} \cdot \bar{y}_2^{(2)} & & \\ p_3^{(2)} \cdot \bar{y}_3^{(3)} & & \\ \vdots & & \\ p_n^{(n-1)} \cdot \bar{y}_n^{(n)} & & \end{vmatrix} + \begin{vmatrix} \gamma_1 \cdot \bar{\lambda}_1 \cdot \bar{\xi}_1 \\ \gamma_2 \cdot \bar{\lambda}_2 \cdot \bar{\xi}_2 \\ \vdots \\ \gamma_n \cdot \bar{\lambda}_n \cdot \bar{\xi}_n \end{vmatrix}$$

Considérons maintenant le système suivant

$$(18)\quad \tilde{A}\, X = \tilde{Y} \qquad \text{avec } A\ (\ell, \ell) \qquad \ell = n \cdot n$$

où :

$$\tilde{A} = \begin{vmatrix} A_1 & & \\ & A_2 & \\ & & \ddots \\ & & & A_n \end{vmatrix}$$

tous les A_i étant identiques et de la forme

$$A_i = \begin{vmatrix} \alpha & 1 & & \\ 1 & \alpha & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & \alpha \end{vmatrix}$$

la solution réellement obtenue <u>peut-être</u> considérée comme solution exacte du système :

$$(20) \qquad (\tilde{A} + \delta\,\tilde{A})\,X = \tilde{Y} + \delta\,\tilde{Y}$$

avec :

$$(21) \qquad \delta\,\tilde{A} = \begin{vmatrix} \delta A_1 & & & \\ & \delta A_2 & & \\ & & \ddots & \\ & & & \delta A_n \end{vmatrix} \qquad \text{et } \delta\,\tilde{Y} = \begin{vmatrix} \delta Y_1 \\ \delta Y_2 \\ \vdots \\ \delta Y_n \end{vmatrix}$$

où les δA_i sont toutes identiques à δA donnée par (16) et où chaque δY_i est donnée par une relation semblable à (17).

<u>Lemme I</u> :

Dans la résolution du système linéaire $\tilde{A}\,X = \tilde{Y}$, si on utilise la méthode de Gauss, la solution réellement obtenue <u>peut-être</u> considérée comme la solution exacte du système $(\tilde{A} + \delta\,\tilde{A})\,X = \tilde{Y} + \delta\,\tilde{Y}$.

$\delta\,\tilde{A}$ et $\delta\,\tilde{Y}$ étant données par (21), (16) et (17).

<u>b Résolution du système $BX = Y$</u>

où :

$$(22) \quad B = \begin{vmatrix} \beta & & & & 1 & & & \\ & \beta & & & & 1 & & \\ & & \ddots & & & & \ddots & \\ 1 & & & \beta & & & & 1 \\ & 1 & & & \beta & & & & 1 \\ & & \ddots & & & \ddots & & & & \ddots \\ & & & 1 & & & \beta & & & & 1 \\ & & & & \ddots & & & \ddots \\ & & & & & 1 & & & \beta \end{vmatrix} \qquad B(\ell,\ell) \quad \ell = n.n$$

(la colonne $n+1$ et la ligne $(n+1)$ sont indiquées par des flèches)

En triangulant la matrice B, les n premières lignes ne sont pas modifiées, quant aux autres à chaque pas $K \geqslant n$ je modifie dans la $(K+1)^{ème}$ ligne les éléments $(K+1, K+1)$ et $(K+1, \ell+1)$, l'élément $(K+1, K+1-n)$ étant pris nul (principe de la méthode) et l'élément $(K+1, K+1+n)$ n'ayant pas changé (toujours égal à 1).

Donc les seules erreurs qui restent sont :

$$\eta^{(K)}_{K+1,K+1} \qquad K = n, \ldots\ldots \ell - 1$$

$$\rho^{(K)}_{K+1} \qquad K = n, \ldots\ldots \ell - 1$$

$$\gamma_i \qquad i = 1, \ldots\ldots \ell$$

le vecteur X réellement obtenu peut-être considéré comme solution exacte du système :

$$(23) \quad (B + \delta B)\, X = Y + \delta Y$$

avec :

$$(24) \quad \delta B = \begin{bmatrix} 0 & & & & 0 \\ & \ddots & & & \\ & & 0 & & \\ & & & \eta_{n+1} \cdot \bar{\lambda}'_{n+1} & \\ 0 & & & & \eta_{n+2} \cdot \bar{\lambda}'_{n+2} \\ & & & & \\ & & & & \eta_\ell \, \bar{\lambda}'_\ell \end{bmatrix}$$

$$(25)\quad \delta Y = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ p_{n+1}^{(n)} \bar{y}_{n+1} \\ \vdots \\ p_{\ell}^{(\ell-1)} \bar{y}_{\ell}^{(\ell)} \end{bmatrix} \qquad \begin{bmatrix} \gamma_1 \cdot \bar{\lambda}'_1 \\ \vdots \\ \gamma_\ell \cdot \bar{\lambda}'_\ell \cdot \bar{\xi}_\ell \end{bmatrix}$$

où $\gamma_K = \gamma^{(K)}_{K+1,K+1}$ et $\bar{\lambda}'_K = \bar{b}^{(K)}_{K,K}$

$b^{(K)}_{i,j}$ étant l'élément (i,j) de la matrice $B^{(K)}$

lemme II

"Dans la résolution du système linéaire $BX = Y$, si on utilise la méthode de Gauss, la solution réellement obtenue peut-être considérée comme la solution exacte du système :

$$(B + \delta B)\, X = Y + \delta Y.\ "$$

δB et δY étant données par (24) et (25)

c.) Supposons que toutes les erreurs relatives sont égales à ϵ.

Alors les égalités (16) et (17) s'écrivent :

$$(26)\quad \delta A = \epsilon \begin{bmatrix} 0 & & & & \\ & \bar{\lambda}_2 & & & \\ & & \bar{\lambda}_3 & & \\ & & & \ddots & \\ & & & & \bar{\lambda}_n \end{bmatrix}.$$

$$(27)\quad \delta Y = \epsilon \begin{bmatrix} 0 \\ \bar{y}_2^{(2)} \\ \vdots \\ \bar{y}_n^{(n)} \end{bmatrix} + \epsilon \begin{bmatrix} \bar{\lambda}_1 \cdot \bar{\xi}_1 \\ \bar{\lambda}_2 \cdot \bar{\xi}_2 \\ \vdots \\ \bar{\lambda}_n \cdot \bar{\xi}_n \end{bmatrix}$$

et chaque δA_i et δY_i des (21) sera donnée par (26) et (27).

Les égalités (24) et (25) deviennent :

$$(28)\quad \delta B = \epsilon \begin{bmatrix} 0 & & & & & & \\ & 0 & & & & & \\ & & \ddots & & & & \\ & & & 0 & & & \\ & & & & \bar{\lambda}'_{n+1} & & \\ & & & & & \bar{\lambda}'_{n+2} & \\ & & & & & & \ddots \\ & & & & & & & \bar{\lambda}'_{\ell} \end{bmatrix}$$

$$(29)\quad \delta Y = \epsilon \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \bar{y}_{n+1}^{(n+1)} \\ \vdots \\ \bar{y}_{\ell}^{(\ell)} \end{bmatrix} + \epsilon \begin{bmatrix} \bar{\lambda}'_1 \cdot \bar{\xi}_1 \\ \vdots \\ \bar{\lambda}'_{\ell} \cdot \bar{\xi}_{\ell} \end{bmatrix}$$

si on pose :

$$\lambda = \begin{bmatrix} \bar{\lambda}_1 & & \\ & \ddots & \\ & & \bar{\lambda}_n \end{bmatrix} \qquad E = \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 1 \end{bmatrix}$$

(29)'

$$T = \begin{bmatrix} 0 & & & & \\ -\frac{1}{\bar{\lambda}_1} & 1 & & & \\ & -\frac{1}{\bar{\lambda}_2} & 1 & & \\ & & \ddots & \ddots & \\ & & & -\frac{1}{\bar{\lambda}_{n-1}} & 1 \end{bmatrix}$$

(26) et (27) s'écrivent alors :

$$(30) \qquad \delta A = \epsilon \lambda . E$$

$$\delta Y = \epsilon T . Y + \epsilon \lambda . X$$

l'écriture (11) devient alors :

$$(31) \qquad (A + \epsilon \lambda E) X = Y + \epsilon T Y + \epsilon \lambda X$$

ou $$(A + \epsilon \lambda E - \epsilon \lambda) X = (I + \epsilon T) Y$$

et comme le signe de ϵ est incertain alors :

$$(31)' \qquad (A + \epsilon \lambda F) X = (I + \epsilon T) Y$$

avec :

$$(32) \quad F = \begin{bmatrix} 1 & & & & \\ & 2 & & 0 & \\ & & 2 & & \\ & & & \ddots & \\ & 0 & & & 2 \end{bmatrix}$$

et le système donné par (20) devient :

$$(33) \quad (\tilde{A} + \epsilon \tilde{\lambda}\, \tilde{F})\, X = (I + \epsilon\, \tilde{T})\, \tilde{Y}$$

avec :

$$(34) \quad \tilde{F} = \begin{vmatrix} F_1 & & & \\ & F_2 & & \\ & & \ddots & \\ & & & F_n \end{vmatrix}$$

$$(35) \quad \tilde{\lambda} = \begin{vmatrix} \tilde{\lambda}_1 & & & \\ & \tilde{\lambda}_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{vmatrix} \quad \text{et } \tilde{T} = \begin{vmatrix} T_1 & & & \\ & T_2 & & \\ & & \ddots & \\ & & & T_n \end{vmatrix}$$

chaque sous-matrice F_i étant donnée par (32), ainsi que les sous-matrices $\tilde{\lambda}_i$ et T_i sont données par (29)'

Lemme III.

Dans la résolution du système linéaire $\tilde{A}\ X = Y$, si on utilise la méthode de Gauss et si toutes les erreurs relatives sont de l'ordre de ϵ sur chaque opération de la forme a + b. c ou (a+b)/c alors la solution réellement obtenue peut-être considérée comme la solution exacte du système

$$(\tilde{A} + \epsilon\ \tilde{\lambda}\ \tilde{F})\ X = (I + \epsilon\ \tilde{T})\ Y$$

$\tilde{\lambda}$, $\tilde{F}$, $\tilde{T}$ étant données par (34) et (35).

Posons maintenant :

$$\lambda' = \begin{vmatrix} \bar{\lambda}'_1 & & & \\ & \bar{\lambda}'_2 & & \\ & & \ddots & \\ & & & \bar{\lambda}'_\ell \end{vmatrix} \qquad E' = \begin{vmatrix} 0 & & & & \overset{n+1}{\downarrow} \\ & \ddots & & & \\ & & 0 & & \\ & & & 1 & \\ & & & & \ddots \\ & & & & & 1 \end{vmatrix}$$

(36)

$$T' = \begin{vmatrix} 0 & & & & & \overset{n+1}{\downarrow} \\ & 0 & & & & \\ & & \ddots & & & \\ & & & 0 & & \\ & & & & 1 & \\ \overset{n+1}{\rightarrow} & \frac{1}{\bar{\lambda}'_1} & & & & \\ & & -\frac{1}{\bar{\lambda}'_2} & & 1 & \\ & & & \ddots & & \ddots \\ & & & & -\frac{1}{\bar{\lambda}'_{\ell-n}} & & 1 \end{vmatrix}$$

λ', E', T' étant des matrices (ℓ, ℓ) ; $\ell = n.n$

Alors les relations (28) et (29) deviennent :

$$\delta\ B = \epsilon\ \lambda'\ E'$$

$$\delta\ Y = \epsilon\ T'\ Y + \epsilon\ \lambda'\ X$$

et le système (23) sera remplacée par :

$$(37) \quad (B + \epsilon \lambda' E') X = Y + \epsilon T' Y + \epsilon \lambda' X$$

ou

$$(B + \epsilon \lambda' E' - \epsilon \lambda') X = (I + \epsilon T') Y$$

et comme le signe de ϵ est incertain, alors :

$$(38) \quad (B + \epsilon \lambda' F') X = (I + \epsilon T') Y$$

avec :

$$(39) \quad F' = \begin{vmatrix} & & & & n+1 \\ & & & & \downarrow \\ 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 2 \\ & & & & & \ddots \\ & & & & & & 2 \end{vmatrix}$$

lemme IV

Dans la résolution du système linéaire $B X = Y$
si l'on utilise la méthode de Gauss et si toutes les erreurs sont de l'ordre de ϵ sur chaque opération de la forme $a + b.c$ ou $(a+b)/c$, alors la solution réellement obtenue peut-être considérée comme la solution exacte du système

$$(B + \epsilon \lambda' F') X = (I + \epsilon T') Y$$

λ', F', T' étant données par (36) et (39)

d. Erreur dans la méthode de Peaceman-Rachford* quand on calcule en flottant avec la méthode de Gauss.

Considérons les deux relations (1) et (2), i.e.

$$(40) \quad \begin{cases} (rI + H) X^{(m+1/2)} = (rI - V) X^{(m)} + S \\ (rI + V) X^{(m+1)} = (rI - H) X^{(m+1/2)} + S \end{cases}$$

avec $S = \begin{vmatrix} S_1 \\ S_2 \\ \vdots \\ S_n \end{vmatrix}$

*Il s'agit ici de résoudre $\Delta U = 0$ dans un rectangle avec un quadrillage ayant n points suivant Ox et n points suivant oy et avec des pas h et k inégaux.

les formules (40) sont de la forme :

$$(41)\quad \begin{cases} \tilde{A}\, X^{(m+1/2)} = C\, X^{(m)} + S_h \\ B\, X^{(m+1)} = D\, X^{(m+1/2)} + S_h \end{cases}$$

avec :

$$(42)\quad \tilde{A} = \begin{vmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_n \end{vmatrix} \qquad \text{et } A_i = \begin{vmatrix} K & 1 & & \\ 1 & K & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & \\ & & 1 & K \end{vmatrix}$$

A étant composée des sous-matrices toutes identiques

$$(43)\quad C = K'' \begin{vmatrix} K' & & & & & n+1 \downarrow \\ & K' & & & & 1 \\ & & K' & & 1 & \\ n+1 \rightarrow 1 & & & K' & & \\ & 1 & & & & \\ & & 1 & & & 1 \\ & & & 1 & & \\ & & & & K' & \end{vmatrix}$$

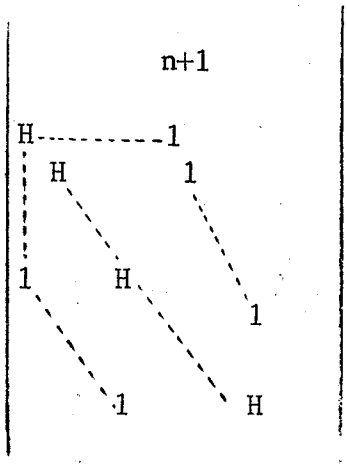
$K = -(r\,h^2 + 2)$

$K' = rh^2 - 2$

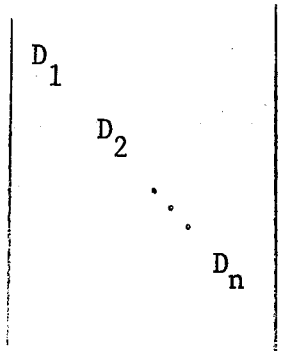
$K'' = -h^2/k^2$

$S_h = -h^2\, S$

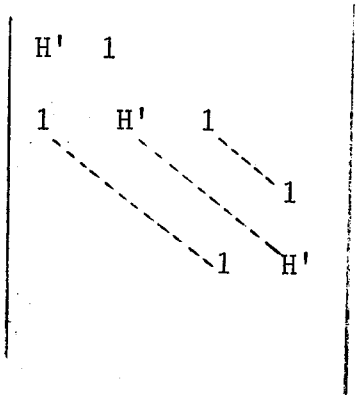
(44) $B =$



(45) $D = H''$



et $D_i =$



D étant composée des sous-matrices toutes identiques

$$H = -(rk^2 + 2)$$

$$H' = rh^2 - 2$$

$$H'' = -\frac{k^2}{h^2} \qquad \text{et } S_k = -k^2 S$$

$$S_p = \begin{vmatrix} \alpha(p,0) & & \\ 0 & & \\ \vdots & & \\ 0 & & \\ \alpha(p,n+1) & & \end{vmatrix} \qquad \text{pour p allant de 2 à n - 1}$$

(46)

$$S_1 = \begin{vmatrix} \alpha(1,0) + \alpha(0,1) & & \\ 0 + \alpha(0,2) & & \\ \vdots & & \\ 0 + \alpha(0,n-1) & & \\ \alpha(1,n+1) + \alpha(0,n) & & \end{vmatrix} \quad \text{et } S_n = \begin{vmatrix} \alpha(n,0) + \alpha(n+1,1) & & \\ 0 + \alpha(n+1,2) & & \\ \vdots & & \\ 0 + \alpha(n+1,n-1) & & \\ \alpha(n,n+1) + \alpha(n+1,n) & & \end{vmatrix}$$

les α désignant les valeurs de la fonction sur la frontière, de (41) on tire :

$$(47)\quad \begin{cases} X^{(m+1/2)} = \tilde{A}^{-1}\, C\, X^{(m)} + \tilde{\tilde{A}}^{-1}\, S_h \\ X^{(m+1)} = B^{-1}\, D\, \tilde{A}^{-1}\, C\, X^{(m)} + B^{-1}\, D\, \tilde{\tilde{A}}^{-1}\, S_h + B^{-1}\, S_h \end{cases}$$

c'est-à-dire :

$$(48)\quad X^{(m+1)} = M\,(r)\, X^{(m)} + N\,(r)$$

Partant de $X^{(m)}$, le vecteur théoriquement trouvé $X^{(m+1)}$ satisfait à (48), mais pratiquement nous obtenons $\tilde{X}^{(m+1)}$ qui vérifie :

$$(49)\quad \tilde{X}^{(m+1)} = M'(r)\, X^{(m)} + N'\,(r).$$

En effet, considérons d'abord :

$$\tilde{A}\, X^{(m+1/2)} = \tilde{Y} \qquad \text{avec } \tilde{Y} = C\, X^{(m)} + S_h$$

le vecteur réellement trouvé est $\tilde{X}^{(m+1/2)}$ qui satisfait à :

$$(\tilde{A} + \varepsilon\, \tilde{\lambda}\, \tilde{F})\, \tilde{X}^{(m+1/2)} = (I + \varepsilon\, \tilde{T})\, \tilde{Y} \quad \text{(formule 33)}$$

mais le calcul de Y fait intervenir les opérations

$$K''\,(x_{i-n} + K'\, x_i + x_{i+n}) \text{ et } -h^2\, s_i \text{ qui font intervenir}$$

des erreurs ; donc si les erreurs relatives sur les opérations de la forme a+b.c ou (a+b)/c sont de l'ordre de ε on aura en fait :

$$K''\left[(x_{i-n}+K'\, x_i)(1+\varepsilon)+x_{i+n}\right](1+\varepsilon) \simeq K''\,(x_{i-n} + K'\, x_i + x_{i+n})\,(1+\eta)$$

et :

$$-h^2\,(1+2\varepsilon)\, s_i = -h^2\,(1+\eta)\, s_i \quad \text{avec } \eta = 2\varepsilon$$

donc

$$\bar{C} = (1+\eta)\, C \text{ et } \bar{S}_h = (1+\eta)\, S_h$$

et nous aurons en fin :

$$(50)\quad (\tilde{A} + \varepsilon\, \tilde{\lambda}\, \tilde{F})\, \tilde{X}^{(m+1/2)} = (I + \varepsilon\, \tilde{T})\,(1+\eta)\, C\, X^{(m)} + (I + \varepsilon\tilde{T})\,(1+\eta) S_h$$

Considérons maintenant la 2ème relation de (41)

$$B\,X^{(n+1)} = Z \qquad \text{avec } Z = D\,\tilde{X}^{(m+1/2)} + S_k$$

le vecteur réellement obtenu $\tilde{X}^{(m+1)}$ satisfait à :

$$(B + \epsilon\,\lambda'\,F')\,\tilde{X}^{(m+1)} = (I + \epsilon\,T')\,Z$$

en faisant le même raisonnement pour le calcul de Z que celui de $Y = CX^{(m)} + S_h$ il vient :

$$(51)\quad (B+\epsilon\,\lambda'F')\,\tilde{X}^{(m+1)} = (I + \epsilon\,T')\,(1+\eta)\,D\,\tilde{X}^{(m+1/2)} + (1+\epsilon T')(1+\eta)S_k$$

De (50) et (51) on tire :

$$(52)\quad \tilde{X}^{(m+1/2)} = (\tilde{A}+\epsilon\tilde{\lambda}\tilde{F})^{-1}(I+\epsilon\tilde{T})(1+\eta)CX^{(m)}+(\tilde{A}+\epsilon\tilde{\lambda}\tilde{F})^{-1}(I+\epsilon\tilde{T})(1+\eta)S_h$$

$$(53)\quad \tilde{X}^{(m+1)} = (B+\epsilon\lambda'F')^{-1}(I+\epsilon\tilde{T}')(1+\eta)D\tilde{X}^{(m+1/2)}+(B+\epsilon\lambda'F')^{-1}(I+\epsilon\tilde{T}')(1+\eta)S_k$$

d'où :

$$(54)\quad \begin{aligned}\tilde{X}^{(m+1)} &= (B+\epsilon\lambda'F')^{-1}(I+\epsilon T')(1+\eta)D(\tilde{A}+\epsilon\tilde{\lambda}\tilde{F})^{-1}(I+\epsilon\tilde{T})(1+\eta)CX^{(m)}\\ &+ (B+\epsilon\lambda'F')^{-1}(I+\epsilon T')(1+\eta)D(\tilde{A}+\epsilon\tilde{\lambda}\tilde{F})^{-1}(I+\epsilon\tilde{T})(1+\eta)\,S_h\\ &+ (B+\epsilon\lambda'F')^{-1}(I+\epsilon T')(1+\eta)\,S_k\end{aligned}$$

(54) est de la forme :

$$(55)\quad \underline{\tilde{X}^{(m+1)} = M'(r)\,X^{(m)} + N'(r)}$$

avec :

$$(56)\quad M'(r) = (1+\eta)^2(B+\epsilon\lambda'F')^{-1}(I+\epsilon T')D(\tilde{A}+\epsilon\tilde{\lambda}\tilde{F})^{-1}(I+\epsilon\tilde{T})C$$

$$(57)\quad \begin{aligned}N'(r) &= (1+\eta)^2(B+\epsilon\lambda'F')^{-1}(I+\epsilon T')D(\tilde{A}+\epsilon\tilde{\lambda}\tilde{F})^{-1}(I+\epsilon\tilde{T})S_h\\ &+ (1+\eta)(B+\epsilon\lambda'F')^{-1}(I+\epsilon T')\,S_k\end{aligned}$$

les matrices $\tilde{A}$, B, C, D, $\tilde{\lambda}$, $\tilde{T}$, λ' et T' données plus haut sont fonction du paramètre d'accélération r.

Finalement le vecteur $X^{(m+1)}$ qui satisfait théoriquement à (48), dans la pratique, il peut-être considéré comme satisfaisant à (55). Alors si nous prenons des r différents au nombre de k par exemple, le même calcul fait plus haut nous permet d'écrire :

$$(58)\quad \begin{cases} \tilde{X}^{(m+1)} = M'(r_1)\,\tilde{X}^{(m)} + N'(r_1) \\ \tilde{X}^{(m+2)} = M'(r_2)\,\tilde{X}^{(m+1)} + N'(r_2) \\ \vdots \\ \tilde{X}^{(m+k)} = M'(r_k)\,\tilde{X}^{(m+k-1)} + N'(r_k) \end{cases}$$

de sorte que :

$$(59)\quad \tilde{X}^{(m+k)} = \left[\prod_{i=k}^{1} M'(r_i)\right] \tilde{X}^{(m)} + \sum_{j=1}^{k-1} \left(\prod_{i=k}^{j+1} M'(r_i)\right) N'(r_j) + N'(r_k)$$

qui est de la forme :

$$(60)\quad \underline{\tilde{X}^{(m+k)} = \mathcal{M}'\,\tilde{X}^{(m)} + \mathcal{N}'}$$

avec :

$$(61)\quad \begin{cases} \mathcal{M}' = \prod_{i=k}^{1} M'(r_i) \\ \mathcal{N}' = \sum_{j=1}^{k-1} \left(\prod_{i=k}^{j+1} M'(r_i)\right) N'(r_j) + N'(r_k) \end{cases}$$

donc le vecteur réellement trouvé $\tilde{X}^{(m+k)}$ à partir de $X^{(m)}$ peut-être considéré comme donné par (60) au lieu de

$$(61)\quad \underline{X^{(m+k)} = \mathcal{M}\,X^{(m)} + \mathcal{N}} \qquad \text{(relation (5) au début de ce chapitre)}$$

et en posant $\tilde{X}^{(m+k)} = \tilde{V}_1$ et $X^{(m)} = \tilde{V}_o$

alors il vient :

$$\tilde{V}_1 = \mathcal{M}' \tilde{V}_o + \mathcal{N}'$$

$$\tilde{V}_{m+1} = \mathcal{M}' \tilde{V}_m + \mathcal{N}'$$

et plus généralement

$$\boxed{\tilde{V}_i = \mathcal{M}'^{i-\alpha} \tilde{V}_\alpha + \sum_{j=o}^{i-\alpha-1} \mathcal{M}'^{j} \mathcal{N}'} \tag{62}$$

et le $i^{ème}$ itéré peut-être considéré comme solution exacte du système (62) au lieu du système (7)

e) Conclusion générale

Dans la résolution de l'équation $\Delta U = 0$ dans un rectangle par la méthode de Peaceman-Rachford, si l'on travaille en flottant, si l'on utilise la méthode de triangulation-élimination de Gauss et si les erreurs relatives sur les opérations de la forme $a + b \bullet c$ et $(a + b)/c$ sont de l'ordre de ε, les vecteurs réellement obtenus $X^{(m+p)}$ et Y_i qui vérifient théoriquement les équations :

$$\begin{aligned} X^{(m+p)} &= M(r_p)\, X^{(m+p-1)} + N(r_p)\, S \\ X^{(m+k)} &= \mathcal{M} X^{(m)} + \mathcal{N} \\ Y_i &= \mathcal{M}^i Y_o + \sum_{j=o}^{i-1} \mathcal{M}^j \mathcal{N} \\ Y_i &= \mathcal{M}^{i-\alpha} Y_\alpha + \sum_{j=o}^{i-\alpha-1} \mathcal{M}^j \mathcal{N} \end{aligned} \tag{63}$$

peuvent-être considérés comme solutions exactes de

$$\left\{ \begin{aligned} \tilde{X}^{(m+p)} &= M'(r_p)\, \tilde{X}^{(m+p-1)} + N'(r_p) \\ \tilde{X}^{(m+k)} &= \mathcal{M}' \tilde{X}^{(m)} + \mathcal{N}' \\ \tilde{V}_i &= \mathcal{M}'^{i} \tilde{V}_o + \sum_{j=o}^{i-1} \mathcal{M}'^{j} \mathcal{N}' \\ \tilde{V}_i &= \mathcal{M}'^{i-\alpha} \tilde{V}_\alpha + \sum_{j=o}^{i-\alpha-1} \mathcal{M}'^{j} \mathcal{V}' \end{aligned} \right. \tag{64}$$

IV - PARTIE

-=-=-

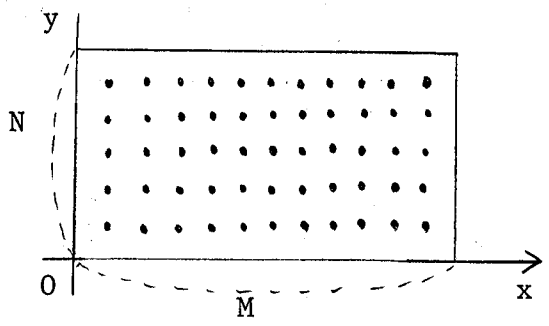
RESULTATS NUMERIQUES

-=-=-=-=-=-

A. - Rappels et définitions

Dans tout ce qui va suivre, nous désignerons par :

. Pb (S,T;M,N) le problème du Laplacien ($\Delta U = 0$) dans un rectangle dont les côtés sont parallèles aux axec de coordonnées et ayant une longueur S suivant ox et T suivant oy ; avec M points intérieurs sur l'axe des x et N points intérieurs sur l'axe des y ; les conditions aux limites étant U (x,y) = SIN (y)* CH (x) sur la frntière Γ du rectangle.



. PR1 (t) la méthode de Peaceman-Rachford avec un seul ensemble de t paramètres.

. PR2 (t) la méthode Peaceman-Rachford avec deux ensembles de t paramètres chacun.

. t le nombre de paramètres sélectionnés.

. NP le nombre de "pas" (ou de tours) effectués.

. NI le nombre d'itérations effectuées (il est clair que NI = t. NP)

. h le pas du quadrillage suivant ox et k le pas du quadrillage suivant oy.

. MSFib. la méthode qui calcule les paramètres accélérateurs au moyen des suites de Fibonaci (page 40)

. MWach. la méthode qui calcule les paramètres accélérateurs au moyen de la formule approximative de Wachpress (page 43 formule 84)

. U(I,J) la vraie solution de l'équation $\Delta U = 0$

. $\bar{U}(I,J)$ la vraie solution du système discrétisé

. U*(I,J) la solution réellement obtenue.

. $ERMAX = \max_{I,J} \left|U(I,J) - U^*(I,J)\right|$ = erreur globale maximum

. ERRELC l'erreur relative correspondante à ERMAX

. $ERRELMAX = \max_{I,J} \left| \frac{U(I,J) - U^*(I,J)}{U(I,J)} \right|$

. DSH la méthode dite "directe SH" étudiée par Melle DI CRESCENZO*

. SUROP la méthode de surrelaxation optimisée (Voir Varga [5] page 105)

Tous les programmes ont été écrits en Fortran et le temps de calcul est donné à une seconde près.

---

* Voir DI CRESCENZO [8]

B. Calcul des paramètres accélérateurs par deux méthodes différentes et comparaison des résultats.

Nous prenons :

$a = 10$ et $b = 400$

Pour $t = 2$, nous trouvons :

(I)

Méthode	r_1	r_2
MSFib	208, 55093	19, 179964
MWach	208, 55959	19, 179171

Pour $t = 4$, nous trouvons

(II)

Méthode	r_1	r_2	r_3	r_4
MSFib	331,13414	116,88275	34,222327	12,079695
MWach	331,15272	116,88520	34,221611	12,079018

Pour $t = 3$, nous ne pouvons utiliser que la deuxième méthode MWach, nous trouvons :

(III)

Méthode	r_1	r_2	r_3
MWach	290,00341	63,245552	13,792942

Nous constatons qu'il y a au moins cinq chiffres exacts et que, même si t n'est pas de la forme 2^k les paramètres sont situés équitablement de part et d'autre de $\sqrt{ab}$ et sont tels :

$$r_i * r_{t-i+1} = ab \qquad \forall_i = 1, \ldots, t$$

Pour le calcul de deux ensembles de paramètres, nous avons pris $\lambda \in [a,b] \equiv [a_o, b_o]$ et $\nu \in [c,d] \equiv [c_o, d_o]$

avec :

$a = 1 \qquad b = 100$

$c = 10 \qquad d = 1000$

et $t = 4$

si nous utilisons la méthode de la page 62, nous trouvons :

(IV)
$$\begin{cases} R_{H_1} = 495,0161 & R_{V_1} = 89,83595 \\ R_{H_2} = 82,11834 & R_{V_2} = 42,27370 \\ R_{H_3} = 23,65535 & R_{V_3} = 12,17757 \\ R_{H_4} = 11,13141 & R_{V_4} = 2,020132 \end{cases}$$

avec $\mu_4 = 0,001316993$

si nous utilisons la méthode de la page 71, nous trouvons :

(V)
$$\begin{cases} R_{H_1} = 495,1317 & R_{V_1} = 89,84017 \\ R_{H_2} = 82,12130 & R_{V_2} = 42,27481 \\ R_{H_3} = 23,65475 & R_{V_3} = 12,17711 \\ R_{H_4} = 11,13088 & R_{V_4} = 2,019665 \end{cases}$$

avec $\mu_4 = 0,001317394$

Nous constatons que, si les R_H et R_V calculés par la première méthode diffèrent très peu de ceux calculés par la deuxième méthode, la différence δ entre la borne du rayon spectral calculée par la première méthode et celle calculée par la deuxième méthode est presque insignifiante :

$$\delta = 0,001317394 - 0,001316993 = 4.10^{-7}$$

C. Temps de Calcul

Considérons le problème Pb (1,1;M,N) ; faisons varier M et N, et en utilisant la méthode PR1 (4) nous obtenons le tableau suivant :

(VI)

M	N	NP	NI	T en secondes
9	9	3	12	1
19	19	3	12	4
29	29	4	16	16
39	39	6	24	39
49	49	6	24	63
59	59	6	24	96
69	69	8	32	149

la dernière colonne de (VI) donne le temps (en secondes) de calcul pour chaque système.

D. Influence du nombre t sur la rapidité de convergence.

Considérons le problème Pb (1,1;9,9) et faisons varier t et arrêtons le calcul chaque fois que le maximum du résidu relatif (page 22) est inférieur à 10^{-5}. Nous obtenons le tableau suivant :

(VII)

t	NP	NI
1	24	24
2	7	14
3	10	30
4	3	12
5	3	15
6	3	18

Considérons maintenant le problème Pb (1,1;39,39) et faisons varier t. Après l'arrêt du programme pour chaque t relevons la valeur de la solution calculée au milieu du carré, c'est-à-dire pour le cas prèsent U* (20,20) ; d'où le tableau suivant :

(VIII)

t	NP	NI	U* (20,20)
1	80	80	0,54061801
2	16	32	0,54061829
4	4	16	0,54061838
5	3	15	0,54061830
6	3	18	0,54061851
7	2	14	0,54061816
8	2	16	0,54061855
9	2	18	0,54061832
10	3	30	0,54061800

la vraie solution de l'équation $\Delta U = 0$ au point U (20,20) étant :

$$\sin(0,5).\ \mathrm{ch}(0,5) = 0,54061266$$

Pour étudier la rapidité de convergence en fonction de t, il suffit de comparer le nombre d'itérations effectuées, pour chaque t, avant l'arrêt du programme. Ainsi en examinant le tableau (VII) , nous constatons que le nombre 4 correspond au meilleur choix, cependant le nombre 2 n'est pas très éloigné du nombre 4 du point de vue rapidité de convergence. En examinant le tableau (VIII) nous constatons qu'en utilisant $\underline{t}$ au nombre de quatre, cinq, sept ou huit, nous aurons à peu près la même rapidité de convergence.

D'autres expériences nous ont permis de constater que les nombres 4 et 5 correspondent toujours au meilleur choix. Cependant, pour les systèmes tels que $M \geq 40$ et $N \geq 40$ le nombre 8 correspond aussi à ce meilleur choix.

*Remarque.

Il est même conseillé d'utiliser, pour des systèmes tels que $M \geqslant 40$ et $N \geqslant 40$, le nombre 8. En effet, il est vrai qu'en utilisant le nombre 4 ou le nombre 8 le programme s'arrête presque après le même nombre d'itérations, mais le nombre de "pas" effectués dans le cas t = 4 est le double de celui dans le cas t = 8 ; de ce fait nous sommes amenés à faire deux fois plus de tests (test d'arrêt) dans le cas t = 4 que dans le cas t = 8, il en résulte pour des systèmes très grands une notable différence dans le temps de calcul : nous avons considéré le problème Pb (1,1;59,59) ; le tableau suivant nous donne une idée sur la différence de temps :

(XI)

t	NP	NI	NT	T en secondes
4	6	24	6	96
8	3	24	3	90

NT désignant le nombre de tests.

E. Cas d'un rectangle.

Considérons le problème Pb (2π, π ; 39,19) et résolvons-le par les deux méthodes PR1 (4) et PR2 (4) ; nous obtenons le tableau suivant :

(X)

Méthode	NP	NI	T en secondes	$U^*(10,20)$
PR1 (4)	6	24	17	11,666661
PR2 (4)	5	20	13	11,666667

la vraie solution de $\Delta U = 0$ au milieu du rectangle étant $\sin(\pi/2).ch(\pi)=11,592038$

Pour le problème Pb (1,1;79,9) nous obtenons le tableau suivant :

(XI)

Méthode	NP	NI	T en secondes	U* (5, 40
PR1 (4)	5	20	15	0,54065279
PR2 (4)	3	12	9	0,54064892

Nous constatons que la méthode PR2 est nettement plus rapide que la méthode PR1, ceci est dû certainement au fait que le rayon spectral de la matrice de PR2 est inférieur à celui de la matrice de PR1. En effet, considérons le problème Pb (1,1; 49,9), en calculant les bornes des valeurs propres de H et V nous obtenons :

a = 9, 7887 b = 390, 21

c = 9, 8664 d =9990, 1

en utilisant deux ensembles de paramètres l'un entre a et b l'autre entre c et d, nous bornerons le rayon spectral de la matrice de PR2 (4) par 0, 00843 et en choisissant un seul ensemble de paramètres entre a et d, nous bornerons le rayon spectral de la matrice de PR1 (4) par 0, 034659 ; de ce fait découle la différence de temps de calcul ; en plus des deux tableaux (X) et (XI), le tableau suivant correspondant au problème Pb (1,1 ; 49,9), en donne une idée.

(XII)

Méthode	NP	NI	T en secondes	U* (5, 25)
PR1 (4)	6	24	10	0,54065122
PR2 (4)	4	16	7	0,54065066

Nous constatons par ailleurs que le rapport de temps de calcul de PR2 et de PR1 est d'autant plus grand que le rapport cd/ab est plus grand ; en effet pour le problème Pb (1,1 ; 79, 9) les bornes sont :

a = 9, 7887 b = 390, 21

c = 9, 8683 d = 25590

le tableau suivant en donne une idée

(XIII)

Problème	RPT	cd/ab
Pb (1,1;79,9)	1,63	65
Pb (1,1;49,9)	1,43	25

RPT indique le rapport des temps de calcul

* Remarques :

1) Considérons le problème Pb (1,1;79,9)

avec :

a = 9, 7887

b = 390, 21

c = 9, 8683

mais au lieu de prendre d = 25590, qui est la vraie borne supérieure des valeurs propres de la matrice V, nous avons pris d = 9990,1 = $\bar{d}$ et nous avons calculés les paramètres avec ces bornes, c'est-à-dire avec a,b,c et $\bar{d}$ et nous avons utilisé la méthode PR1 (4), c'est-à-dire nous avons calculé 4 paramètres entre a et $\bar{d}$; vu que $\bar{d} < d$ le $\min_i - \max_x$ de la fraction

$$\left| \prod_1^4 \left(\frac{x - r_i}{x + r_i} \right) \right|^2$$ est plus petit dans le cas où

$d = \bar{d} = 9990,1$ que dans le cas où d = 25590 et la méthode PR1 (4) devait-être plus rapide dans le premier cas que dans le second ; or un programme de la méthode PR1 (4) avec a et $\bar{d}$ s'arrête après 18 secondes alors qu'avec a et d il s'arrête après 15 secondes, ceci est très logique car dans le cas où on considère a et $\bar{d}$ la quantité

$$\min_i \quad \max_x \quad \left| \prod_1^4 \left(\frac{x - r_i}{x + r_i} \right) \right|^2$$ n'est plus une borne du rayon spectral.

2) Nous avons remplacé dans la méthode PR2, les R_{H^i} par les R_{V^i}

et réciproquement et nous avons constaté que la rapidé de convergence et la solution calculée n'ont pas été modifiées.

E' Comparaison avec d'autres méthodes

Considérons le problème Pb (2π, π ; M, N), le tableau suivant indique le temps de calcul en secondes :

(XIV)

h	k	M	N	SUROP	DSH	PR1 (4)	PR2 (4)
$\pi/10$	$\pi/10$	19	9	7"	4"	3"	2"
$\pi/20$	$\pi/20$	39	19	42"	30"	18"	14"

Pour le cas où M = 39 et N = 19, nous avons comparé les valeurs obtenues en deux points : la valeur au milieu du rectangle U* (10, 20) et la plus grande valeur à l'intérieur du rectangle que nous désignons par PGV, et si ϵ désigne l'erreur entre deux valeurs l'une calculée par DSH, l'autre par PR2 (4) nous aurons le tableau suivant :

(XV)

Méthode	PGV	U* (10, 20)
DSH	228, 89990	11, 666551
PR2 (4)	228, 90334	11, 666667
ϵ	0, 00344	0, 000118

Nous constatons que le rapport des temps entre DSH et PR2 (4) est $30/14 \simeq \underline{2,2}$ et le rapport des temps entre SUROP et PR2 (4) est $42/14 = \underline{3}$

* Remarque :

Les programmes de la méthode DSH et de la méthode SUROP ont été écrits par Melle DI CRESCENZO [8] en Algol et le temps de calcul obtenu est :

136 secondes pour DSH

188 secondes pour SUROP

pour avoir le temps correspondant en Fortran nous avons divisé le temps (en Algol) par 4,5 .

F. Utilisation des valeurs propres de H et V

Nous avons vu (page 55) que pour que HV = VH, il faut que le domaine dans lequel on cherche une solution de l'équation

$$-\frac{\partial}{\partial x}\left(P(x,y)\ \frac{\partial}{\partial x}\ U(x,y)\right) - \frac{\partial}{\partial y}\left(\varphi(x,y)\ \frac{\partial}{\partial y}\ U(x,y)\right)$$

$$+\sigma(x,y)\ U(x,y) = S(x,y)$$

soit un rectangle dont les côtés sont parallèles aux axec de coordonnées, de plus il faut que P (x,y) soit fonction de x seul, φ (x,y) fonction de y seul et σ (x,y) une constante positive ; c'est-à-dire l'équation ci-dessus se trouve réduite à la forme suivante :

$$-\frac{\partial}{\partial x}\left(f(x)\ \frac{\partial}{\partial x}\ U(x,y)\right) - \frac{\partial}{\partial y}\left(g(y)\ \frac{\partial}{\partial y}\ U(x,y)\right) + \sigma\ U(x,y) = S(x,y)$$

si f(x) = g(y) = constante il est ftrès facile de démontrer que la matrice H aura la forme suivante :

$$h^2 * H = \begin{vmatrix} H_1 & & & \\ & H_2 & & \\ & & \ddots & \\ & & & H_n \end{vmatrix}$$

où :

$$H_1 = H_2 = \ldots = H_n = \begin{vmatrix} c & -1 & & & \\ -1 & C & -1 & & \\ & -1 & C & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & C \end{vmatrix}$$

et la matrice V aura la forme (23) de la page 15 mais le nombre 2 étant remplacé par C alors les valeurs propres de H et V sont connues et données par :

$$\lambda_j = \frac{1}{h^2}\left[C - 2\cos\left(\frac{j\,\Pi}{n+1}\right)\right]$$

$$\nu_j = \frac{1}{k^2}\left[C - 2\cos\left(\frac{j\,\pi}{m+1}\right)\right]$$

Nous avons eu l'idée d'utiliser ces valeurs propres comme paramètres accélérateurs : soient :

$$\lambda_1 < \lambda_2 < \ldots\ldots < \lambda_n$$

les valeurs propres de H (qui sont aussi valeurs propres de V si le domaine est un carré).

Nous avons considéré le problème Pb (1,1;M,N), nous avons fait varier M et N et nous avons arrêté le calcul chaque fois que le résidu relatif devient plus petit que $5.\ 10^{-6}$, d'où le tableau suivant :

(XVI)

M	N	NI	Résidu relatif	Temps en secondes
9	9	4	$4,\ 4.\ 10^{-6}$	1
19	19	6	$1,\ 2.\ 10^{-6}$	3
29	29	5	$1,\ 4.\ 10^{-6}$	5
59	59	4	$3,\ 7.\ 10^{-6}$	12
69	69	4	$0,28.\ 10^{-6}$	16
79	79	4	$3,\ 7.\ 10^{-6}$	21

Les valeurs propres ont été utilisées de la façon suivante : la première itération a été faite avec $r_1 = \lambda_1$, la deuxième avec $r_2 = \lambda_2$, etc ... la $i^{ème}$ avec $r_i = \lambda_i$.

Remarques

1. Nous avons pensé utiliser les valeurs propres dans l'ordre décroissant et pas dans l'ordre croissant comme il a été fait ci-dessus ; en désignant par PR3C la méthode de Peaceman-Fachford exécutée avec les valeurs propres dans l'ordre croissant et par PR3D la même méthode mais avec les valeurs propres dans l'ordre décroissant nous aurons le tableau suivant :

(XVII)

Méthode	M	N	NI	Résidu relatif	temps
PR3C	69	69	4	$0,28.10^{-6}$	16
PR3D	69	69	69	$2,7.10^{-6}$	272

avec la méthode PR3D le programme ne s'arrête qu'après la 69 ième itération avec un résidu relatif de $2,7.10^{-6}$ et un temps de calcul de 272 secondes soit 17 fois le temps de calcul avec PR3C.

2. Nous avons considéré le problème Pb (1, 1; 69, 69) et nous avons utilisé la méthode PR3C un peu modifiée : c'est-à-dire nous avons fait

la première itération avec $r_1 = \lambda_2$

la deuxième itération avec $r_2 = \lambda_1$

la troisième itération avec $r_3 = \lambda_3$

la quatrième itération avec $r_4 = \lambda_4$

⋮

la 69 ème itération avec $r_{69} = \lambda_{69}$

le programme s'arrête alors après $\underline{5}$ itérations, un résidu relatif de $\underline{0,45.10^{-6}}$ et un temps de calcul de 20 secondes (contre 16 secondes avec la PR3C non modifiée)

3. Nous avons considéré le problème Pb (1, 1 ; 29, 29) et nous avons utilisé la méthode PR3C de la façon suivante :

$$r_1 = \lambda_1$$

$$r_2 = \lambda_2$$

$$r_3 = \lambda_1$$

$$r_4 = \lambda_2$$

$$\vdots$$

$$r_{2i+1} = \lambda_1 \qquad \forall_i = 0, \ldots\ldots, 14$$

$$r_{2i} = \lambda_2 \qquad \forall_i = 1, \ldots\ldots, 14$$

le programme s'arrête après la 13ème itération avec un résidu relatif de $0,45.10^{-6}$ et un temps de calcul de 12 secondes (contre 5 pour la PR3C non modifiée)

Toujours pour le problème Pb (1, 1 ; 29, 29) avec la méthode PR3D le programme s'arrête après la 29 ème itération avec un résidu relatif de $0,46.10^{-6}$ et un temps de calcul de 25 secondes (contre 5 pour la PR3C).

De tout ce qui précède nous concluons que si nous utilisons la méthode de Peaceman-Rachford avec les valeurs propres de H ou de V comme paramètres, à condition de les prendre dans l'ordre croissant, la rapidité de convergence est [illegible] et le temps de calcul est très réduit même par rapport à celui de PR1 (4), en effet, considérons le problème Pb (1, 1; 69,69) d'où le tableau suivant :

(XVIII)

Méthode	Temps
PR1 (4)	149
PR3 C	16

4. Considérons le problème Pb (2π, π; 39, 19) en utilisant la méthode PR3C avec les valeurs propres de H, le programme s'arrête après la deuxième itération avec un résidu relatif de $0,6.10^{-6}$ et un temps de calcul de 2 secondes.

Le tableau suivant donne une idée sur la différence entre les temps de calcul des différentes méthodes déjà citées :

(XIX)

Méthode	SUROP	DSH	PR1 (4)	PR2(4)	PR3C
temps	42"	30"	18"	14"	2"

G. Cas où le domaine n'est pas un rectangle

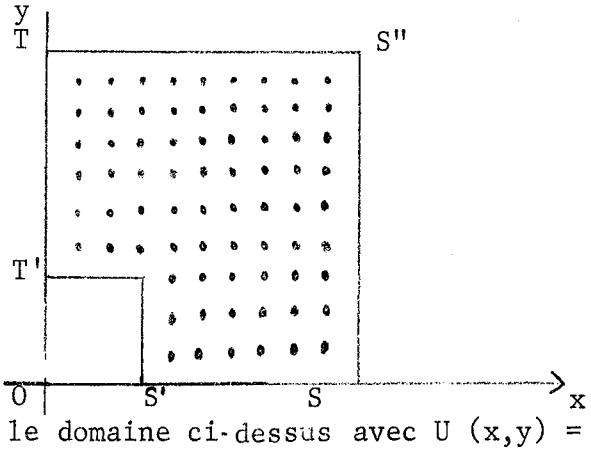
Si le domaine, dans lequel on cherche une solution de $\Delta U = 0$, n'est plus un rectangle la méthode de Peaceman-Rachford n'est plus avantageuse.

En effet, les matrice H et V ne commutent plus et alors les valeurs propres de la matrice Tr ne sont plus de la forme

$$\left(\frac{\lambda_i - r}{\lambda_i + r}\right)\left(\frac{\nu_i - r}{\nu_i + r}\right)$$

et on ne peut plus se prononcer sur le problème de minimisation du rayon spectral de Tr et par conséquent le paramètre $\bar{r}$ correspondant au meilleur choix devient inconnu (de plus on ne sait pas si ce $\bar{r}$ appartient ou pas à l'intervalle $[a,b]$ des valeurs propres de H et V)

Nous avons quand même considéré le problème suivant :



$\Delta U = 0$ dans le domaine ci-dessus avec $U(x,y) = \sin y \operatorname{ch} x$ sur la frontière

0 S = 1

0 S'= 0, 3

0 T = 1

0 T'= 0, 3

les pas du quadrillage étant h = k = 0, 1

Nous aurons alors :

6 points entre S' et S
6 points entre T' et T
9 points entre S et S"
9 points entre T et S"

et nous avons pris pour paramètre r = 61, 8 c'est-à-dire le paramètre correspondant au meilleur choix pour le cas du carré entier ; le programme s'arrête après 70 itérations et 7 secondes de calcul, alors qu'il s'arrête pour le cas du carré après la 17 ème itération et $\leq$ secondes de calcul.

H. Erreur Globale

L'erreur globale, c'est-à-dire la différence entre la vraie solution U (I,J) de l'équation différentielle et la solution U* (I,J) réellement obtenue, est proportionnelle à h^2 (h étant le pas du quadrillage).

En effet, considérons le problème Pb (2π, π ; M, N) et résolvons-le par la méthode PR2 (4) et faisons varier M et N de la façon suivante :

pour M = 19 N = 9
pour M = 39 N = 19
pour M = 59 N = 29
pour M = 79 N = 39

et pour chaque cas, calculons ERMAX ; nous obtenons le tableau suivant :

(XX)

M	N	h = k	ERMAX	ERRELC	I	J
19	9	$\pi/10$	0, 80202	0,0076870	5	17
39	19	$\pi/20$	0, 20183	0,0019344	10	34
59	29	$\pi/30$	0, 090027	0,000958	15	51
79	39	$\pi/40$	0, 050724	0,0005275	20	67

les deux dernières colonnes du tableau ci-dessus indiquent respectivement le lieu de l'erreur (ERMAX). Nous constatons bien que l'erreur est proportionnelle à h^2 et que, par ailleurs, cette erreur (ERMAX) se trouvent au même endroit du rectangle.

Il est évident qu'à ERMAX ne correspond pas ERRELMAX, en effet le tableau suivant en donne une idée

(XXI)

M	N	h = k	ERRELMAX
19	9	$\pi/10$	0,035749
39	19	$\pi/20$	0,00891
59	29	$\pi/30$	0,003965
79	39	$\pi/40$	0,00223

Aussi, ici, ERRELMAX est proportionnelle à h^2

Le graphique (I) donne une idée sur les plages d'erreurs absolues (globales) et le graphique (II) donne une idée sur les plages d'erreurs relatives ; nous avons considéré le problème Pb (2π, π; 39, 19) donc ERMAX = 0,201 et ERRELMAX = 0, 00891 :

Graphique (I) : Nous avons divisé ERMAX par 10 et partagé ainsi le triangle en dix zones : la première contenant les erreurs inférieures à 0, 02 : $\epsilon_a < 0{,}02$

la deuxième : $0,\ 02 < \epsilon_a < 0,\ 04$

la troisième : $0,\ o4 < \epsilon_a < 0,\ 06$

⋮

la dixième : $0,\ 18 < \epsilon_a < 0,\ 2$

Graphique (II) :

Nous avons divisé ERRELMAX par 10 :

ϵ = ERRELMAX

ϵ = E/10

Remarques

1. Nous constatons que l'erreur globale est symétrique si la solution est symétrique : graphique (I).

2. Le graphique (II) montre que l'erreur relative (globale) est presque constante (même ordre de grandeur) par colonne, ceci est une conséquence logique de la remarque déjà faite plus haut.

3. La répartition de l'erreur absolue et celle de l'erreur relative dépendent de la valeur de la solution.

I. Erreur de calcul

Il s'agit ici de la différence entre la vraie solution $\bar{U}$ (I,J) du système discrétisé et la solution U* (I,J) réellement obtenue, cette différence étant due à la capacité limitée de la machine. Pour mettre en évidence cette erreur, que nous avons déjà étudiée théoriquement, nous allons utiliser un sous-programme TRONC, conçu par B. LIOT* et qui a pour effet d'augmenter l'erreur de calcul élémentaire. Nous avons constaté que :

1. L'erreur de calcul est proportionnelle à n^2, n^2 étant l'ordre du système linéaire

2. L'erreur de calcul est proportionnelle à l'erreur élémentaire commise sur chaque opération.

En effet, considérons le problème Pb (π, π; M,N), et pour chaque couple (M,N) résolvons-le une fois avec la précision maximale et une fois en utilisant le sous-programme TRONC et pour chaque cas calculons l'erreur commise sur l'élément qui figurent au milieu du carré, nous obtenons le tableau suivant

(XXII)

M = N	Erreur
9	$46.\ 10^{-6}$
19	$176.\ 10^{-6}$
29	$444.\ 10^{-6}$
39	$792.\ 10^{-6}$

ces calculs ont été faits en tronquant 7 bits c'est-à-dire en travaillant avec 5 décimales.

Les nombres qui figurent dans la colonne erreur du tableau (XXII) indiquent l'erreur moyenne, car d'une itération à une autre l'erreur varie, bien qu'elle conserve le même ordre de grandeur, le graphique (III) donne une idée pour le cas M = N = 29. D'après le tableau XXII nous constatons que l'erreur de calcul est proportionnelle à M·N

*Voir LIOT [4]

Nous avons, par ailleurs considéré le cas M = N = 19 et nous avons fait augmenter l'erreur élémentaire grace au sous-programme TRONC et nous avons constaté que l'erreur de calcul est proportionnelle à l'erreur élémentaire, comme l'indique le tableau suivant :

(XXIII)

TRONC	NCS	E.E	Erreur de calcul
7	5	ϵ	$176.\ 10^{-6}$
10	4	10ϵ	$160.\ 10^{-5}$
13	3	100ϵ	$178.\ 10^{-4}$

la colonne TRONC indique le nombre de bits tronqués, la colonne NCS le nombre de chiffres significatifs correspondant et la colonne E.E l'erreur élémentaire correspondante.

Remarques.

1. Si la solution est symétrique, nous avons constaté que l'erreur de calcul est aussi symétrique.

2. L'erreur de calcul dépend des paramètres r_i choisis et de leur nombre, le tableau (VIII) le montre facilement, mais il nous a été difficile d'expliciter une relation entre les r_i et cette erreur.

3. Considérons le problème Pb (1, 1; M,N) et faisons M = N = 80 puis M = N = 90 et puis M = N = 100 et calculons pour chaque cas la borne μ_t du rayon spectral en fonction du nombre de paramètre t, nous obtenons le tableau suivant :

(XXIV)

M = N	μ_1	μ_4
80	0, 925	0, 0565
90	0, 933	0, 0627
100	0, 939	0, 0686

Pour éviter alors que la méthode ne diverge pour t = 1, car μ_1 est très voisin de 1, il suffit simplement de sélectionner 4 paramètres ; et avec 4 paramètres nous pouvons résoudre n'importe quel système quelle que soit sa grandeur (dans les limites de capacité de la machine).

J. Conclusion

1. Les nombres 4 et 5 correspondent généralement au meilleur choix des paramètres.

2. Dans le cas d'un rectangle la méthode de Peaceman-Rachford est plus rapide que n'importe quelle autre méthode.

Cependant pour le cas où le domaine n'est plus un rectangle il est avantageux d'utiliser la méthode dite "directe SH" ou même la méthode de surrelaxation successive optimisée.

La méthode de Peaceman-Rachford garde toujours un avantage sur la méthode "directe SH" celui de considérer un quadrillage à pas h et k inégaux.

3. En utilisant les valeurs propres de H et V dans l'ordre croissant nous obtenons une convergence nettement plus rapide que celle obtenue avec le meilleur choix de 4 paramètres.

4. L'erreur globale est proportionnelle à h^2

5. L'erreur de calcul est proportionnelle à M.N
et à l'erreur élémentaire commise sur chaque opération.

6. En utilisant 4 paramètres (ou plus) la méthode ne manque jamais de converger faute de capacité limitée de la machine, et ceci quelle que soit la grandeur du système discrétisé.

BIBLIOGRAPHIE

[1] FORSYTHE et WASOW — Finite - Difference Methods for partial Differential Equations. - John Wiley and sons New-York 1960

[2] N. GASTINEL — Matrices du second degré et normes générales en analyse numérique linéaire. Thèse de Doctorat ès Sciences. Université de Grenoble 1960

[3] N. GASTINEL — Sur le meilleur choix des paramètres de surrelaxation (Procédé de Peaceman-Rachford) CHIFFRES - 2 - 1962 - Faculté des Sciences de Grenoble.

[4] B. LIOT — Etude de la propagation des erreurs de calcul dans deux méthodes classiques de résolution de l'équation de la chaleur. Thèse de 3ème cycle Grenoble 1964.

[5] R.S. VARGA — Matrix itérative Analysis - Prentice Hall Englewood Chiffs (EUA) 1962)

[6] E. WACHPRESS — Extended Application of Alternating Direction Implicit Itération Model Problem Theory - Knolls Atomic Power Laboratory, Général Electric Company ; Schenectady - New-York 4/12/63

[7] N. GASTINEL — Analyse Numérique linéaire Cours de l'Université de Grenoble - 1964 -

[8] Melle DI CRESCENZO — "Sur la solution d'un système linéaire aux différences associé au problème de Dirichlet pour l'équation de Laplace". Thèse de 3ème cycle Grenoble 1965.

$\Delta U = 0$

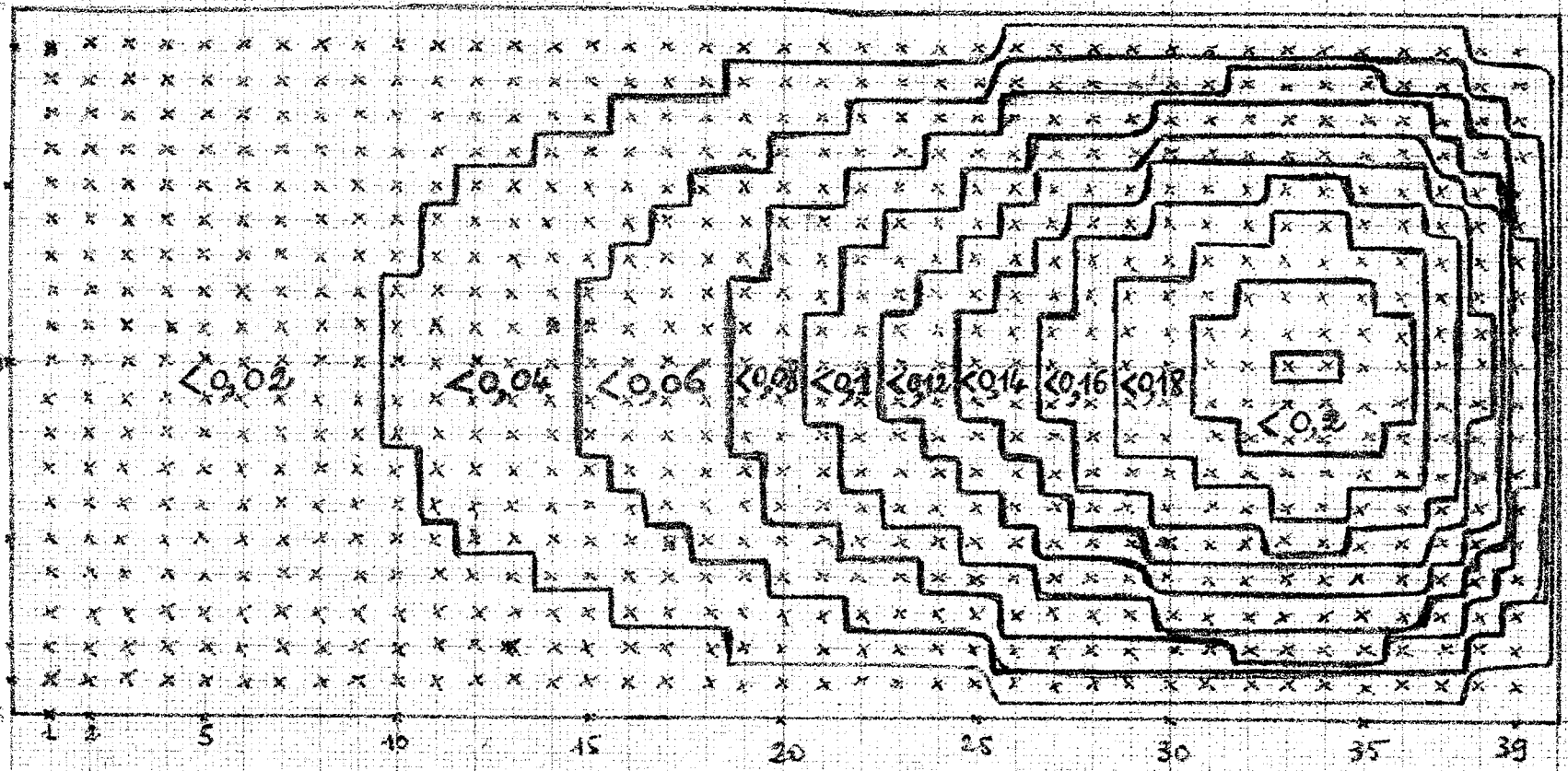
Conditions aux limites $F = \sin y \, \mathrm{ch}\, x$

dans un rectangle de longueur $S = 2\pi$ et de largeur $T = \pi$

Nombre de points suivant ox : $M = 39$ — pas $H = \pi/20$

Nombre de points suivant oy : $N = 19$ — pas $K = \pi/20$

Graphique (I)



Plages d'erreurs (absolues)

Erreur Absolue Maximum = 0,201

Erreur relative correspondante = 0,50193

$\Delta U = 0$

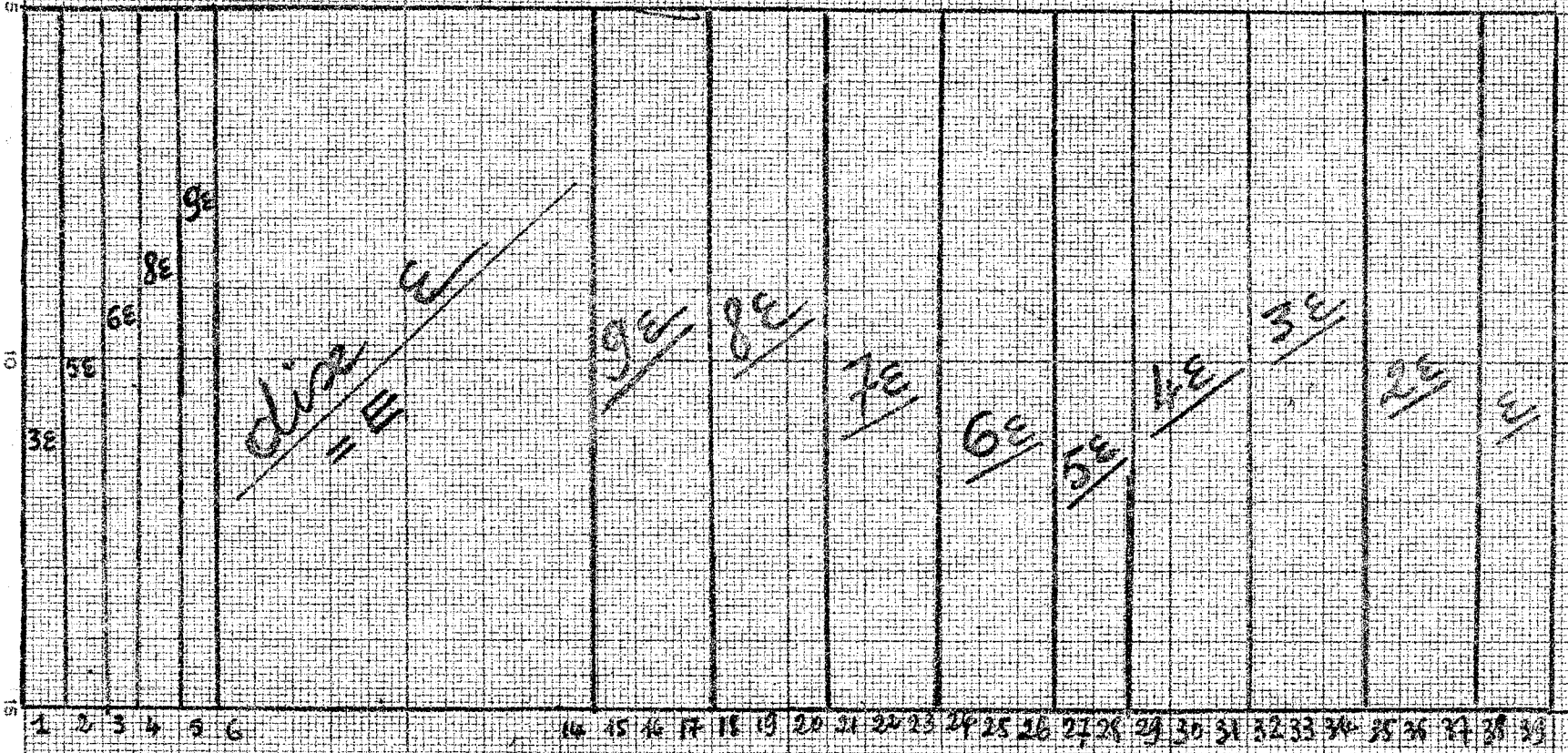
Conditions aux limites $F = \sin y \, \mathrm{ch}\, x$

Graphique (II)

dans un rectangle de longueur $S = 2\pi$ et de largeur $T = \pi$

Nombre de points suivant ox : $M = 39$ pas $H = \pi/20$

Nombre de points suivant oy : $N = 19$ pas $K = \pi/20$



Plages d'erreurs (relatives)

Erreur relative maximum = $E = 0,00891$

$\varepsilon = \frac{E}{10}$

Graphique (III)

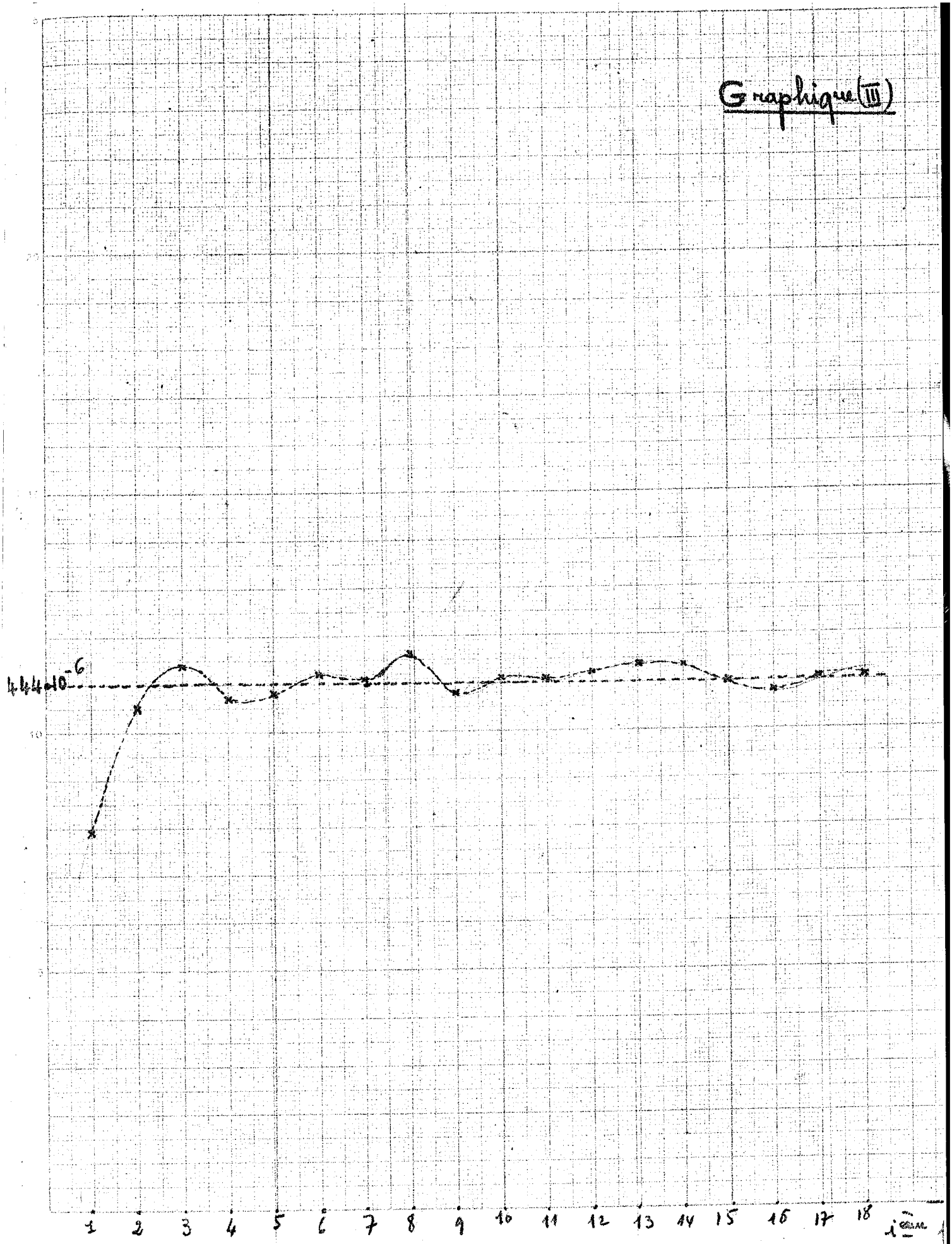


TABLE DES MATIERES

pages

I. Méthode des direction alternées 1

A. Rappel de quelques définitions et théorèmes 1

B. Principe de la méthode 7

C. Convergence de la méthode 13

C.1. Convergence 13

C.2. Calcul de ρ (Tr) 13

C.2.1. Théorème de convergence 16

C.2.2. Théorème de Frobénius 16

C.2.3. Expression de ρ (Tr) 19

C.2.4. Minimisation de ρ (Tr) 20

C.3. Calcul des valeurs et vecteurs propres de H et de V 22

II. Méthode de Peaceman - Rachford ..... 27'

A. Principe 27

B. Calcul de ρ ($\mathfrak{C}$) 30

B.1. Borne du rayon spectral de $\mathfrak{C}$ 30

B.2. Théorème d'existence et d'unicité 31

B.2.1. Lemme 31

B.2.2. Lemme 33

B.2.3. Remarque fondamentale 34

B.2.4. Allure de la courbe 34

B.2.5. Théorème 35

C. Calcul des paramètres accélérateurs 38

C.2. Cas où $t = 2^k$ 40

Cas où t est quelconque 44

pages

III. Etude théorique des erreurs dans la méthode de Peaceman - Rachford .75

1. Notions préliminaires 76

2. Calcul en flottant 79

a) Résolution du système AX = Y 79

b) Résolution du système BX = Y 86

c) Cas où toutes les erreurs relatives sont égales à ε .. 88

d) Erreur dans la méthode de Peaceman - Rachford 93

IV. Résultats numériques 100

A. Rappels et définitions 101

B. Calcul des paramètres accélérateurs 103

C. Temps de calcul 105

D. Influence du nombre t sur la rapidité de convergence 105

E. Cas d'un rectangle 107

E' Comparaison avec d'autres méthodes 110

F. Utilisation des valeurs propres de H et V 111

G. Cas où le domaine n'est pas un rectangle 115

H. Erreur globale 117

I. Erreur de calcul 119

-:-:-:-:-:-:-:-

pages

E. Cas générale - Domaine conforme 46

E.1. Cas où H V = V H 46

E.2. Calcul direct de H V et VH 49

E.3. lemme : condition sur l'équation 51

E.4. Lemme : condition sur l'équation 53

E.5. lemme : condition sur le domaine 53

E.5.1. Corollaire : Domaine conforme 55

E.6. Théorème : condtions nécessaires et suffisantes pour que H et V commutent 55

F. Cas d'un rectangle 56

F.2. Choix d'un seul jeu de paramètres 59

F.3. Choix de deux jeux de paramètres 60

- Choix des meilleurs paramètres $\{R_{V^k}\}$ et $\{R_{H^k}\}$ 62

- Cas où $t = 2^n$ 67

- Cas où t est quelconque 71

Vu,

Grenoble, le

Le Président de la Thèse.

Vu,

Grenoble, le

Le Doyen de la Faculté des Sciences

Vu et permis d'imprimer,

Le Recteur de l'Académie de GRENOBLE