1ère thèse: Matrices du second degré et normes générales en analyse numérique linéaire

Louis Noël Gastinel

T H E S E S

présentées

A LA FACULTE DES SCIENCES

DE L'UNIVERSITE DE GRENOBLE

Pour obtenir

Le grade de Docteur ès Sciences Mathématiques

Par

Noël GASTINEL

Agrégé de l'Université

1° THESE

MATRICES DU SECOND DEGRE ET NORMES GENERALES EN ANALYSE NUMERIQUE LINEAIRE

2° THESE

Propositions données par la Faculté

- LE THEOREME DE STONE - WEIRSTRASS

Thèse soutenue le 12 décembre 1960 devant la Comission d'examen

M. CHABAUTY, Président

MM. DE POSSEL, KUNTZMANN, HACQUES } examinateurs

Je tiens à assurer de ma respectueuse reconnaissance :

Mon maitre, Monsieur le professeur J.KUNZTMANN, directeur du laboratoire de calcul de l'Université de GRENOBLE, sous la direction du quel ce travail a été effectué, pour la bienveillante attention et le soutien constant qu'il m'a accordés, et surtout pour le goût du Calcul qu'il ne cesse de donner à tous ses élèves,

Monsieur le professeur DE POSSEL, de l'Université de PARIS, qui m'a fait l'honneur de se joindre à mon jury ,

Monsieur le professeur CHABAUTY, qui a bien voulu en accepter la présidence,

Monsieur HACQUES, maitre de conférences à la faculté des sciences de GRENOBLE.

Je remercie tous mes camarades du Laboratoire de calcul de l'Université de GRENOBLE pour l'aide inappréciable qu'ils m'ont apportée lors de l'élaboration de ce travail.

I° THESE

MATRICES DU 2° DEGRE

&

NORMES GENERALES

EN ANALYSE NUMERIQUE LINEAIRE

I N T R O D U C T I O N

Le Calcul Numérique présente deux aspects apparemment distincts :

la partie consacrée à la recherche de méthodes ou de procédés algorithmiques, partie algébrique ;

la partie consacrée au contrôle des calculs et des approximations, partie topologique.

Sur le plan théorique, l'Analyse Numérique présuppose donc l'algèbre et la topologie. La raison profonde en est la définition même des nombres réels en laquelle ce double point de vue se fait déjà sentir.

Ce travail se propose d'utiliser deux idées dirigées dans les deux voies dont nous venons de parler. La première est de montrer le rôle important joué dans la plupart des processus de calcul linéaire par des matrices de polynômes minimaux du second degré, dont l'inverse est, par ce fait, immédiatement calculable. La seconde est de montrer comment la considération de normes générales peut permettre de définir des notions importantes comme lesconditionnements numériques et d'obtenir un contrôle plus fin des erreurs de calcul.

Les deux premiers chapitres présentent le "matériel" théorique utilisé. Le chapitre I contient, après la détermination des matrices du 2° degré, un certain nombre de propositions prouvant que, par des produits à gauche par de telles matrices particulièrement simples, l'on peut trianguläriser, et même, en général, diagonaliser une matrice quelconque. (algorithmes de GAUSS et de JORDAN). Le chapitre II présente les notions de normes de vecteur et de matrice. Après un rappel de résultats classiques, dont certains sont dûs à OSTROWSKI, nous avons cherché à caractériser les normes symétriques de matrices associées à des normes de vecteurs. Il résulte des théorèmes (en particulier du théorème VI) que ces normes sont les plus "serrées" des normes multiplicatives possibles : donc les plus utilisables en calcul. Il est signalé deux caractérisations : l'une (théorème V) par l'intermédiaire d'une double association de norme de vecteur et de matrice, l'autre (théorème VII) par la propriété, pour ces normes, d'être lesnormes multiplicatives minimales.

Dans le chapitre III, se trouve définie la notion générale de conditionnement numérique. L'idée en est très simple : on veut trouver la solution d'une équation, c'est-à-dire trouver une inconnue x afin qu'une expression T(x) soit nulle. Puisque, en calcul numérique, et quel que soit le moyen de calcul utilisé effectivement, toute quantité inférieure à la capacité minimum de ce moyen est considérée comme nulle. Laquestion essentielle est de savoir si le fait qu'une norme de T(x) soit faible implique la proximité de x et de la solution théorique. Le conditionnement général est défini comme rapport de deux constantes qui figurent dans le théorème d'équivalence de normes sur un espace vectoriel topologique. Géométriquement, il apparaît comme une coefficient de forme, pour le lieu des x tels que une normede T(x) ait une valeur donnée. On étudie ensuite les conditionnements de système d'équations linéaires. Des comparaisons des différents types deconditionnements entre eux sont établies. Le théorème III de ce chapitre établit l'équivalence de cette définition et de celle de TURING sur les variations relatives de la solution pour des variations des coefficients.

Un conditionnement particulier C(A) et son complémentaire C'(A) sont étudiés au chapitre IV. Leur définition etleurs propriétés géométriques montrent que ces nombres, toujours compris entre zéro et un, ne sont égaux à un que si la matrice A est orthogonale en lignes. Ils peuvent donc servir à la définition d'un système "bien" ou "mal" conditionné d'une façon plus intrinsèque que les conditionnements généraux. On montre comment lescalculer et l'on étudie ensuite les opérations qui les laissent invariants. Des exemples de calcul de conditionnement C(A) sont donnés : matrice d'HILBERT, matrice d'interpolation d'opérateur différentiel. Dans ce dernier cas, on prouve un résultat bien souvent constaté en calcul : si l'on remplace un problèmediffé-rentiel aux limites par un système linéaire dû à une interpolation de l'opérateur différentiel, plus l'on augmente la précision en diminuant le pas, plus le système linéaire voit son conditionnement normalisé diminuer.

L'étude des erreurs dans la résolution des systèmes linéaires par élimination est faite dans le chapitre V. Nous établissons la forme de la "matrice d'erreur" en utilisant les matrices du 2° degré que l'on a déterminées au chapitre I, aussi bien pour un calcul en virgule fixe qu'en virgule flottante. La matrice d'erreur pourl'inversion en virgule fixe d'une matrice est aussi déterminée. Le théorème VIII indique, en fonction des "matrices d'erreur" et du "second membre d'erreur" la forme générale de la double inégalité à laquelle satisfait une norme de l'erreur absolue, ce qui met en évidence le rôle important joué par le conditionnement de la matrice du premier membre.

Le théorème IX indique une condition suffisante pourchoisir la capacitéd'un calcul en virgule fixe, afin d'obtenir une précision désirée sur la solution. La forme de la matrice d'erreur montre comment il faut choisir les différents "pivots" dans une résolution en virgule fixe. Des expériences de calcul viennent confirmer cette règle. En virgule flottante, l'on obtient des résultats analogues. Enfin, nous indiquons comment faire en sorte qu'une élimination ne fasse pas diminuer le conditionnement d'un système.

Le chapitre VI traite de la méthode de résolution par orthogonalisation. Une forme de matrice d'erreur est encore établie. Une règle de choix des différentes variantes dans ces calculs est énoncée.

Parl'intermédiaire de la notion de décomposition d'une norme, nous montrons, au chapitre VII, que l'on peut trouver différents procédés de résolution itératifs d'un système linéaire dont la convergence est directement fonction des conditionnements étudiés. L'un est original, les deux autres sont des variations sur les méthodes de gradient, très fréquemment utilisées.

Le chapitre VIII est consacré à l'étude des éléments caractéristiques des matrices. En utilisanttoujours des matrices du 2° degré, nous obtenons une forme particulièrement simple de transmutation. Deux méthodes de détermination du polynôme caractéristique sont ensuite exposées. La première est basée sur la formule (I) du chapitre I, § 3-d, et permet d'obtenir, en général, une matrice semblable à la matrice donnée de la forme deFROBENIUS. La seconde recherche une matrice semblable, de forme triple-diagonale. Enfin, l'on indique comment obtenir une matrice d'erreur, pour la première méthode, parun raisonnement assez analogue à celui qui est fait au chapitre V. Nous terminons en indiquant comment exploiter cette matrice pour obtenir les variations correspondantes des éléments caractéristiques de la matrice donnée.

CHAPITRE I

Matrices du 2° degré

I°) Anneaux de matrices du 2° degré

Définition : On dit qu'une matrice est du 2° degré si son polynôme minimal $m(u)$, est du 2° degré; nous écrirons ce polynôme :

$$m(u) = u^2 - pu - q \quad .$$

A partir d'une matrice K du 2° degré, de type (n,n), à éléments appartenant au corps C des nombres complexes, de polynôme minimal $m(u)$, formons l'ensemble $\mathcal{A}_K$ des matrices $Z = a\,I + b\,K$ avec $a,b \in C$, I unité de type (n,n).

Théorème I : $\mathcal{A}_K$ est un anneau de matrices. Une condition nécéssaire et suffisante pour que Z soit inversible dans cet anneau est que, pour $b \neq 0$, $m(\mu) \neq 0$, si $-\mu = a/b$. L'inverse de $Z = a\,I + b\,K$ est alors $Z' = a'I + b'K$ avec :

$$(1) \quad \begin{cases} a' = \dfrac{1}{b} \dfrac{p-\mu}{m(\mu)} \\ b' = -\dfrac{1}{b} \dfrac{1}{m(\mu)} \end{cases}$$

En effet, si $Z = aI + bK$, $Z' = a'I + b'K$, $Z + Z' = (a+a')I + (b+b')K$ prouve que l'addition dans $\mathcal{A}_K$ est une loi de groupe commutatif.

$$Z.Z' = (aI+bK).(a'I+b'K) = aa'I + (ab'+ba')K + bb'K^2$$
$$= (aa'+qbb')I + (ab'+ba'+pbb')K$$

puisque $K^2 = pK + qI$, qui est bien élément de $\mathcal{A}_K$.

La matrice Z est donc inversible si et seulement si

$$a a' + q b b' = 1$$
$$a b' + b a' + p b b' = 0$$

c'est à dire pour a,b tels que :

$$\text{Det} \begin{vmatrix} a & qb \\ b & a+pb \end{vmatrix} \neq 0$$

ou : $a^2 + pab - qb^2 \neq 0$

Pour a,b,p,q donnés, a', b' se déterminent d'une manière unique par :

$$a' = \frac{a+pb}{a^2+pab-qb^2} \quad , \qquad b' = \frac{-b}{a^2+pab-qb^2}$$

Pour $b \neq 0$ ces formules sont les formules (I).

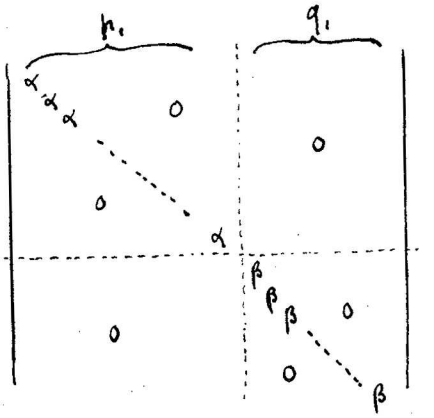
C'est la facilité du calcul de l'inverse d'une matrice Z qui fait l'importance de telles matrices en analyse numérique. Ces matrices sont construites à partir de matrices K, nous allons voir comment les déterminer.

2°) Détermination des matrices du 2° degré, de type (n,n) sur le corps C

La théorie classique de la réduction à la forme de JORDAN [I], nous conduit à décomposer $m(u) = u^2 - pu - q$ en $m(u) = (u-\alpha)(u-\beta)$ dans C.

Deux cas sont possibles:

a) $\alpha \neq \beta$. Le polynôme minimal n'ayant que des facteurs irréductibles du premier degré, la matrice K de polynôme minimal m(u) n'est autre que la transmuée par une matrice T, non singulière, d'une matrice diagonale D de la forme:

$D =$ 

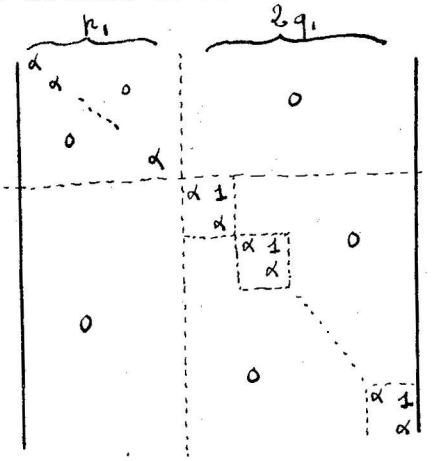
$p_1 + q_1 = n$

Toute matrice $K = T^{-1} D\, T$ convient.

b) $\alpha = \beta$. Le polynôme minimal est $m(u) = (u-\alpha)^2$. La chaîne des diviseurs élémentaires sera de la forme:

$$\underbrace{(u-\alpha)^2, \ldots, (u-\alpha)^2}_{q_1}, \quad \underbrace{(u-\alpha), \ldots (u-\alpha)}_{p_1}.$$

Par suite, la matrice K n'est autre que la transmuée par une matrice T non singuliére d'une matrice de JORDAN de la forme:

$J =$ 

$p_1 + 2q_1 = n$

Toute matrice $K = T^{-1} J\, T$ convient.

Définition : On nommera matrice antiscalaire une matrice :

$$K = X.Y^T \qquad X \text{ et } Y \text{ colonnes non nulles}$$

Théorème II : Toute matrice antiscalaire est semblable à l'une des matrices :

$$D = \begin{vmatrix} \lambda & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{vmatrix}, \lambda \neq 0 \quad ; \qquad J = \begin{vmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{vmatrix}.$$

et réciproquement.

1°) Pour la propriété directe, on écrira :

$$K = X\,Y^T \qquad K^2 = X(Y^T.X)\,Y^T = \lambda.X\,Y^T$$

puisque $Y^T X = \lambda$ est un scalaire.

Il en résulte : $K^2 - \lambda K = 0$

Soit U un vecteur colonne. $K\,U = 0$ est vérifié si $Y^T U = 0$, donc pour des vecteurs formant un sous-espace de dimension n-1. Il en résulte que 0 est racine d'ordre au moins n-1 pour l'équation caractéristique. Si 0 est racine d'ordre n-1, la forme réduite de JORDAN est D. Si 0 est racine d'ordre n, la matrice est nulle (ce qui exige $X = 0$ ou $Y = 0$, et se trouve donc exclu) ou bien la forme réduite de JORDAN est J.

2°) Réciproquement,

Supposons que K soit semblable à D ou J. Il existe une matrice T inversible telle que

$$K = T^{-1}\,D\,T \quad \text{ou} \quad K = T^{-1}\,J\,T.$$

Posons $T = (t_{ij})$ et $T^{-1} = (\tau_{ij})$ pour $i,j = 1, \ldots\ldots\ldots, n$

Alors :

$$T^{-1} D T = \begin{vmatrix} \tau_{11} & \cdots & \tau_{1n} \\ \tau_{21} & \cdots & \tau_{2n} \\ \tau_{31} & & \tau_{3n} \\ \vdots & & \vdots \\ \tau_{n1} & \cdots & \tau_{nn} \end{vmatrix} \cdot \begin{vmatrix} \lambda & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{vmatrix} \cdot \begin{vmatrix} t_{11} & \cdots & t_{1n} \\ t_{21} & \cdots & t_{2n} \\ t_{31} & & t_{3n} \\ \vdots & & \vdots \\ t_{n1} & \cdots & t_{nn} \end{vmatrix}$$

$$= \lambda \cdot \begin{vmatrix} \tau_{11}t_{11} & \tau_{11}t_{12} & \cdots & \tau_{11}t_{1n} \\ \tau_{21}t_{11} & \tau_{21}t_{12} & \cdots & \tau_{21}t_{1n} \\ \vdots & & & \\ \tau_{n1}t_{11} & & \cdots & \tau_{n1}t_{1n} \end{vmatrix} = U \circ V^T \quad , \text{ si } U = \lambda \begin{vmatrix} \tau_{11} \\ \tau_{21} \\ \vdots \\ \tau_{n1} \end{vmatrix} \qquad V = \begin{vmatrix} t_{11} \\ t_{12} \\ \vdots \\ t_{1n} \end{vmatrix}$$

Ou bien :

$$T^{-1} J T = \begin{vmatrix} \tau_{11} & \cdots & \tau_{1n} \\ \tau_{21} & & \vdots \\ \vdots & & \\ \tau_{n1} & & \tau_{nn} \end{vmatrix} \cdot \begin{vmatrix} 0 & 1 & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{vmatrix} \cdot \begin{vmatrix} t_{11} & \cdots & t_{1n} \\ t_{21} & & \vdots \\ \vdots & & \\ t_{n1} & & t_{nn} \end{vmatrix} = \begin{vmatrix} \tau_{11}t_{21} & \tau_{11}t_{22} & \cdots & \tau_{11}t_{2n} \\ \tau_{21}t_{21} & & & \\ \vdots & & & \\ \tau_{n1}t_{21} & & & \tau_{n1}t_{2n} \end{vmatrix}$$

$$= U \circ V^T \quad , \qquad \text{si } U = \begin{vmatrix} \tau_{11} \\ \tau_{21} \\ \vdots \\ \tau_{n1} \end{vmatrix} \qquad V = \begin{vmatrix} t_{21} \\ t_{22} \\ \vdots \\ t_{2n} \end{vmatrix}$$

Dans les deux cas $K = X \circ Y^T$

On distinguera les deux formes par :

1°) $V^T \circ U = 0$ pour la forme J

2°) $V^T \circ U \neq 0$ pour la forme D

Nous allons montrer quelques propriétés importantes des matrices $X.Y^T$.

3°) Utilisation de matrices $X.Y^T$, antiscalaires.

— a) Décomposition d'une matrice quelconque en une somme de matrices antiscalaires.

Soient e_i pour i=I.....n. les vecteurs de la base fondamentale de C^n, c'est à dire les vecteurs-colonnes

$$e_i = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{vmatrix} \leftarrow i°$$

On voit que les matrices unitaires E_{ij} ,base de l'espace à n^2 dimensions des matrices carrées sur C,

$$E_{ij} = \begin{vmatrix} & & j° & \\ & 0 & \vdots & 0 \\ i° \rightarrow & \cdots\cdots & 1 & \cdots\cdots \\ & 0 & \vdots & 0 \end{vmatrix}$$

ne sont autres que les produits antiscalaires :

$$E_{ij} = e_i . e_j^T$$

Soit alors u_k , k=I.....n ,un système de n vecteurs linéairement indépendants, puisqu'ils forment une base de C^n, on peut écrire:

$$e_i = \sum_{k=1}^{k=n} \beta_{ki}\, u_k$$

et,puisque si A est une matrice quelconque de terme général (a_{ij}) , i,j=I.....n , $A = \sum_{i,j} a_{ij} E_{ij}$

D'où $$A = \sum_{i,j} a_{ij} \left(\sum_k \beta_{ki} u_k\right)\left(\sum_{k'} \beta_{k'j} u_{k'}^T\right) = \sum_{\substack{m,e \\ =1,\dots n}} \gamma_{m,e}\, u_m . u_e^T$$

Proposition I: Etant donné un système de n vecteurs linéairement indépendants, u_k , k=I.....n, toute matrice peut être écrite sous la forme

$$A = \sum_{m,e=1,\dots n} \gamma_{m,e}\, u_m . u_e^T$$

c'est à dire sous la forme d'une somme de matrices de type $X.Y^T$.

Nous verrons ,à la fin de ce chapitre une proposition sur cette décomposition un peu moins générale.

b) L'algorithme d'élimination de GAUSS.

Proposition II: Etant donnée une matrice A quelconque,il existe une suite de matrices du 2° degré, G_ℓ ,en nombre fini,telles que le produit

$$T = G_1 G_2 G_3 \ldots\ldots\ldots G_m A$$

soit une matrice triangulaire supérieure.On peut choisir les G_ℓ de sorte que Det(A) = Produit des éléments diagonaux de T.

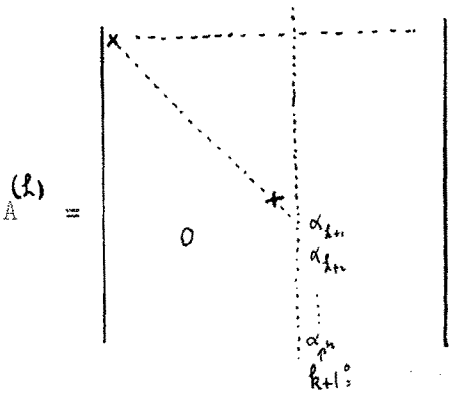
Nous allons utiliser des matrices de type $Z_{ij} = I + \lambda E_{ij}$ avec $\lambda \in C$, E_{ij} antiscalaire Il est clair que se sont bien des matrices du 2° degré, dontl'inverse, pour i≠j,n'est autre que $Z'_{ij} = I - \lambda E_{ij}$.

L'effet produit par multiplication àdroite ou à gauche d'une matrice quelconque par Z est bien connu .

M.Z = Matrice M où la j°colonne est remplacée par la somme de celle-ci et de la i°,multipliée par λ .

Z.M = Matrice M où la i°ligne est remplacée par la somme de celle-ci et de la j°,multipliée par λ . [3]

Soit alors une matrice $A^{(k)}$ de la forme ci-dessous:



c'est à dire ayant des 0 dans ses k premières colonnes et au dessous de la diagonale . Soient $\alpha_{k+1}, \alpha_{k+2}, \ldots \alpha_n$, les termes dans la k+I° colonne sur la diagonale et au dessous.

a) Supposons $\alpha_{k+1} \neq 0$. Il est clair que si l'on considère

$$G_{k+1,k+2} = I - \frac{\alpha_{k+2}}{\alpha_{k+1}} E_{k+2,k+1}$$

$$\vdots$$

$$G_{k+1,n} = I - \frac{\alpha_n}{\alpha_{k+1}} E_{n,k+1}$$

le produit $G_{k+1,n} \cdot G_{k+1,n-1} \ldots G_{k+1,k+2} \cdot A^{(k)}$ est de la forme : $A^{(k+1)}$ c'est à dire n'a que des 0 dans ses k+I premières colonnes et au dessous de la diagonale.

Si l'on pose :

$$J^{(\ell)} = \begin{vmatrix} 1 & & & \overset{k+1^\circ}{} & & \\ & 1 & & \vdots & & \\ & & 1 \ddots & \vdots & 0 & \\ & 0 & & \vdots & & \\ & & & 1 & \cdots\cdots & \leftarrow \ell+1^\circ \\ & & & -\frac{\alpha_{\ell+2}}{\alpha_{\ell+1}} & \ddots & \\ & 0 & & \vdots & & \ddots \\ & & & -\frac{\alpha_n}{\alpha_{\ell+1}} & & 1 \quad 1 \end{vmatrix}$$

On peut vérifier que $J^{(\ell)} = G_{\ell+1,n}.G_{\ell+1,n-1}.\ldots\ldots.G_{\ell+1,\ell+2}$. précédents.
Cette forme nous sera ûtile pour la suite. On peut remarquer aussi que J est aussi du 2° ordre et de la forme $I+X.Y^T$,avec

$$X = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -\frac{\alpha_{\ell+2}}{\alpha_{\ell+1}} \\ -\frac{\alpha_{\ell+3}}{\alpha_{\ell+1}} \\ \vdots \\ -\frac{\alpha_n}{\alpha_{\ell+1}} \end{vmatrix} \leftarrow \ell+2^\circ \qquad\qquad Y = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{vmatrix} \leftarrow \ell+1^\circ \; = \; e_{\ell+1}$$

b) Si $\alpha_{\ell+1}=0$. Ou bien tous les éléments de la k+I° colonne au dessous de la diagonale sont nuls: $A^{(k)}$ est de type $A^{(\ell+1)}$.
Ou bien il existe $\alpha_\ell \neq 0$ dans cette k+I° colonne
Soit $V_{ij} = I - E_{ii} - E_{jj} + E_{ij} + E_{ji}$,cette matrice multipliant à gauche une matrice M quelconque donne une matrice identique à M à cela près que les lignes i et j sont échangées.(En multipliant à droite on échange les i° et j°colonnes) D'ou $V_{ij}^2 = I$, V est encore du 2° degré.Son déterminant vaut -I. Je considère $U = I - 2E_{nn}$,qui est du 2° degré et par multiplication à droite par une matrice quelconque change les signes des éléments de la dernière ligne de M.

Alors si l'on fait $U.V_{\ell,k+1}A^{(\ell)} = A_1^{(\ell)}$

On est ramené au cas précédent.

Les remarques a) et b) étant faites, on peut poser $A^{(0)}$= A donnée,passer ainsi que l'on vient de le prouver à une $A^{(1)}$, puis $A^{(2)}$,etc......On aboutit à une $A^{(n-1)}$qui est une forme triangulaire supérieure.

Il est remarquable de voir que les $G_{\ell,i}$ du a) ou les $U.V_{n,\ell}$du b) sont des matrices de déterminant +I. Donc la matrice du produit des G_ℓ de l'énoncé a un déterminant +I.Cela prouve la fin de la proposition.

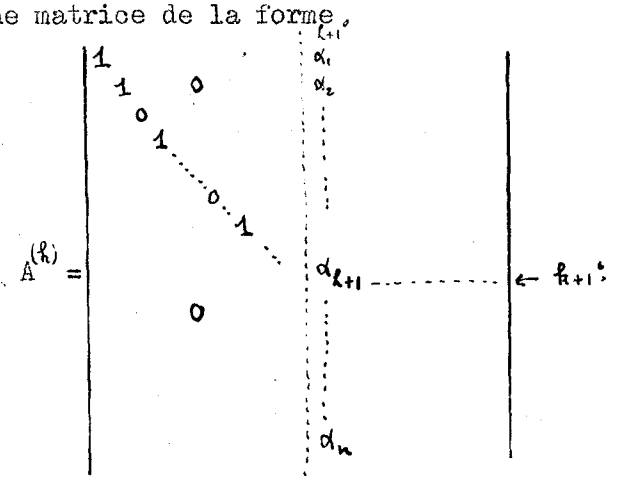
<u>Remarque</u> : Dans l'utilisation pratique de l'algorithme le cas b) est vraiment exeptionnel

b) L'algorithme d'élimination de JORDAN

Proposition III Pour une matrice A quelconque il existe toujours un nombre fini de matrices du 2° degré telles que A en soit le produit. Dans le cas général, A est le produit de n matrices du 2° degré de la forme $\left(I + X.e_k^T\right)^{-1}$ [2]

La propriété, comme la précédente, provient de la théorie de l'élimination, selon JORDAN, cette fois.

Soit une matrice de la forme,



c'est à dire ayant des 0 dans ses k premières colonnes, sauf certains éléments de la diagonale qui peuvent être des 1.

$\alpha_1, \alpha_2, \ldots, \alpha_{k+1}, \ldots \alpha_n$, désignent les éléments de sa k+I° colonne

a) Supposons $\alpha_{k+1} \neq 0$. Considérons les matrices

$$G_{k+1,1} = I - \frac{\alpha_1}{\alpha_{k+1}} E_{1,k+1}$$

$$\vdots$$

$$G_{k+1,k+1} = I - \left(1 - \frac{1}{\alpha_{k+1}}\right) E_{k+1,k+1}$$

$$\vdots$$

$$G_{k+1,n} = I - \frac{\alpha_n}{\alpha_{k+1}} E_{n,k+1}$$

Il est clair que $G_{k+1,1} G_{k+1,2} G_{k+1,3} \ldots\ldots\ldots G_{k+1,n} A^{(k)} = A^{(k+1)}$

c'est à dire a des 0 dans ses k+I°colonnes, sauf en sa diagonale où il y a des I

b) Si $\alpha_{k+1} = 0$ Ou bien tous les éléments de la k+I°colonne sont nuls, $A^{(k)}$ est de la forme $A^{(k+1)}$. Ou bien il existe $\alpha_\ell \neq 0$ dans cette colonne. Dans ce cas je multiplie par $V_{k+1,\ell}$ et je suis ramené au cas précédent.

En conclusion, par une suite de produits à gauche par des matrices du 2° degré on peut toujours passer de $A^{(k)}$ à $A^{(k+1)}$. Par suite, partant de $A^{(0)} = A$ donnée, on passe à $A^{(1)}$, puis à $A^{(2)}$, etc.....$A^{(n)}$ est soit l'unité I, soit l'unité dans la quelle des 0 figurent dans la diagonale. Soit J une telle matrice.

Il est clair que toutes les matrices G ou V envisagées sont du 2° degré et inversibles. Donc, de $G_{e,m} . G_{e',m'} \ldots V^{(1)} \ldots G_{e,1} . A = J$

On tire: $A = G_{e,1}^{-1} \ldots V^{(1)-1} \ldots G_{e',m'}^{-1} . G_{e,m}^{-1} J.$

Remarquons pour finir que J est du 2° degré, car $J^2 = J$; cela nous donne la démonstration de la 1° partie de la proposition.

Remarquons aussi que, dans le cas général, où b) ne se produit pas, le passage $A^{(k)}$ vers $A^{(k+1)}$ se fait par multiplication par la matrice $J^{(k)}$:

$$J^{(k)} = \begin{vmatrix} 1 & & & -\frac{\alpha_1}{\alpha_{k+1}} & & \\ & 1 & & \vdots & & \\ & & 1 & & & \\ & & & \frac{1}{\alpha_{k+1}} & & \\ & & & -\frac{\alpha_{k+2}}{\alpha_{k+1}} & 1 & \\ & & & \vdots & & \ddots \\ & & & -\frac{\alpha_n}{\alpha_{k+1}} & & 1 \end{vmatrix} \leftarrow (k+1)^{\circ}$$

qui n'est autre que le produit $J = G_{k+1,1}\, G_{k+1,2}\, G_{k+1,3} \cdots\cdots\cdots G_{k+1,n}$, dans ce cas.

Or cette matrice peut s'écrire:

$$J^{(k)} = I + X_{k+1} e_{k+1}^{\top}$$

avec:

$$X = \begin{vmatrix} -\frac{\alpha_1}{\alpha_{k+1}} \\ \vdots \\ -\frac{\alpha_k}{\alpha_{k+1}} \\ \frac{1}{\alpha_{k+1}} - 1 \\ \vdots \\ -\frac{\alpha_n}{\alpha_{k+1}} \end{vmatrix}, \qquad e_{k+1} = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{vmatrix} \leftarrow (k+1)$$

Cela prouve que dans le cas général,

$$A = \left(I + X_1 \cdot e_1^{\top}\right)^{-1} \cdot \left(I + X_2 \cdot e_2^{\top}\right)^{-1} \cdot \cdots\cdots \left(I + X_n \cdot e_n^{\top}\right)^{-1} \quad ?$$

Nous trouverons usage de toutes ces propriétés par la suite.

c) L'orthogonalisation en "lignes" et les matrices du 2nd degré.

Soit $A^{(k)}$ unematrice à éléments réels, que nous écrirons :

$$A^{(k)} = \begin{vmatrix} h_1 \\ h_2 \\ \vdots \\ h_k \\ a_{k+1} \\ \vdots \\ a_n \end{vmatrix}$$

où les $h_1, \dots h_k, a_{k+1}, \dots, a_n$, désignent les différentes lignes de $A^{(k)}$ (par la suite, ces symboles seront compris comme des vecteurs-lignes) et supposons que les k premières lignes de $A^{(k)}$ soient orthogonales, c'est à dire:

$$h_i \cdot h_j^T = 0$$

pour $i \neq j$, $(i,j = 1, 2 \dots k)$

Considérons alors la matrice:

$$S_k = \begin{array}{|cccccc|} 1 & & & & & \\ & 1 & & & 0 & \\ & & 1 & & & \\ & & & \ddots & & \\ & 0 & & & & \\ \leftarrow k+1 & \lambda_{k+1,1} & \cdots & \lambda_{k+1,k} & 1 & 0 \\ & & & & & 1 \\ & 0 & & & & \ddots \\ & & & & & 1 \end{array}$$

ou $S_k = I + \Lambda_k$ avec:

$$\Lambda_k = \begin{array}{|ccccc|} & & & k+1 & \\ & & & & \\ & 0 & & & \\ & & & & \\ \leftarrow k+1 & \lambda_{k+1,1} & \cdots & \lambda_{k+1,k} & 0 \cdots \\ & & & & \\ & 0 & & & \end{array}$$

on voit immédiatement que $\Lambda_k^2 = 0$, donc, S_k , est du 2° degré et l'inverse de S_k n'est autre que $I - \Lambda_k$

Notons aussi l'analogie entre ces matrices et les J_k (S_{k-1} et J_{k-k} sont en quelque sorte "symétriques" par rapport à la 2° diagonale) .
(une // à la)

Cherchons le produit $S_k \cdot A^{(k)}$, on peut l'écrire enlignes:

$$S_k \cdot A^{(k)} = \left| \begin{array}{l} h_1 \\ h_2 \\ \vdots \\ h_k \\ a_{k+1} + \lambda_{k+1,1} h_1 + \lambda_{k+1,2} h_2 + \dots + \lambda_{k+1,k} h_k = h_{k+1} \\ a_{k+2} \\ \vdots \\ a_n \end{array} \right.$$

Le produit ne diffère de $A^{(k)}$ que par sa $k+1^{\circ}$ ligne.

Peut-on choisir les $\lambda_{k+1,1}, \lambda_{k+1,2} \ldots\ldots \lambda_{k+1,k}$, de sorte que cette nouvelle $k+1^{\circ}$ ligne soit orthogonale aux k premières ? Il suffit d'écrire:

$$h_{k+1} \cdot h_i^T = 0 \qquad (i = 1, 2, \ldots, k)$$

et en tenant compte de l'orthogonalité des k premières lignes entr-elles, il reste: $a_{k+1} \cdot h_i^T + \lambda_{k+1,i} \cdot h_i \cdot h_i^T = 0$

et si l'on suppose (matrices réelles) qu'aucune des k premières lignes n'est nulle on obtient:

$$\lambda_{k+1,i} = - \frac{a_{k+1} \cdot h_i^T}{h_i \cdot h_i^T} \qquad (i = 1, 2, \ldots, k) .$$

Soit alors une matrice A à élément réels, non singulière,

$$A = \begin{vmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{vmatrix} = (a_{ij}) \qquad (i, j = 1, \ldots n)$$

Posons $A^{(1)} = A$, et considérons la suite des matrices:

$$A^{(2)} = S_1 \cdot A^{(1)}, \ldots\ldots, A^{(k)} = S_{k-1} \cdot A^{(k-1)}, \ldots, A^{(n)} = S_{n-1} \cdot A^{(n-1)}$$

Il est clair que:

$$A^{(k)} = S_{k-1} \cdot S_{k-2} \ldots\ldots S_1 \cdot A$$

Puisque les déterminants de toutes les matrices S_k sont égaux à 1 la non-singularité de A impose la non-singularité des $A^{(k)}$, pour tout k donc quels que soient les λ_{ij}, il est impossible qu'une des k premières lignes de $A^{(k)}$ soit nulle, cela permet de voir que pour A non singulière le procédé suivant, dit de SCHMIDT, permet l'orhogonalisation:

Partant de $A^{(1)} = A$ on détermine les λ_{ij} par

$$\lambda_{ij} = - \frac{a_i \cdot h_j^T}{h_j \cdot h_j^T} .$$

il est clair que $A^{(n)} = H$ est orthogonale en lignes.

D'où:

Proposition IV Pour une matrice A, non singulière il existe une suite, $S_1, \ldots\ldots, S_{n-1}$ de matrices du 2° degré, de la forme indiquée, telle que le produit

$$H = S_{n-1} \cdot S_{n-2} \ldots\ldots S_1 \cdot A$$

soit orthogonal en lignes.

d) Tansmutations par des matrices du 2° degré .

. La plus part des problèmes de détermination numérique de valeurs propres ou de vecteurs propres sont résolus en transmuant la matrice donnée A en A' , avec $A' = M.A.M^{-1}$ et M non singulière . si l'on prend pour M, des matrices non simples le travail numérique est considérable et introduit une grosse quantité d'erreurs , d'où l'intéret de chercher des M et M^{-1} pour les quelles il n'y a pas d'erreur de calcul.

Or si $M = \alpha.I + \beta.K \in \mathcal{Q}_K$

$$M^{-1} = \alpha' I + \beta' K$$

avec :

$$\alpha' = \frac{1}{\beta} \cdot \frac{p - p}{m(p)}$$

$$\beta' = -\frac{1}{\beta} \cdot \frac{1}{m(p)}$$

Par suite:

$$A' = (\alpha I + \beta K).A.(\alpha' I + \beta' K) = \alpha\alpha' H + \alpha'\beta.KA + \alpha\beta' AK + \beta\beta' K.A.K$$

et enfin :

$$(I) \quad \boxed{A' = \frac{p^2 - pp}{m(p)}.A + \frac{p - p}{m(p)} K.A + \frac{p}{m(p)} A.K - \frac{1}{m(p)} K.A.K}$$

Première application : Matrices - Test :

Afin de vérifier, en calcul numérique, la qualité d'une méthode de détermination des valeurs et vecteurs propres , il est indispensable de traiter par cette méthode des matrices dont on connaisse à l'avance les valeurs propres et les vecteurs propres .

Les matrices antiscalaires permettent d'écrire un programme capable de faire sortir , sous forme de données immédiatement ré-utilisables , les éléments d'une matrice dont on se donne à l'avance l'ordre n , la suite $\lambda_1, \lambda_2, \ldots\ldots \lambda_n$, des valeurs propres réelles, mais pas forcement distinctes , et dont on connaisse les colonnes propres et les lignes propres. ?

En effet, soit :

$$M = B.\Lambda.B^{-1}$$

avec :

$$\Lambda = \begin{vmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{vmatrix} \quad , \quad B = \begin{vmatrix} \beta & 1 & 1 & \cdots & 1 \\ 1 & \beta & 1 & \cdots & 1 \\ 1 & 1 & \beta & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & \beta \end{vmatrix}$$

alors :

$$B = (\beta - 1)I + K \qquad \text{si:} \quad K = \sum_{i,j} E_{ij} = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & & & & \vdots \\ 1 & & & & 1 \end{vmatrix} = X.X^T \quad , \quad X = \begin{vmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{vmatrix}$$

Il vient pour : $M = (m_{ij}) \quad (i,j = 1, \ldots n)$

$$m_{ij} = \delta_{ij}.\lambda_j + \frac{1}{\beta - 1}\lambda_j - \frac{1}{\beta - 1 + n}\lambda_i - \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_n}{(\beta - 1)(\beta - 1 + n)}$$

Pour cette matrice, les valeurs propres sont $\lambda_1, \lambda_2, \ldots \lambda_n$ données, et à la valeur propre λ_i sont associées les colonnes et lignes propres :

$$\lambda_i \longleftrightarrow u_i = \begin{vmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ \beta \leftarrow i^{\circ} \\ \vdots \\ 1 \end{vmatrix} \quad , \quad v_i = \underset{\downarrow\, i^{\circ}}{(-1, -1, \ldots, \beta + n - 2, \ldots, -1)} \cdot \frac{1}{(\beta - 1)(\beta - 1 + n)}$$

Remarquons que faisant varier β de $+1$ à $+\infty$ on obtient un n-èdre des vecteurs propres qui varie assez régulièrement de la "fermeture" à l'orthogonalité . Enfin pour $\lambda_1, \ldots \lambda_n ; n$, donnés, il y a une seule matrice qui soit symétrique , elle correspond à un cas où $\beta = 1 - \frac{n}{2}$.

CHAPITRE I I

Normes de vecteurs et de matrices

I°) Normes de vecteurs. [4]

Etant donné un espace vectoriel $\mathcal{E}$ sur R (ou C), on appelle norme sur cet espace une application $x \longrightarrow p(x)$ de celui-ci dans R_+ telle que:

$$\begin{cases} \text{I)} & p(x)=0 \text{ est équivalent à } x=0 \\ \text{II)} & p(\lambda x) = |\lambda| p(x) \quad \text{pourtout } \lambda \in R \text{ (ouC)} \\ \text{III)} & p(x+y) \leqslant p(x) + p(y) \end{cases}$$

quels que soient les éléments x,y de cet espace vectoriel.

Soit donc sur l'espace vectoriel $\mathcal{E}$ une norme p(x). Donnons nous, de plus une application linéaire T(x) de $\mathcal{E}$ dans lui-même que nous supposerons complet

A $x \in \mathcal{E}$ faisons correspondre: $q(x)=p(T(x))$; c'est une application de $\mathcal{E}$ dans R_+. Deux questions se posent:

a) Est-elle une norme sur $\mathcal{E}$?

b) Est-elle équivalente à p(x)?

Proposition I Pour que q(x) soit une norme sur $\mathcal{E}$, il faut et il suffit que T(x) ne soit pas singulière (c'est à dire qu'il n'y ait pas d'éléments $\neq 0$ de tels que $T(x)=0$).

En effet à cause de la linéarité de T(x) on a:

$$\text{II)} \quad q(\lambda x) = p[T(\lambda x)] = p[\lambda T(x)] = |\lambda| p(T(x)) = |\lambda| \cdot q(x)$$

puis,

$$\text{III)} \quad q(x+y) = p[T(x+y)] = p[T(x)+T(y)] \leqslant p[T(x)] + p[T(y)] \leqslant q(x) + q(y)$$

cela prouve II et III pour q(x).

La condition I résulte du fait que p(T(x))=0 est équivalent à T(x)=0, et que cela ne peut être, à son tour, équivalent à x=0 que si et seulement si T(x) n'est pas singulière.

PropositionII Pour que la norme q(x) soit équivalente à p(x), il faut et il suffit que T(x) soit continue dans $\mathcal{E}$.

I) La condition est suffisante. Supposons que T(x) soit continue. Elle l'est pour 0 de $\mathcal{E}$. Si l'on se donne la boule $B_q(\beta)$, ensemble des x tels que $q(x)=p(T(x)) \leqslant \beta$, il existe α tel que

$$p(x) \leqslant \alpha \Longrightarrow \quad p(T(x)) \leqslant \beta \qquad \text{ou} \quad B_q(\beta) \supset B_p(\alpha)$$

donc la topologie définie par q n'est pas plus fine que celle définie par p.

Inversement, soit $B_p(\alpha)$ donnée, je dis qu'il existe un β tel que si $x \in B_q(\beta)$ cela implique $x \in B_p(\alpha)$. Il suffit de prouver un a positif tel que quelque soit x : $p(x) \leqslant a\,q(x)$. Or cela résulte d'un corollaire [5] d'un théorème de BANACH [6] indiquant que toute application linéaire continue et biunivoque d'un espace vectoriel normé et complet dans lui-même est un isomorphisme. Donc, si l'on prend $\beta = \frac{\alpha}{a}$ On est sûr que $B_q(\frac{\alpha}{a}) \subset B_p(\alpha)$, la topologie définie par p est donc moins fine que celle définie par q. D'après cela les deux normes p et q sont bien équivalentes.

2) La condition est nécéssaire: Si p et q sont équivalentes c'est qu'il existe deux constantes positives a et b, telles que [7] :

$$a\,p(x) \leqslant p(T(x)) \leqslant b\,p(x)$$

Si ε est donné, il suffit de prendre $p(x) \leqslant \frac{\varepsilon}{b}$ pour assurer que $p(T(x)) \leqslant \varepsilon$ d'où la continuité en 0, qui implique la continuité en tout point.

Proposition III. Si T(x) est une application linéaire continue et biunivoque d'un espace vectoriel normé et complet Ξ sur lui-même et $p_1(x)$ et $p_2(x)$ deux normes équivalentes sur Ξ. Il existe alors deux constantes a et b telles que quel que soit x

$$a\,p_1(x) \leqslant p_2(T(x)) \leqslant b\,p_1(x) \quad .$$

Il suffit de remarquer que par la propositionII les deux normes $p_1(x)$ et $p_2(T(x))$ sont équivalentes.

Cette propriété est la base théorique de la définition des conditionnements numériques que nous étudierons plus loin.

Après avoir rappelé des propriétés relatives à des espaces vectoriels généraux, nous nous intéresserons dans la suite à des espaces de dimensions finies.

En analyse numérique l'espace Ξ est toujours R^n (ou C^n à la rigueur) et trois normes de vecteurs sont à envisager:

Si $x \in R^n$ a pour éléments (ξ_i) et l'on pose $\varphi_r(x) = \left(\sum_{i=1}^{i=n} |\xi_i|^r \right)^{\frac{1}{r}}$

(norme de HOLDER pour r)

I) La norme $\varphi_1(x) = \Phi(x) = \sum_{i=1}^{i=n} |\xi_i|$

II) La norme $\varphi_2(x)$ ou norme " Euclidienne" $\varphi_2(x) = \|x\| = \left(\sum_{i=1}^{i=n} |\xi_i|^2 \right)^{\frac{1}{2}} = (x^T.x)^{\frac{1}{2}}$

III) La norme $\varphi_\infty(x) = \max_{i=1,\dots n} |\xi_i| = \mathfrak{M}_6(x)$

Pour ces trois normes il est important de connaître les nombres a et b tels que : $(a\,\varphi_i(x) \leqslant \varphi_j(x) \leqslant b\,\varphi_i(x))$ les voici:

I) $\|x\| \leqslant \Phi(x) \leqslant \sqrt{n}\,\|x\|$, $\mathcal{M}(x) \leqslant \Phi(x) \leqslant n.\mathcal{M}(x)$

II) $\frac{1}{\sqrt{n}}\Phi(x) \leqslant \|x\| \leqslant \Phi(x)$, $\mathcal{M}(x) \leqslant \|x\| \leqslant \sqrt{n}.\mathcal{M}(x)$

III) $\frac{1}{n}\Phi(x) \leqslant \mathcal{M}(x) \leqslant \Phi(x)$, $\frac{1}{\sqrt{n}}\|x\| \leqslant \mathcal{M}(x) \leqslant \|x\|$

Dans ces inégalités il est toujours un choix pour x qui assure une égalité.

Enfin, signalons qu'au point de vue géométrique, donner une norme $p(x)$ de vecteurs sur R^n ou sur C^n, revient à se donner un corps convexe, la boule: $B_p(a) = \{x\,;\, p(x) \leqslant a\}$, symétrique par rapport à l'origine et telle que toute demi-droite issue de O la coupe en un point distinct de O (O est intérieur à $B_p(a)$) . Toutes ces boules, si a varie, sont homothétiques de l'une $B_p(1)$ (par exemple) , dont la connaissance suffit pour définir la norme [43]

28) Normes de matrices

Soit $\mathcal{M}_n$ l'espace vectoriel à n^2 dimensions des matrices carrées sur R(ouC). Une norme générale est, comme on l'adit plus haut une application N(A) de $\mathcal{M}_n$ dans R_+ telle

$$\mathcal{G}_N \begin{cases} \text{I) } N(A)=0 \text{ équivaut à } A=0 \\ \text{II) } N(\lambda A) = |\lambda| N(A) \\ \text{III) } N(A+B) \leqslant N(A)+N(B) \end{cases}$$

quels que soient $\lambda \in R$; $A, B \in \mathcal{M}_n$.

Une Norme multiplicative , d'après OSTROWSKI [9] est une application M(A) de $\mathcal{M}_n$ dans R_+ telle

$$\mathcal{G}_M \begin{cases} \text{I) Il existe un } A_0 \text{ tel que } M(A_0) \neq 0 \\ \text{II) } M(\lambda A) = |\lambda| M(A) \\ \text{III) } M(A+B) \leqslant M(A)+M(B) \\ \text{IV) } M(A.B) \leqslant M(A).M(B) \end{cases}$$

Notations : $\mathfrak{N}$ désigne l'ensemble des normes sur $\mathfrak{M}_n$

$\mathfrak{N}.\mathfrak{M}$ désigne l'ensemble des normes multiplicatives sur $\mathfrak{M}_n$

$\mathcal{V}$ désigne l'ensemble des normes de vecteurs sur $\mathbb{R}^n$ (ou $\mathbb{C}^n$)

Théorème I Toute norme multiplicative est une norme générale dans $\mathfrak{M}_n$. [8]

- a) Soit I l'unité de $\mathfrak{M}_n$ on a/: $M(A_0.I) \leqslant M(A_0).M(I)$ (IV)
  puisque $M(A_0)>0$ cela implique $M(I) \geqslant 1$
- b) Si E_{ij} sont les matrices unitaires (telles que: $A = \sum_{i,j} \alpha_{ij} E_{ij}$)
  on a $I = \sum_{i=1}^{n} E_{ii}$
  donc $0 < 1 \leqslant \sum_i M(E_{ii})$: il est au moins i_0 tel $M(E_{i_0 i_0})>0$
- c) Mais $E_{i_0 i_0} = E_{i_0 j} E_{jj} E_{j i_0}$ donc: $M(E_{i_0 i_0}) \leqslant M(E_{i_0 j}).M(E_{jj}).M(E_{j i_0})$ $\forall\, i \neq j$
  Par suite tous les E_{jj} sont tels que: $M(E_{jj})>0$
- d) Mais $E_{ij} = E_{ji} E_{ii} E_{ij}$; $M(E_{ij}) \leqslant M(E_{ji}) M(E_{ii}) M(E_{ij})$ donc quelques soient les i, j, $M(E_{ij})>0$
  Or si $A = (\alpha_{ij})$, il est clair que : $\alpha_{ij} E_{kk} = E_{ki} A . E_{jk}$
  donc $|\alpha_{ij}| . M(E_{kk}) \leqslant M(E_{ki}) . M(A) . M(E_{jk})$

  d'où $M(A)>0$ dès qu'il existe un $\alpha_{ij} \neq 0$

Théorème II Toutes les normes sur $\mathfrak{M}_n$ (comme sur tout espace vectoriel de dimension finie) sont équivalentes . C'est à dire que, étant données lés deux normes $N_1(A)$ et $N_2(A)$, il est deux constantes positives a et b telles:

$$a \leqslant \frac{N_2(A)}{N_1(A)} \leqslant b$$

- a) La démonstration est classique: si $A = (\alpha_{ij})$, $B = (\beta_{ij})$
  $|N(A) - N(B)| \leqslant N(A-B) \leqslant \sum_{i,j} |\alpha_{ij} - \beta_{ij}| N(E_{ij}) \longrightarrow$ continuité de $N(A)$
- b) de plus si $N^*(A) = \max_{i,j} |\alpha_{ij}|$ et $\mu_1 = \sum_{i,j} N(E_{ij})$
  $N(A) \leqslant N^*(A) . \mu_1$
  D'autre part , si les (α_{ij}) sont liés qu'ils varient de sorte que $N^*(A) = 1$, le point (α_{ij}) dans l'espace à n^2 dimensions décrit la frontière compacte d'un parallèlotope, donc il existe pour un choix des α_{ij} une valeur minimum non nulle pour N(A) , soit μ' . Mais l'homogénéité de $\frac{N(A)}{N^*(A)}$ prouve que $\frac{N(A)}{N^*(A)} = N(A')$ si $A' = \frac{1}{N^*(A)} . A$, telle que $N^*(A') = 1$
  et par suite
  $$\frac{N(A)}{N^*(A)} \geqslant \mu' \quad , \quad \mu' \leqslant \frac{N(A)}{N^*(A)} \leqslant \mu_1$$
- c) Mais , de $a_1 \leqslant \frac{N_1(A)}{N^*(A)} \leqslant b_1$, $a_2 \leqslant \frac{N_2(A)}{N^*(A)} \leqslant b_2$, on tire:
  $$\frac{N_2(A)}{N_1(A)} = \frac{N_2(A)}{N^*(A)} \cdot \frac{N^*(A)}{N_1(A)} \quad , \quad \frac{a_2}{b_1} \leqslant \frac{N_2(A)}{N_1(A)} \leqslant \frac{b_1}{a_1}$$ q.e.d.

Il résulte de la démonstration du Théorème II que si $N(A)$ est donnée, il existe deux constantes α , β positives, telles que :

$$(1) \qquad \alpha \leqslant \frac{N(A)}{N^*(A)} \leqslant \beta$$

Or, on vérifie immédiatement que N^* est telle que :

$$N^*(A.B) \leqslant n\, N^*(A) \,.\, N^*(B) \quad \text{pour tout A et B d'ordre n}$$

donc :

$$N(A.B) \leqslant \frac{\beta}{\alpha^2} \,.\, n \,.\, N(A) \,.\, N(B)$$

Par suite, si $\lambda = \frac{\beta}{\alpha^2} n$, $\lambda\, N(A.B) \leqslant \lambda . N(A) \,.\, \lambda . N(B)$ (2)

puisque $\lambda\, N(A) = N_1(A)$ est une norme de matrice (2), s'écrit :

$$N_1(A.B) \leqslant N_1(A) \,.\, N_2(B) \quad \text{donc :}$$

Proposition I : Etant donné une norme générale η , il existe toujours un nombre λ tel que $\lambda\,\eta$ soit une norme multiplicative.

Nous allons voir une autre construction de normes multiplicatives différente au paragraphe 3°).

Signalons enfin une autre propriété des normes multiplicatives :

Théorème III : Pour toute norme multiplicative $M(A)$ sur $\mathcal{E}$, on a : $M(A) \geqslant \lambda_A$, si λ_A désigne la valeur absolue maximum des valeurs propres de A . [8]

En effet, soit x un vecteur propre correspondant à la valeur propre λ , je peux écrire la relation $A.x = \lambda . x$, où, si X est la matrice (n,n) associée à x, c'est ainsi que je nomme la matrice

$$X = \begin{vmatrix} x_1 & 0 & \cdots & 0 \\ x_2 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ x_n & 0 & \cdots & 0 \end{vmatrix} \qquad \text{si} \qquad x = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{vmatrix} \qquad A.X = \lambda . X$$

Donc :

$$M(A).M(X) \geqslant \lambda\, M(X)$$

et, parsuite :

$$M(A) \geqslant \lambda_A$$

3°) Normes de matrices associées à des normes de vecteurs.

Dans ce qui va suivre, les vecteurs sont ceux de R^n (ou C^n) et les matrices sont toutes de type (n,n) sur R (ou sur C).

Soient p(x) et q(x) deux normes ; formons le rapport $\frac{p(Ax)}{q(x)}$

Posons :

$$S_{pq}(A) = \operatorname{Max}_{x \neq 0} \left(\frac{p(Ax)}{q(x)} \right)$$

Théorème IV $S_{pq}(A)$ est une norme sur $\mathfrak{M}_n$. Si $p(x) = q(x)$, cette norme est multiplicative.

Prouvons d'abord que, quel que soient p et q, c'est une norme.

N 1) est satisfait car si $S_{pq}(A) = 0$, c'est que, quelque soit x, $p(Ax)=0$; $A = 0$

N 11) est satisfait car

$$S_{pq}(\lambda A) = \operatorname{Max}_{x \neq 0} \left(\frac{p(\lambda Ax)}{q(x)} \right) = |\lambda| \operatorname{Max}_{x \neq 0} \left(\frac{p(Ax)}{q(x)} \right) = \lambda . S_{pq}(A)$$

N(111) est satisfait puisque :

$$\frac{p((A+B)x)}{q(x)} \leqslant \frac{p(Ax)}{q(x)} + \frac{p(Bx)}{q(x)} \leqslant S_{pq}(A) + S_{pq}(B)$$

donc $S_{pq}(A)$ est bien une norme.

Montrons que S_{pq} n'est pas multiplicative en général.

Si p(x) est une norme, $\frac{1}{C}\, p(x)$ en est une autre, prenons $q(x) = \frac{1}{C}\, p(x)$

$$S_{pq}(A.B) \leqslant S_{pq}(A) \cdot S_{pq}(B)$$

exige :

$$S_{pp}(A.B) \leqslant C . S_{pp}(A) \cdot S_{pp}(B)$$

Il y a, certainement, des valeurs de C pour lesquelles cette inégalité n'est pas vérifiée.

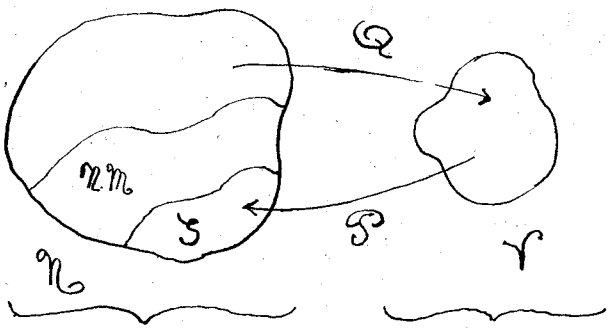
Par contre, de : $\frac{p(ABx)}{p(x)} = \frac{p(ABx)}{p(Bx)} \cdot \frac{p(Bx)}{p(x)}$

on tire $S_{pp}(AB) \leq S_{pp}(A) \cdot S_{pp}(B)$

donc $S_{pp}(A)$ est <u>multiplicative</u>

Pour résumer la situation:

Ayant une norme de vecteur, p, on peut lui faire correspondre une normê multiplicative S_{pp} de matrice. Cela définit une application $\mathcal{P}$ de $\mathcal{V}$ sur l'ensemble $\mathcal{S}$ des normes multiplicatives de la forme S_{pp} .



D'autre part, ayant une norme de matrice N(A) à $x \in R^n$, faisons correspondre la matrice X associée et posons $p_N(x) = N(X)$

Il est immédiat de voir que p_N est bien une norme de vecteur . Cela définit une application $\mathcal{Q}$ de $\mathcal{N}$ dans $\mathcal{V}$

Pour caractériser l'ensemble $\mathcal{S}$ on a le :

<u>Théorème V</u> : $\mathcal{S}$ est l'ensemble des invariants de l'application composée $\mathcal{P}\circ\mathcal{Q}$ de $\mathcal{N}$ dans lui-même.

I) Soit $S_{pp} \in \mathcal{S}$, par $\mathcal{Q}$ il correspond une norme $p_1(x)$ à cette S_{pp} • $p_1(x) = S_{pp}(X)$

Or $Xy = \eta_i x$ si $y = (\eta_i)$ $(i = 1, 2 \ldots n)$

donc

$$p_1(x) = S_{pp}(X) = \max_{y \neq 0}\left(\frac{p(\eta_i x)}{p(y)}\right) = \left(\max_{y \neq 0}\frac{|\eta_i|}{p(y)}\right) \cdot p(x) = C \cdot p(x)$$

en posant $C = \max_{y \neq 0}\frac{|\eta_i|}{p(y)}$

Par suite $S_{p_1 p_1}$ qui correspond à p_1 par $\mathcal{P}$ est telle que

$$S_{p_1 p_1}(A) = \max_{y \neq 0}\left(\frac{p_1(Ay)}{p_1(y)}\right) = \max_{y \neq 0}\left(\frac{C \cdot p(Ay)}{C \cdot p(y)}\right),$$

et par cela $S_{p_1 p_1} = S_{pp}$ donc $\mathcal{P}\circ\mathcal{Q}(S_{pp}) = S_{pp}$

II) Si une norme est invariante par $\mathcal{P}\circ\mathcal{Q}$ et puisque $\mathcal{P}$ n'a ses valeurs que dans $\mathcal{S}$, elle est du type S_{pp}

Théorème VI (OSTROWSKI): Soit une norme multiplicative M(A), si $\rho_M = Q(M)$ c'est-à-dire si $\rho_M(x) = M(X)$ et $S_{\rho_M \rho_M}$ la norme image de M par $P \circ Q$, pour toute matrice A :

$$S_{\rho_M \rho_M}(A) \leqslant M(A)$$

Autrement dit, la norme $S_{\rho_M \rho_M}$ est toujours plus "serrée" que M.

En effet

$$S_{\rho_M \rho_M}(A) = \operatorname*{Max}_{y \neq 0} \left(\frac{\rho_M(Ay)}{\rho_M(y)} \right) = \operatorname*{Max}_{y \neq 0} \left(\frac{M(Y_1)}{M(Y)} \right)$$

si Y_1 est l'associée de A.y

Or, il est clair que $Y_1 = A.Y$

donc :

$$\frac{M(AY)}{M(Y)} \leqslant M(A) \cdot \frac{M(Y)}{M(Y)}$$

et, par suite :

$$S_{\rho_M \rho_M}(A) \leqslant M(A)$$

D'après la propriété (Prop. I, parag. 2), nous posons :

Définitions : 1) Une norme multiplicative M, de matrice, est dite "critique" s'il n'y a pas de constante, $0 < k < 1$, telle que $k.M(A)$ soit encore une norme multiplicative.

2) Une norme multiplicative M de matrice est dite "minimale" s'il n'existe pas de norme multiplicative $M' \neq M$, telleque $M'(A) \leqslant M(A)$ pour tout A.

Pour toute norme multiplicative $M(I) \geqslant 1$ (Démonstration Th. I).
Donc, toute norme M telle $M(I) = 1$ est critique. Les S_{pp} sont critiques car $S_{pp}(I) = 1$. Le théorème VI d'OSTROWSKI prouve qu'il existe toujours une S_{pp} inférieure à toute norme multiplicative.

Théorème VII : Toute norme S_{pp} est une norme multiplicative minimale.

Nous allons d'abord prouver que si, pour tout A, on a :

$$S_{qq}(A) \leqslant S_{rr}(A)$$

c'est que forcément :

$$S_{rr} = S_{qq}$$

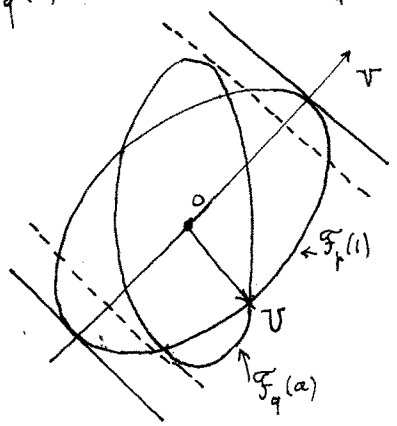
Prenons $A = U \cdot V^T$, c'est-à-dire matrice antiscalaire où U et V sont deux colonnes quelconques de R^n.

On a, par définition :

$$\operatorname*{Max}_{x \neq 0} \left(\frac{|V^T x|}{p(x)} p(U) \right) \geqslant \operatorname*{Max}_{y \neq 0} \left(\frac{|V^T y|}{p(y)} \cdot q(U) \right) \quad : (1)$$

puisque : $U \cdot V^T x = (V^T x).U$, $p(V^T x . U) = |V^T x| \cdot p(U)$

$\mathcal{F}_p(a)$ désigne la frontière de la boule : $B_p(a)$ (ensemble des x tels $p(x) \leqslant a$)
$\mathcal{F}_q(a)$ " " " " " " : $B_q(a)$ (" " " " " $q(x) \leqslant a$)



Soit $U \in \mathcal{F}_p(1)$, donc $p(U) = 1$
alors $q(U) = a$ et $U \in \mathcal{F}_q(a)$
On peut écrire la condition (1)

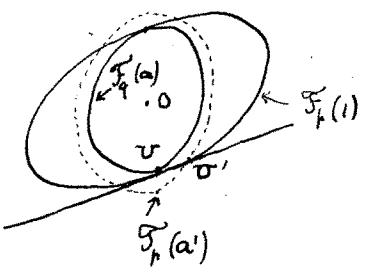
$$(1)' : \operatorname*{Max}_{x \in \mathcal{F}_p(1)} |V^T x| \geqslant \operatorname*{Max}_{y \in \mathcal{F}_q(a)} |V^T y|$$

puisque

$$\operatorname*{Max}_{y \neq 0} \left(\frac{|V^T y|}{q(y)} q(U) \right) = \operatorname*{Max}_{y \in \mathcal{F}_q(a)} |V^T y|$$

(1') se traduit géométriquement en disantque les deux hyper-plans d'appui perpendiculaires à V pour $B_p(1)$, contiennent les deux hyper-plans d'appui perpendiculaires à V pour $B_q(a)$ et cela pour tout V. Donc $B_q(a)$ est toute dans $B_p(1)$ et tout hyper-plan d'appui en U pour $B_p(1)$ l'est pour $B_q(a)$. [43]

Or, si a varie, lesboules $B_q(a)$ se déduisent de l'une d'elles par des homothéties. Supposons que pour un point $U \in \mathcal{F}_p(1)$, $q(U) = a$



Je dis qu'en un autre point $U' \in \mathcal{F}_p(1)$
on a aussi : $q(U') = a$

En effet, si $q(U') = a'$, avec $a' > a$, par exemple, pour $B_q(a')$ il existe des points strictement extérieurs à $B_p(1)$, puisqu'il y a des points de $B_q(a')$ du côté de l'hyper-plan d'appui en U commun non situés du même côté que O. Donc $B_q(a')$ ne serait pas tout dans $B_p(1)$, donc ne peut avoir que $a = a'$.

Il en résulte que $B_p(1)$ est une $B_q(a)$ pour un a déterminé.

Donc : $q(x) = a \cdot p(x)$, $S_{pp} = S_{qq}$.

D'autre part, si (2) : $M(A) \leqslant S_{pp}(A)$ quel que soit A, on sait qu'il existe $p_n(x)$ telle que :

$$S_{p_n p_n}(A) \leqslant M(A) \leqslant S_{pp}(A)$$

Le résultat ci-dessus prouve que $S_{p_n p_n} = S_{pp}$

donc (2) implique : $M(A) = S_{pp}(A)$.

Les normes $S_{pp}(A)$ sont les normes multiplicatives <u>minimales</u>.

<u>4°) Groupes de matrices associés à des normes de vecteurs ,ou de matrices</u>

L'ensemble $\mathfrak{S}$ des normes S_{pp} nous semble si important qu'il est intéressant d'essayer de le caractériser d'un autre point de vue.

a) Soit p une norme de vecteur dans R^n (ou C^n) , $\mathcal{G}_p$ désigne l'ensemble des matrices carrées Q telles:

$$p(Qx) = p(x) \qquad \text{pour tout } x \in R^n \ (\text{ou } C^n)$$

Il est évident que:

- I) $I \in \mathcal{G}_p$
- II) si Q_1 et $Q_2 \in \mathcal{G}_p$, $p(Q_1 Q_2 x) = p(Q_1(Q_2 x)) = p(Q_2 x) = p(x)$ donc $Q_1 \cdot Q_2 \in \mathcal{G}_p$
- III) si $Q \in \mathcal{G}_p$, elle ne peut être singulière ,car si non il existerait $x_0 \neq 0$ tel $Qx_0 = 0$; par suite :

$p(x_0) = p(Qx_0) = 0$ ce qui est absurde.

Alors , si dans $p(x) = p(Qx)$ on fait $y = Qx$ on aura $p(Q^{-1}y) = p(y)$ donc $Q^{-1} \in \mathcal{G}_p$.

Donc $\mathcal{G}_p$ est un groupe multiplicatif de $\mathfrak{M}_n$, il est dit <u>associé</u> à la norme de vecteur p .

b) Soit N une norme de matrice . On prouve pareillement que l'ensemble $\mathcal{G}_N$ des matrices carrées Q telles:

$$N(QA) = N(A) \qquad \text{pour toute } A \in \mathfrak{M}_n$$

est un groupe multiplicatif de $\mathfrak{M}_n$.

(les points I et II sont évidents ,pour III,si $Q \in \mathcal{G}_N$ elle ne peut être singulière ,car si non il existe $x_0 \neq 0$ auquelest associée $X_0 \neq 0$ telle que $Q \cdot X_0 = 0$ donc $N(X_0) = N(QX_0) = 0$ ce qui est absurde,lasuite comme dans la démonstration précédente). Il est dit <u>associé</u> à la norme N.

<u>Théorème VIII</u> Etant donnée la norme N ,si l'on forme $r_N = Q(N)$, on a: $\mathcal{G}_N \subset \mathcal{G}_{r_N}$

Le groupe associé à la norme N est un sous-groupe du groupe associé à la norme de vecteur correspondant à N par Q .

Car si $Q \in \mathcal{G}_N$ on a pour tout x : $r_N(Qx) = N(QX) = N(X) = r_N(x)$

Mais en généralsi $Q \in \mathcal{G}_{r_N}$, $r_N(Qx) = r_N(x) \rightarrow N(QX) = N(X)$

X n'est pas une matrice quelconque et l'on ne peut affirmer $N(A) = N(QA)$ pour tout A .

<u>Thèorème IX</u> Si l'on se donne une norme de matrice $S_{rr} \in \mathcal{S}$, le groupe $\mathcal{G}_{S_{rr}}$ est identique au groupe $\mathcal{G}_{rS_{rr}}$,lui-même identique à $\mathcal{G}_r$.

En effet, je sais que $r_{S_{rr}}$ est définie par $r_{S_{rr}}(x) = S_{rr}(X)$ et j'ai prouvé dans la démonstration du Th.V que $r_{S_{rr}} = c \cdot r = r_1$

donc $\mathcal{G}_{rS_{rr}} = \mathcal{G}_r$

Soit donc $Q \in \mathcal{G}_r = \mathcal{G}_{rS_{rr}}$

on a:

$$S_{rr}(Q \cdot A) = \max_{y \neq 0} \left(\frac{r(QAy)}{r(y)} \right) = \max_{y \neq 0} \left(\frac{r(Ay)}{r(y)} \right) = S_{rr}(A)$$

et par suite: $Q \in \mathcal{G}_{S_{rr}}$

Donc $\mathcal{G}_{rS_{rr}} \subset \mathcal{G}_{S_{rr}}$ et d'après le théorème précédent comme $\mathcal{G} \subset \mathcal{G}_{S_{rr}}$ on a bien le résultat énoncé.

<u>Théorème X</u> Pour une norme de matrice $S_{rr} \in \mathcal{S}$,toute matrice $Q \in \mathcal{G}_r$ est telle que $S_{rr}(QA) = S_{rr}(AQ) = S_{rr}(A)$

On vient de voir que $S_{rr}(QA) = S_{rr}(A)$

D'autre part,

$$S_{rr}(AQ) = \max_{x \neq 0} \left(\frac{r(AQx)}{r(x)} \right) = \max_{x \neq 0} \left(\frac{r(AQx)}{r(Qx)} \right)$$

puisque Q est inversible Qx prend toute valeur $y \in R^n$ (ou C^n)

donc: $S_{rr}(AQ) = \max_{y \neq 0} \left(\frac{r(Ay)}{r(y)} \right) = S_{rr}(A)$

On peut remarquer aussi que $S_{rr}(Q) = +1$ pour tout $Q \in \mathcal{G}_r$

car $S_{rr}(Q) = \max_{x \neq 0} \left(\frac{r(Qx)}{r(x)} \right) = +1$

5°) Normes de matrices complètement invariantes.

Soit $N(A)$ une norme générale sur l'ensemble des matrices carrées d'ordre n et à éléments réels, c'est-à-dire une application de l'ensemble des matrices carrées d'ordre n sur R, dans R_+ et satisfaisant aux axiomes de ce chapitre . On peut se demander s'il existe des normes "complètement invariantes", c'est-à-dire telles

$$N(A) = N(SAS^{-1})$$

pour toute A et pour toute S non singulière.

Théorème XII : Il est impossible de trouver pour $n \geqslant 2$ une norme complètement invariante.

En effet, considérons la matrice (d'ordre 2)

$$A = \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix}$$

et la matrice $A' = \begin{vmatrix} 0 & \frac{a}{c} \\ 0 & 0 \end{vmatrix}$, $(c \neq 0,\ a \neq 0)$

Il est clair que :

$$\begin{vmatrix} a & b \\ 0 & c \end{vmatrix} \cdot \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix} = \begin{vmatrix} 0 & a \\ 0 & 0 \end{vmatrix}$$

et que :

$$\begin{vmatrix} 0 & \frac{a}{c} \\ 0 & 0 \end{vmatrix} \cdot \begin{vmatrix} a & b \\ 0 & c \end{vmatrix} = \begin{vmatrix} 0 & a \\ 0 & 0 \end{vmatrix}$$

Donc, si $S = \begin{vmatrix} a & b \\ 0 & c \end{vmatrix}$, $SA = A'S$

S n'étant pas singulière $A' = S.A.S^{-1}$

Il est clair que si N est une norme quelconque

$$N(A') = \left|\frac{a}{c}\right| . N(A)$$

puisque

$$A' = \frac{a}{c} . A$$

Donc, on ne peut avoir $N(A') = N(A)$ quel que soit A et S

Cette démonstration s'étend visiblement à n quelconque.

Cela prouve qu'il est impossible de trouver une norme de matrice dont lavaleur ne dépendrait que de la transformation linéaire qu'elle représente, puisque les matrices $S.A.S^{-1}$ peuvent s'interprêter comme définissant la même transformation linéaire.

Cela explique les difficultés qu'il y a dans le "dosage" de la proximité des éléments propres d'une matrice (c'est-à-dire attachés à la transformation qu'elle représente) par des mesures de voisinage sur les coefficients de celle-ci.

6°) Normes dematrices usuelles en calcul numérique linéaire.

Il est clair que pour l'ensemble des matrices A d'ordre n, à éléments dans R (ou C), en correspondance biunivoque avec R^{n^2} (ou (C^{n^2})), on peut prendre une norme de vecteur dans R^{n^2} (ou C^{n^2}).

On obtient ainsi :

$$N_1 = \sum_{i,j} |a_{ij}| \quad , \quad N_2 = \sqrt{\sum_{i,j} |a_{ij}|^2} \quad , \quad N_\infty = \operatorname*{Max}_{i,j} |a_{ij}|$$

qui correspondant aux trois normes de HOLDER; φ_1 , φ_2 , φ_∞

Les normes de matrices obtenues à partir de la construction du paragraphe 3 seront utilisées en détail dans le chapitre suivant.

CHAPITRE III

Les conditionnements numériques

I) Notion de conditionnement .

Après avoir présenté le matériel théorique dont nous nous servirons nous abordons les problèmes de calcul numèrique que nous nous proposons d'étudier. Nous commençons par la notion de conditionnement numérique.

Nous allons considérer, pour commencer, un exemple cité par **BODEWIG[9]**

Soient les trois systèmes de deux équations à deux inconnues du I° degré:

$$\text{I} \begin{cases} 3x + 4y = 7 \\ 3x + 4{,}0000\text{I}\, y = 7{,}0000\text{I} \end{cases} \quad \text{ou} \quad \begin{cases} 3x + 4y - 7 = 0 \\ 3x + 4{,}0000\text{I}\, y - 7{,}0000\text{I} = 0 \end{cases}$$

$$\text{II} \begin{cases} 3x + 4y - 7 = 0 \\ 3x + 3{,}99999\, y - 7{,}00004 = 0 \end{cases}$$

$$\text{III} \begin{cases} 3x + 4y - 7 = 0 \\ 3x + 3{,}999992\, y - 7{,}000042 = 0 \end{cases}$$

Ces systèmes ont pour solution exacte

I $x = \text{I}$, $y = \text{I}$.

II $x = 7 + 2/3$, $y = -4$.

III $x = 9 + \text{I}/3$, $y = -(5 + \text{I}/4)$

Il est d'usage de dire que de tels systèmes sont "mal" conditionnés.

Avant de quitter cet exemple, voici un tableau donnant les valeurs des deux formes des équations I, II, III, lorsque l'on y substitue les valeurs solutions de I, II, III.

	Sys I	Sys II	Sys III
Sol I	0 0	0 -0,00005	0 -0,00005
Sol II	0 -0,00005	0 0	0 -0,00001
Sol III	0 -0,0000625	0 0,0000105	0 0

2°) Examen des différentes définitions de conditionnement.

La nécessité d'une définition correcte du conditionnement d'un système linéaire a été ressentie depuis longtemps par les calculateurs.

Remarquons d'abord que la valeur absolue du déterminant d'une matrice n'apprend rien sur le conditionnement du système. Soit, en effet, un système (d'ordre 10 par exemple), de déterminant égal à 1 par exemple ; multiplions par 1/10 toutes les équations, le déterminant du nouveau système est 10^{-10}. Or, la résolution du nouveau système n'est pas numériquement plus délicate que celle du système de départ.

TURING [13] et LONSETH [14] ont le point de vue suivant. Ils cherchent à savoir si, faisant varier de très peu les éléments de A, la solution exacte du système : $(A + \delta A)\, x - b = 0$ diffère de peu ou de beaucoup de la solution exacte $A\, x - b = 0$.

Ce point de vue conduit à une définition du conditionnement voisine de la nôtre (cf. Th. III).

TURING [13] considère deux nombres pouvant donner une mesure du conditionnement. (cf. aussi sur ce sujet TODD , [15])

le premier est : $T_1(A) = n \cdot \underset{i,j}{\mathrm{Max}} |a_{ij}| \cdot \mathrm{Max} |a'_{ij}|$. , $A^{-1} = (a'_{ij})$

le second est : $T_2(A) = \frac{1}{n} N(A) \cdot N(A^{-1})$, $N(A) = \left(\sum_{i,j} a_{ij}^2 \right)^{\frac{1}{2}}$

Pour VON NEUMANN et GOLDSTINE [16] , la notion de conditionnement vient à l'issue d'un calcul d'erreur dans la résolution numérique des systèmes linéaires à matrice symétrique. Leur nombre de conditionnement égal au rapport de la plus grande à la plus petite valeur propre de la matrice A, est un cas particulier de ceux que nous définissons (à l'inverse près).

3°) Définition générale des conditionnements numériques.

Soit $\mathcal{E}$ un espace vectoriel topologique, normé et complet, N_1 et N_2 deux normes équivalentes sur $\mathcal{E}$. Soit T une application de $\mathcal{E}$ dans lui-même, satisfaisant en un point x_0 à la propriété:

(P) Il existe un voisinage V_0 de x_0, tel que pour tout $x \in V_0$, deux nombres positifs p_1 et p_2 puissent être trouvés tels que:

$$0 < p_1 \leqslant \frac{N_1\left(T(x)-T(x_0)\right)}{N_2(x-x_0)} \leqslant p_2 \qquad \text{pour tout } x \in V_0$$

Il est clair que si (P) est vérifiée, on ne fait pas de restriction en supposant que p_1 est la Lim des p_1 y satisfaisant

que p_2 est la Lim des p_2 y satisfaisant.

On suppose qu'il en est ainsi dans toute la suite, autrement dit, que les bornes p_1 et p_2 qui figurent dans l'énoncé de (P) sont les "meilleures"

Théorème I : Toute application T satisfaisant à (P) est continue en x_0 et l'équation

$$T(x) = T(x_0)$$

a dans V_0 une et une seule solution $x=x_0$.

a) La continuité est très simple à prouver: si ε est donné

comme $N_1\left[T(x)-T(x_0)\right] \leqslant p_2 N_2(x-x_0)$

il suffit de choisir x de sorte que: $N_2(x-x_0) \leqslant \frac{\varepsilon}{p_2}$

mais puisque N_1 et N_2 sont équivalentes il existe a tel $N_2(x-x_0) \leqslant a.N_1(x-x_0)$: donc il suffit de prendre x dans le voisinage:

$$V_0 \cap \left\{ (x) \text{ tels } N_1(x-x_0) \leqslant \frac{\varepsilon}{p_2 a} \right\}$$

b) Puisque $p_1 \neq 0$ $p_1 N_2(x_1-x_0) \leqslant N_1\left(T(x_1)-T(x_0)\right)$ implique si $T(x_1)=T(x_0)$

$$N_2(x_1-x_0) = 0$$

donc $x_1 = x_0$

Expliquons l'importance en calcul numérique des nombres p_1 et p_2 de la propriété (P):

Tout calcul effectif se conduit avec un certain "moyen de calcul" Nous entendons par là, soit une machine, soit, dans le cas d'un calcul"a la main "un certain nombre de règles que doit obligatoirement se donner le calculateur. Dans un cas, comme dans l'autre ce "moyen" est toujours à capacité limitée, aussi bien supérieurement qu'inférieurement. Autrement dit, dans tout calcul effectif, on est obligé de considérer comme nuls des nombres qui ne sont réellement qu'inférieurs à une certaine valeur ε.

Si donc le calcul nous a conduits, pour la solution du problème:

$$T(x) = T(x_0)$$

à un x_1 tel $\quad \eta_1 \leq N_1(T(x_1) - T(x_0)) \leq \eta_2$

tout ce que l'on peut dire relativement à la proximité de x_1 et x_0 est que

$$N_2(x_1 - x_0) \leq \frac{\eta_2}{\rho_1} = \varepsilon_{Max}$$

$$N_2(x_1 - x_0) \geq \frac{\eta_1}{\rho_2} = \varepsilon_{min}$$

Donc: $\frac{\varepsilon_{min}}{\varepsilon_{Max}} = \frac{\eta_1}{\eta_2} \cdot \frac{\rho_1}{\rho_2}$

Propriété I : Le rapport entre les extremums des erreurs en norme N_2, que l'on peut commettre est égal à $\frac{\eta_1}{\eta_2} \cdot \frac{\rho_1}{\rho_2}$. [10]

Si l'ordre de grandeur de $N_1(T(x_1) - T(x_0))$ est connu $= \eta = \eta_1 = \eta_2$ on peut dire que ce rapport est égal à : $\frac{\rho_1}{\rho_2}$.

Définition I: On appelle conditionnement général d'une équation $T(x) = T(x_0)$, avec T satisfaisant à (P), pour x_0 et relativement aux normes N_1 et N_2 le rapport $\frac{\rho_1}{\rho_2}$.

4°) Cas d'applications linéaires

D'après les propriétés exposées au chapitre II, si T est linéaire et non singulière la propriété (P) est toujours satisfaite, les nombres p_1 et p_2 ne dépendent pas de x_0 . La définition I est donc toujours applicable . Il est cependant obligatoire de particulariser pour les applications numériques afin de rendre utilisable en calcul cette définition.

Soit alors A une matrice carrée non singulière définissant une application linéaire de R^n dans R^n; formons les 9 rapports

$$r(\varphi_i, \varphi_j, A, x) = \frac{\varphi_i(Ax)}{\varphi_j(x)} \qquad (j,i = 1, 2, \infty)$$

On sait que $\underset{x}{Max}\, r(\varphi_i, \varphi_j, A, x)$ est une norme de matrice sur $\mathcal{M}_n$ c'est ce qu'au chapitre II on a posé = $S_{\varphi_i \varphi_j}(A) = S_{i,j}(A)$

(elle n'est multiplicative que si $i=j$)

Donc le p_2 de la propriété (P) n'est autre que $S_{i,j}(A)$.

D'autre part, écrivons

$$r(\varphi_i, \varphi_j, A, x_1) = \frac{\varphi_i(x_1)}{\varphi_j(A^{-1}x_1)} \quad , \quad x_1 = Ax \quad , \quad r(\varphi_i, \varphi_j, A, x) = \frac{1}{r(\varphi_j, \varphi_i, A^{-1}, x_1)}$$

Puisque A n'est pas singulière

Donc le p_1 de la propriété (P) n'est autre que $S_{j,i}(A^{-1})^{-1}$

D'où le théorème important :

Théorème II : Le conditionnement général d'une matrice A non singulière pour les normes φ_i et φ_j est :

$$\gamma_{ij} = 1 / S_{ij}(A) \cdot S_{ji}(A^{-1})$$

Définition : Si l'on se donne un système d'équations linéaires dans R^n :

$$A\,x - b = 0$$

ayant une solution unique x_0, son conditionnement général n'est autre que celui de sa matrice de premier membre, puisque le système s'écrit :

$$A\,x = A\,x_0$$

Remarquons quele théorème II peut, très généralement, conduire à poser pour définition d'un nombre de conditionnement

$$\Gamma_{12} = 1 / N_1(A) \cdot N_2(A^{-1})$$

N_1 et N_2 étant deux normes générales.

C'est ce qu'a fait TURING (à l'inverse près).

Les différentes interprêtations de ce nombre sont plus difficiles.

Nous nous bornerons à l'étude de normes de matrices du type S_{ij}.

Enfin, pour rapprocher les points de vue, on a :

Théorème III : Soit x_0 la solution exacte d'un système linéaire d'ordre n

$$A\,x - b = 0$$

Soit x_1 la solution exacte du système varié

$$(A + \delta A)x - b = 0$$

φ_i et φ_j deux normes de vecteurs sur R^n (ou C^n).

(1) Si les variations δA sont prises de sorte que $\dfrac{\varphi_i(\delta A\, x_1)}{\varphi_i(x_1)}$ soit toujours d'un même ordre de grandeur η ;

alors, le rapport des variations relatives extrêmes de la solution, en normes φ_i et φ_j, n'est autre que le conditionnement général γ_{ij}.

— En effet, de $A\,x_0 = b$ et $(A + \delta A)\,x_1 - b = 0$

on tire :

$$A(x_1 - x_0) + \delta A\, x_1 = 0$$

donc :

$$\varphi_i\left(A\,(x_1 - x_0)\right) = \varphi_i(\delta A\, x_1)$$

$$\frac{\varphi_i(A(x_1-x_0))}{\varphi_j(x_1-x_0)} = \frac{\varphi_i(\delta A x_1)}{\varphi_i(x_1)} \cdot \frac{\varphi_i(x_1)}{\varphi_j(x_1-x_0)} \# \eta \cdot \frac{\varphi_i(x_1)}{\varphi_j(x_1-x_0)}$$

si p_1, p_2 sont les barres du rapport du 1er terme.

$$E_{2,Max} = \max_{x_1} \left(\frac{\varphi_j(x_1-x_0)}{\varphi_i(x_1)} \right) \# \eta \cdot \frac{1}{p_1} \quad ,$$

$$E_{2,min} = \min_{x_1} \left(\frac{\varphi_j(x_1-x_0)}{\varphi_i(x_1)} \right) \# \eta \cdot \frac{1}{p_2} \quad ,$$

$$\frac{E_{2,min}}{E_{2,Max}} \# \frac{p_1}{p_2} = \gamma_{ij} \, .$$

En ce sens, le γ_{ij} apparaît comme un <u>indice de forme</u>.

<u>Nota</u> : La condition (1) est toujours satisfaite en ordre de grandeur, ce qui est en calcul la chose la plus importante.

<u>5°) Comparaison des différents conditionnements généraux d'une matrice.</u>

Soient, pour une matrice A donnée non singulière, les deux rapports :

$$r_{ij} = \frac{\varphi_i(Ax)}{\varphi_i(x)} \quad , \quad r_{kl} = \frac{\varphi_k(Ax)}{\varphi_l(x)} \quad , \quad (i,j,k,l = 1,2,\infty)$$

on sait que :

$$(1) \qquad a_{ki} \; \varphi_k(X) \leq \varphi_i(X) \leq b_{ki} \; \varphi_k(X)$$

donc ,

$$a_{ki} \; \varphi_k(Ax) \leq \varphi_i(Ax) \leq b_{ki} \; \varphi_k(Ax)$$

Mais

$$a_{lj} \; \varphi_l(x) \leq \varphi_j(x) \leq b_{lj} \; \varphi_l(x)$$

d'où

$$a_{ki} \; \frac{\varphi_k(Ax)}{\varphi_l(x)} \leq \frac{\varphi_i(Ax)}{\varphi_j(x)} \leq b_{ki} \; \frac{\varphi_k(Ax)}{\varphi_l(x)}$$

puis

$$\frac{a_{ki}}{b_{lj}} \frac{\varphi_k(Ax)}{\varphi_l(x)} \leq \frac{\varphi_i(Ax)}{\varphi_j(x)} \leq \frac{b_{ki}}{a_{lj}} \cdot \frac{\varphi_k(Ax)}{\varphi_l(x)} \; ; \quad \frac{a_{ki}}{b_{lj}} \, p_{1,kl} \leq p_{1,ij} \leq p_{2,ij} \leq \frac{b_{ki}}{a_{lj}} \cdot p_{2,kl}$$

enfin, en posant γ_{ij} et γ_{kl} pour les conditionnements relatifs aux couples de normes φ_i, φ_j et φ_k, φ_l , on peut énoncer :

Thèorème IV Les conditionnements généraux d'une même matrice A ,par rapport à deux couples de normes φ_i, φ_j et φ_k, φ_ℓ sont toujours tels:
(les constantes a_{pr} et b_{pr} sont définies par(1))

$$\gamma_{ij} \cdot \frac{a_{j\ell}}{b_{j\ell}} \cdot \frac{a_{ik}}{b_{ik}} \leq \gamma_{k\ell} \leq \gamma_{ij} \cdot \frac{b_{\ell j}}{a_{\ell j}} \cdot \frac{b_{ki}}{a_{ki}}$$

Nous allons écrire ces relations pour les cas envisagés:

Voici les tableaux relatifs aux valeurs des a_{ij} et des b_{ij}.

a_{ij}

$i \downarrow$ $j \rightarrow$ a_{ij}	1	2	∞
1	1	$\frac{1}{\sqrt{n}}$	$\frac{1}{n}$
2	1	1	$\frac{1}{\sqrt{n}}$
∞	1	1	1

b_{ij}

$i \downarrow$ $j \rightarrow$ b_{ij}	1	2	∞
1	1	1	1
2	$\sqrt{n}$	1	1
∞	n	$\sqrt{n}$	1

D'après ces tableaux on voit que la double inégalité la moins "sérrée" est obtenue pour: $k.\ell = 1$, $i = j = \infty$
elle s'écrit: $\gamma_{\infty\infty} \frac{1}{n^2} \leq \gamma_{11} \leq \gamma_{\infty\infty} \cdot n^2$
cela nous permet de conclure:

Propriété II Tous les conditionnements généraux que nous venons de définir sont équivalents.

Théorème V Le conditionnement γ_{22} relatif aux deux normes φ_2 et φ_2 ,identiques à la norme euclidienne sur R^n, d'une matrice A non singulière n'est autre que le rapport $\sqrt{\lambda_1}/\sqrt{\lambda_n}$,où λ_1 et λ_n désignent la plus petite et la plus grande des valeurs propres de la matrice $A^T.A$.

En effet dans ce cas on peut écrire: $\gamma_{22} = \frac{\|Ax\|}{\|x\|}$

ou si X désigne la colonne des composantes de x

$$\gamma_{22}^2 = \frac{X^T A^T A X}{X^T X}$$

Or, la matrice $A^T.A$ est définie positive et symétrique, et non singulière, donc a des valeurs propres $0 < \lambda_1 \leq \lambda_2 \cdots\cdots \leq \lambda_n$ strictement positives.

D'autre part, il existe une matrice Q, orthonormale ($Q^T.Q = I$) telle que :

$$Q^T A^T . A Q = \Lambda$$

ou

$$\Lambda = \begin{vmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \lambda_n \end{vmatrix}$$

Or, si l'on pose $X = Q\,Y$, on a :

$$r_{22}^2 = \frac{Y^T . Q^T A^T . A . Q . Y}{Y^T Q^T Q Y} = \frac{Y^T \Lambda Y}{Y^T Y} = \frac{\sum_{i=1}^{i=n} \lambda_i y_i^2}{\sum_{i=1}^{i=n} y_i^2}$$

sous cette forme, on voit que :

$$r_{1,22}^2 = \lambda_1 \quad , \quad r_{2,22}^2 = \lambda_n$$

donc :

$$\gamma_{22} = \sqrt{\frac{\lambda_1}{\lambda_n}}$$

Si A est donnée définie positive et symétrique si $0 < u_1 \leq \cdots\cdots \leq u_n$ sont les valeurs propres de A, elle-même, $u_i^2 = \lambda_i$

donc, dans ce cas : $$\gamma_{22} = \frac{u_1}{u_n}$$

C'est le nombre de VON NEUMANN. Il est assez laborieux à calculer.
Les nombres deTURING le sont un peu moins.

Il paraît normal de considérer un système comme le mieux conditionné possible si le système de "plans" dont l'intersection est la solution du système linéaire est un système orthogonal. Le chapitre qui va suivre expose les propriétés de nombres ayant les qualités d'être assez faciles à construire, et d'indiquer la condition d'orthogonalité simplement .

CHAPITRE IV

Le conditionnement normalisé C(A)

1°) Notations : Dans ce qui va suivre, je vais particulièrement étudier le rapport

$$r_{12} = r(\varphi_1, \varphi_2, A, x) = \frac{\varphi_1(Ax)}{\varphi_2(x)}$$

On posera :

$$r_{12}(x) = \frac{\Phi(Ax)}{\|x\|} \quad , \quad m_A = \min_{x \neq 0} \left(r_{12}(x)\right) \quad , \quad M_A = \max_{x \neq 0} \left(r_{12}(x)\right) ,$$

$$\gamma_{12} = \gamma = \frac{m_A}{M_A} .$$

L'intérêt de ces normes dissymétriques est d'avoir des propriétés géométriques simples comme on va le voir.

Afin de mieux utiliser cette interprêtation géométrique, nous commencerons par l'étude d'un cas particulier :

2°) Etude géométrique du rapport $\frac{m_A}{M_A}$, pour A "normée en lignes"

Nous disons que A est "normée en lignes", si tout vecteur-ligne $A_{i.}$ est de module I :

$$\sum_{k=1}^{k=n} a_{ik}^2 = 1 \qquad (i = 1, 2, \ldots n)$$

On suppose pour la suite A non singulière.

Ici $$\Phi(x) = \sum_{i=1}^{i=n} \left| \sum_{k=1}^{k=n} a_{ik} x_k \right| = \sum_{i=1}^{i=n} \left| f_i(x) \right|$$

Pour trouver m_A et M_A il suffit de déplacer le point (x) sur le lieu des points (x) tels que $\Phi(x) = 1$

Or ce lieu est la frontière d'un polyèdre $\mathcal{P}$ convexe ayant l'origine pour centre de symétrie, cette frontière : Π est constituée par des "faces" planes qui sont des parties des plans dont les équations sont :

$$\sum_{i=1}^{i=n} \varepsilon_i f_i(x) = 1 \qquad (\varepsilon_i = \pm 1)$$

La condition de normalisation des lignes de A entraine que

Π est le lieu des points dont la somme des distances géométriques aux plans P_i d'équation $f_i(x) = 0$ est égale à I

Par suite :

a) $m_A = \min_{(x) \text{ tel que } \Phi(x)=1} \left(\frac{1}{\|x\|} \right) = \frac{1}{d_M}$, si d_M est la distance maximun de l'origine à un point de Π, et puis que Π limite un polyèdre convexe, d_M est la distance de O, au sommet

le plus éloigné de O;

b) $M_A = \underset{(x)\,\text{tel que}\,\Phi(x)=1}{\text{Max}} \left(\frac{1}{\|x\|} \right) = \frac{1}{h_m}$, si h_m est la distance minimum de O à un point de Π . Toujours à cause de la convexité de $\mathcal{P}$, h_m est la distance de O à la face la plus proche de O.

3°) Etude analytique:

a) Soient D_i , $(i = 1,2, \ldots, n)$ les n droites :

$$D_i = \bigcap_j P_j \qquad (j = 1,2, \ldots, i-1, i+1, \ldots n)$$

les plans P_i , d'équations $f_i(x) = 0$.

Le fait que A soit non singulière assure que les droites D_i peuvent être prises comme support d'un système de vecteurs $(\vec{e}_1, \vec{e}_2, \ldots \vec{e}_n)$ formant une base pour R^n . Nous choisissons un tel système de sorte que $\|\vec{e}_i\|^2 = 1$.

b) Détermination des faces du polyèdre $\mathcal{P}$ dans le le système d'axes $(o; \vec{e}_1, \vec{e}_2, \ldots \vec{e}_n)$.

Un plan coupant les droites D_i en des points A_i d'abscisse α_i , a pour équation : $\sum_{i=1}^{i=n} \frac{x_i}{\alpha_i} = 1$,

ou si V , est la colonne : $\begin{vmatrix} \frac{1}{\alpha_1} \\ \vdots \\ \frac{1}{\alpha_n} \end{vmatrix}$, $x = \begin{vmatrix} x_1 \\ \vdots \\ x_n \end{vmatrix}$

cela s'écrira : (1) $x^T . V = 1$

Or le produit scalaire de deux vecteurs de composantes (x_i) et (y_i) sécrit: $p = (x_1 \vec{e}_1 + \ldots + x_n \vec{e}_n) \cdot (y_1 \vec{e}_1 + \ldots + y_n \vec{e}_n) = x^T . G . y$

où $G = (g_{ij})$ avec $g_{ij} = \vec{e}_i \cdot \vec{e}_j$ $(i = 1, \ldots n; j = 1, \ldots n)$

(G: matrice de "métrique")

Si $\vec{U}$ est un vecteur de module I ,on aura entre ses composantes

(2) $u^T G u = 1$

Considèrons le plan orthogonal en H ,de coordonnées $h.u$ $(h > 0)$ au vecteur U , son équation est : $(x^T - h.u^T) . G . u = 0$

ou $x^T G u = h$ (3)

En identifiant (3) et (1) cela donne: $h . V = G u$

ou $u = h G^{-1} . V$

et portant dans (2) ,puisque G^{-1} est symétrique : $h^2 V^T G^{-1} V = 1$

Donc :

$$(4) \qquad \frac{1}{h^2} = V^T G^{-1} V \ .$$

Mais les points A_i sont tels que le plan $\frac{x_i}{\alpha_i} = 1$ soit à la distance I de O, car le polyèdre $\mathcal{P}$ n'est autre que le lieu des points dont la somme des distances géométriques auxplans de notre nouveau système de coordonnées est inférieure ou égale à I.

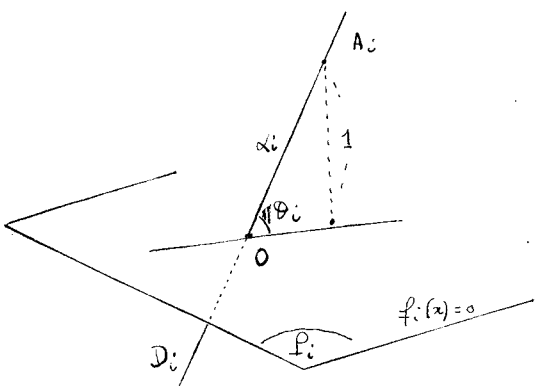
Donc si $V_i = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ \frac{1}{\alpha_i} \leftarrow i^{\circ} \\ \vdots \\ 0 \end{vmatrix}$

Il résulte de (4) que :

$$1 = V_i^T G^{-1} V_i$$

Donc : $\frac{g^{ii}}{\alpha_i^2} = 1$, $\alpha_i = \varepsilon_i \sqrt{g^{ii}}$, si $G^{-1} = (g^{ii})$

On peut remarquer que : I°) $g^{ii} \geqslant 1$



Car si θ_i est l'angle aigu de D_i avec P_i d'équation $f_i(x) = 0$ on a $|\alpha_i| = \frac{1}{\sin \theta_i}$.

2°) L'indétermination de signe provient du fait que sur D_i, le point A_i et son symétrique par rapport à O sont sommets de $\mathcal{P}$. Les faces de $\mathcal{P}$ sont les simplexes de sommets : n points A_i pris de sorte que deux d'entre-eux ne soient pas symètriques par rapprort à O

c) Calcul des d_M et h_m.

I°) Il est clair que : $d_M = \underset{i}{Max} |\alpha_i| = \underset{i}{Max} \sqrt{g^{ii}}$

2°) $h_m = \min(h)$, h, précédemment trouvé fromule (4)

. Or celle-ci, s'écrit en remarquant que $V = V_1 + V_2 + \cdots + V_n$

$$\frac{1}{h^2} = (V_1 + \cdots + V_n)^T G^{-1} (V_1 + \cdots + V_n)$$

Ou

$$\frac{1}{h^2} = n + 2 \sum_{\substack{i<j \\ (1,\cdots n)}} \frac{g^{ij}}{\alpha_i \alpha_j} = n + 2 \sum_{\substack{i<j \\ (1,\cdots n)}} \varepsilon_i \varepsilon_j \frac{g^{ij}}{\sqrt{g^{ii}} \sqrt{g^{jj}}} \qquad (6)$$

Etudions les diverses valeurs de h par (6)

Posons $Z_{ij} = \frac{g^{ij}}{\sqrt{g^{ii}} \sqrt{g^{jj}}}$

et

$$L = \sum_{i<j} \varepsilon_i \varepsilon_j Z_{ij} \quad , \quad \frac{1}{h^2} = n + 2L$$

D'autre part, écrivons :

$$L = \varepsilon_1\left[\varepsilon_2 Z_{12} + \varepsilon_3 Z_{13} + \cdots + \varepsilon_n Z_{1n}\right] + \varepsilon_2\left[\varepsilon_3 Z_{23} + \cdots + \varepsilon_n Z_{2n}\right] + \cdots\cdots + \varepsilon_{n-2}\left[\varepsilon_{n-1} Z_{n-2,n-1} + \varepsilon_n Z_{n-2,n}\right] + \varepsilon_{n-1}\varepsilon_n Z_{n-1,n}$$

Je regarde le signe de $Z_{n-1,n}$ et je choisis les signes $\varepsilon_{n-1}, \varepsilon_n$ de sorte que le dernier terme soit positif, puis je regarde le signe du crochet du terme précédent (une fois que $\varepsilon_{n-1}, \varepsilon_n$ sont fixés ce crochet a un signe bien déterminé) je choisis ε_{n-2} de sorte que ce terme soit positif , etc ..... Il est clair que de proche en proche je peux choisir les ε_i de sorte que $L \geqslant 0$

Donc on peut affirmer, la plus courte distance de l'origine à une face de $\mathcal{P}$ est obtenue certainement pour un choix ε_i tels que $L \geqslant 0$

Donc : $$\frac{1}{h_m^2} \geqslant n$$

Par suite on a toujours :

$$\frac{m_A^2}{M_A^2} = \frac{h_m^2}{d_M^2} = \frac{1}{\operatorname{Max}_i |\alpha_i|^2} \cdot h_m^2 \leqslant \frac{1}{n}$$

Enfin si $\frac{m_A^2}{M_A^2} = \frac{1}{n}$ c'est que $\operatorname{Max}_i |\alpha_i|^2$ qui est, comme on l'a vu toujours $\geqslant 1$ est effectivement tel $= 1$ (puisque $h_m^2 \leqslant \frac{1}{n}$)

Mais alors si $|\alpha_k|^2 = \operatorname{Max}_i |\alpha_i|^2 = 1$

donc : $|\alpha_i| = 1$ pour tout i

et puisque $|\alpha_i| = \frac{1}{\sin \theta_i}$

cela entraîne

$$\theta_i = \frac{\pi}{2}$$

D'où :

Théorème I Pour une matrice A , à termes réels, normée en lignes, non singulière

on a $$\gamma = \frac{m_A}{M_A} \leqslant \frac{1}{\sqrt{n}}$$

et une condition nécéssaire et suffisante pour qu'elle soit **orthonormale** est que $$\frac{m_A}{M_A} = \frac{1}{\sqrt{n}}$$

4°) Etude du rapport $\frac{M_A}{M_i}$ pour A à termes réels, non singulière.

Nous allons généraliser la démomstration précédente au cas où A n'est pas normalisée en lignes. Prenons encore le même système d'axes de coordonnées $(O; \vec{e}_1, \vec{e}_2 \cdots\cdots, \vec{e}_n)$.

Pour le plan, $\sum_{i=1}^{i=n} \frac{x_i}{\alpha_i} = 1$

la distance à l'origine est toujours donnée par :

$$\frac{1}{h^2} = V^T . G^{-1} . V$$

où $V = \begin{vmatrix} \frac{1}{\alpha_1} \\ \vdots \\ \frac{1}{\alpha_i} \\ \vdots \\ \frac{1}{\alpha_n} \end{vmatrix}$.

Le sommet A_i sur la droite D_i, est tel que les composantes de $\overrightarrow{OA_i}$ sont: $\overrightarrow{OA_i} \begin{cases} 0 \\ 0 \\ \vdots \\ \alpha_i \leftarrow i \\ \vdots \\ 0 \end{cases}$

Il est aussi tel que $f_j(A_i) = 0$ pour $j \neq i$ $\quad (j = 1, \ldots, i-1, i+1, \cdots n)$

donc $\Phi(A_i) = |f_i(A_i)| = d_i \cdot \rho_i$

si l'on pose $\rho_i = \|A_i\| = \sqrt{a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2}$

Pour un sommet du polyèdre $\mathcal{P}$: $\Phi(x) = 1$, on aura $d_i = \frac{1}{\rho_i}$.

D'autre part, le plan $\frac{x_i}{\alpha_i} = 1$ (parallèle à $f_i(x) = 0$ et passant par A_i) est à la distance d_i de O.

Par suite

$$\frac{1}{d_i^2} = V_i^T G^{-1} V_i \qquad \text{si :} \quad V_i = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ \frac{1}{\alpha_i} \\ \vdots \\ 0 \end{vmatrix} \leftarrow i$$

donc : $\frac{g^{ii}}{\alpha_i^2} = \frac{1}{d_i^2} = \rho_i^2 \qquad$ et $\alpha_i = \varepsilon_i \frac{\sqrt{g^{ii}}}{\rho_i}$

On aura encore :

1°) $d_m = \max_i |\alpha_i| = \max_i \frac{\sqrt{g^{ii}}}{\rho_i}$

2°) $\frac{1}{h^2} = \sum_{i=1}^{i=n} V_i^T G^{-1} V_i + 2 \cdot \sum_{\substack{i<j \\ (1 \cdots n)}} V_i^T G^{-1} V_j$

$$\frac{1}{h^2} = \sum_{i=1}^{i=n} \rho_i^2 + 2 \sum_{i<j} \varepsilon_i \varepsilon_j \frac{g^{ij}}{\sqrt{g^{ii}} \cdot \sqrt{g^{jj}}} \rho_i \cdot \rho_j$$

Le raisonnement précédemment fait prouve donc que h_m (puisqu'on peut faire un choix des ε_i de sorte que le 2° terme soit $\geqslant 0$)

$$\frac{1}{h_m^2} \geqslant \sum_{i=1}^{i=n} \rho_i^2 = N^2(A)$$

si

$$N(A) = \sqrt{\sum_{i,j} a_{ij}^2}$$

Donc $\frac{m_A^2}{M_A^2} = \frac{\rho_m^2}{\max_i |a_i|^2} \leqslant \frac{1}{N^2(A)} \cdot \frac{1}{\max_i |a_i|^2}$

Or posons $g^{kk} = \max_i g^{ii} \geqslant 1$ et $\rho_0 = \min_i (\rho_i)$

on aura certainement: $\frac{m_A^2}{M_A^2} \leqslant \frac{\rho_0^2}{N^2(A)} \cdot \frac{1}{g^{kk}} \leqslant \frac{\rho_0^2}{N^2(A)}$

Si $\frac{m_A^2}{M_A^2} = \frac{\rho_0^2}{N^2(A)}$ c'est que $g^{kk} = 1$ donc tous les $g^{ii} = 1 \rightarrow \theta_i = \frac{\pi}{2}$
donc A sera orthogonale en lignes d'où:

Théorème II Pour une matrice A, à termes réels, non singulière, on a $\gamma = \frac{m_A}{M_A} \leqslant \frac{\rho_0}{N(A)}$

et une condition nécéssaire et suffisante pour que A soit orthogonale en lignes est que : $\frac{m_A}{M_A} = \frac{\rho_0}{N(A)}$.
où $\rho_0 = \min_i \|A_{i.}\|$ et $N(A) = \left(\sum_{ij} a_{ij}^2\right)^{\frac{1}{2}}$. (10)

Ce théorème va nous fournir la possibilité de définir un conditionnement normalisé.

5°) Le conditionnement normalisé C(A)

Soit une matrice A quelconque, à termes réels

Posons $\begin{cases} C(A) = 0 & \text{, si A est singulière} \\ C(A) = \frac{m_A}{M_A} \Big/ \frac{\rho_0}{N(A)} & \text{si A est non singulière} \end{cases}$

Théorème III On a toujours : $0 \leqslant C(A) \leqslant 1$.

Pour que A soit singulière il faut et il suffit $C(A) = 0$,

pour que A soit orthogonale en lignes il faut et il suffit $C(A) = 1$,

si A est telle que ses éléments varient de sorte que $A \rightarrow A_1$ et A_1 singulière, sans ligne de longueur $= 0$, $C(A) \rightarrow 0$

Les trois premières parties de l'énoncé du théorème II et de la définition de C(A) , Pour démontrer la dernière partie il suffit de remarquer que $C^2(A) \leqslant \frac{1}{g^{kk}} = \sin^2 \theta_k$

d'après les relations vues plus haut.

Or si $A \rightarrow A_1$ singulière (sans ligne nulle), l'un des $\theta_i \rightarrow 0$, donc $\theta_k \rightarrow 0$

$\left(|\sin \theta_k| = \min_i |\sin \theta_i|\right)$

Définition: le nombre C(A) sera appelé "conditionnement normalisé" ou, conditionnement (sans plus) si aucune confusion n'est à craindre.

C'est l'étude de cette fonction de matrice que nous allons faire dans ce qui va suivre.

6°) Détermination de γ et de $C(A)$.

Pour le calcul effectif de $C(A)$, on utilisera les déterminations suivantes de m_A et M_A.

I) M_A.

Soit $\Phi(Ax) = \sum_{i=1}^{i=n} |f_i(x)|$ et formons $r_{12} = \dfrac{\Phi(Ax)}{\|x\|}$

si je pose $Ax = X$ (A non singulière)

on aura: $r_{12} = \dfrac{\Phi(X)}{\|A^{-1}X\|}$

Je peux pour chercher le maximum de ce rapport me placer dans l'espace de point courant $X = (X_i)$ $(i = 1, \dots n)$

Soit Σ l'hyper-ellipsoïde d'équation $\|A^{-1}X\|^2 = 1$, $X^T (A^{-1})^T A^{-1} X = 1$

Si X_0 se déplace sur celui-ci, chercher le maximum de r_{12}, revient à chercher le maximun de $|X_1| + |X_2| + \dots + |X_n| = \Phi(X)$. Or les polyèdres $\Phi(X) = \Phi(X_0)$, varient tous en restant homothétiques et leur convexité assure que le maximum cherché est certainement obtenu lorsque l'une des faces de ce polyédre est tangente à Σ.

Or les faces de ces polyèdres, sont perpendiculaires aux vecteurs

$$E = \begin{vmatrix} \varepsilon_1 = \pm 1 \\ \varepsilon_2 = \pm 1 \\ \vdots \\ \varepsilon_n = \pm 1 \end{vmatrix}$$

Dont les composantes sont des ± 1 . (2^n tels vecteurs)

D'autre part, au point X_0 le plan tangent à Σ a pour équation:

$$X^T (A^{-1})^T A^{-1} X_0 = 1$$

Ce plan est perpendiculaire à E , s'il existe un scalaire k tel:

$$(A^{-1})^T A^{-1} X_0 = k.E$$

donc $X_0 = k.AA^TE$.

. Pour déterminer k , il suffit d'écrire que $X_0 \in \Sigma$

ce qui donne : $k^2 (AA^TE)^T (A^{-1})^T A^{-1} (AA^TE) = 1$

d'où $k^2 . E^T A A^T E = 1$, $k^2 = \dfrac{1}{\|A^TE\|^2}$

D'autre part, le plan $k . X^T E = 1$ ou $X_1 \varepsilon_1 + \dots + X_n \varepsilon_n = \dfrac{1}{k}$

coupe l'un des axes du système de référence en un point ℓ_i d'abscisse sur e_i ce point est tel que pour lui, $\Phi(\ell_i) = \frac{1}{|k|}$

par suite le maximum cherché, n'est autre que le maximum de $1/|k|$

d'où:

Théorème IV Le maximum M_A de γ n'est autre que $\max_E \| A^T E \|$ où E parcourt l'ensemble des vecteurs dont les composantes ont pour valeur absolue des I .

II) m_A

Pour déterminer m_A je considère toujours e_{12} mais je suppose déplacer le point X_1 sur le polyèdre $\Phi(X) = 1$
la valeur de $\| A^{-1} X_1 \|$ varie et, toujours à cause de la convexité le maximum de $\| A^{-1} X \|$ donc le minimum de γ ne peut être atteint qu'en un sommet de $\Phi(X) = 1$. Or ces sommets sont les points

$$e_i = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ 1 \leftarrow i^\circ \\ 0 \\ \vdots \\ 0 \end{vmatrix}$$

et la valeur de

$$m_A = \frac{1}{\max_i \| A^{-1} e_i \|}$$

d'où :

Théorème V Le minimum m_A de γ n'est autre que $1/\max_i \| A^{-1} e_i \|$, c'est à dire, l'inverse de la longueur de la colonne de plus grande longueur de la matrice inverse de A .

Par suite, on aura

$$(I) \qquad \boxed{C(A) = \frac{N(A)}{p_0(A)} \cdot \frac{1}{\max_i \| A^{-1} e_i \|} \cdot \frac{1}{\max_E \| A^T E \|}}$$

C'est cette formule qui en définitive nous servira à calculer des C(A), dès que l'inverse d'une matrice est calculable il est en effet assez commode de l'utiliser, bien que $\max_E \| A^T E \|$ soit, en général, assez difficile à évaluer.

Nous étudierons dans la suite des propriétés générales de C(A), soit au cours de transformations de A, soit au cours d'opérations matricielles sur A.

7°) Le conditionnement normalisé complémentaire C'(A) .

Nous allons rapidement voir comment on peut définir un autre nombre de conditionnement C'(A) qui a des propriétés analogues à celles de C(A) précédemment défini. Cette fois partons du rapport $r_{\omega 2}(x)$

on posera $r_{\omega 2} = \dfrac{m_b(Ax)}{\|x\|}$, $m'_A = \min_{x \neq 0} \left(r_{\omega 2}(x) \right)$, $M'_A = \max_{x \neq 0} \left(r_{\omega 2}(x) \right)$

$\gamma_{\omega 2} = \gamma' = \dfrac{m'_A}{M'_A}$

Comme précédemment étudions avec les mêmes notations le rapport γ'.

Pour déterminer m'_A et M'_A il suffit encore ,pour A non singulière de déplacer le point (x) sur le lieu des points (x) tels que $m_b(Ax) = 1$
Ce lieu est la frontière Π' d'un parallèlotope $\mathcal{P}'$, ayant l'origine pour centre de symétrie, les faces planes de celui-ci sont des parties des plans d'équation:

$$f_i(x) = \pm 1$$

Donc encore $m'_A = \dfrac{1}{d'_M}$ et $M'_A = \dfrac{1}{h'_m}$

mais ici,

I°) h'_m est très facile à déterminer: c'est $\min_i \left(\dfrac{1}{p_i} \right) = \dfrac{1}{\max_i (p_i)} = \dfrac{1}{p^\circ}$

2°) Pour pouvoir calculer d'_M il suffit de remarquer que les sommets de $\mathcal{P}'$ sont les points dont les coordonnées sont les nombres α_i du § 3°) dans $(O; \vec{e}_1, \vec{e}_2 \ldots \vec{e}_n)$

Si $V'_i = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ \alpha_i \leftarrow i^\circ \\ \vdots \\ 0 \end{vmatrix}$

Le vecteur $U' = \begin{vmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{vmatrix}$ joint O à un sommet de $\mathcal{P}'$

Le carré de la longueur est fourni par: $\ell^2 = U'^T . G . U'$

ou: $\ell^2 = \sum_{i=1}^{i=n} V_i'^T G V_i' + 2 \sum_{\substack{i<j \\ (1,\ldots n)}} V_i'^T G V_j$

D'après cela,

$$\ell^2 = \sum_{i=1}^{i=n} \frac{g^{ii}}{p_i^2} + 2 \sum_{i<j} \varepsilon_i \varepsilon_j \frac{\sqrt{g^{ii}} \cdot \sqrt{g^{jj}}}{p_i p_j} g_{ij}$$

puisque (d'après le même raisonnement que celui fait plus haut)

$$\ell_M^2 \geqslant \sum_{i=1}^{i=n} \frac{1}{\sin^2 \theta_i \cdot p_i^2}$$

l'on peut endéduire:

$$d'^2_M \geqslant \sum_{i=1}^{i=n} \frac{1}{p_i^2} = 1/N'^2(A) \quad (*)$$

(*) La notation $N'^2(A)$ n'est utilisée que par analogie; $N'(A)$ n'est pas une norme de matrice.

Par suite : $\frac{m'^2_A}{M'^2_A} = \frac{h'^2_m}{d'^2_M} = \frac{1}{\rho^{\circ 2}} \cdot \frac{1}{d'^2_M} \leqslant \frac{1}{(\rho^\circ)^2} \cdot \frac{1}{\sum_i \left(\frac{1}{p_i}\right)^2}$

d'où : $\frac{m'_A}{M'_A} \leqslant \frac{1}{\rho^\circ} \cdot \left(\frac{1}{\sum \left(\frac{1}{p_i}\right)^2} \right)^{\frac{1}{2}} = \frac{N'(A)}{\rho^\circ}$

Inversement, supposons que A soit telle que : $\frac{m'_A}{M'_A} = \frac{N'(A)}{\rho^\circ}$

puisque $M'_A = \rho^\circ$

et que toujours $m'^2_A = \frac{1}{d'^2_m} \leqslant 1 \Big/ \sum_{i=1}^{i=n} \frac{1}{\sin^2\theta_i \cdot p_i^2} \leqslant N^{-2}(A)$

C'est que $\sin^2 \theta_i = 1$ pour tout θ_i .

Donc, on peut aussi, en posant :

$$\begin{cases} C'(A) = 0 & \text{si A est singulière} \\ C'(A) = \frac{m'_A}{M'_A} \frac{\rho^\circ}{N'(A)} = \frac{m'_A}{N'(A)} & \text{si A est non singulière} \end{cases}$$

énoncer :

Théorème VI : On a toujours $0 \leqslant C'(A) \leqslant 1$

Pourque A soit singulière, il faut et il suffit que $C'(A) = 0$.

Pour que A soit orthogonale en lignes, il faut et il suffit que $C'(A) = 1$.

Si A tend vers une matrice singulière A_0, sans ligne nulle, alors $C'(A) \to 0$.

On peut appeler $C'(A)$ le conditionnement normalisécomplémentaire de $C(A)$

Etudions m'_A d'une autre façon :

$$m'_A = \min_x \left(\frac{\mathcal{M}(Ax)}{\|x\|} \right) = \min_X \left(\frac{\mathcal{M}(X)}{\|A^{-1}X\|} \right)$$

Si je déplace le point X sur la frontière de $\mathcal{M}(X) \leqslant 1$, le maximum de $\|A^{-1}X\|$ ne peut arriver que'en un sommet de ce polyèdre, dont les coordonnées sont les nombres

$$E = \begin{vmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{vmatrix} \qquad \text{donc} \qquad m'_A = \frac{1}{\operatorname{Max}_E \|A^{-1}E\|}$$

et

$$\boxed{C'(A) = 1 \Big/ N'(A) \cdot \operatorname{Max}_E \|A^{-1}E\|}$$

formule à rapprocher de celle du théorème V.

8°) Relations entre ces deux conditionnements.

Les rapports γ et γ' sont liés (théorème IV du chapitre III)

par : $\frac{1}{n}\gamma' \leq \gamma \leq n\gamma'$

ou $\frac{1}{n}\gamma \leq \gamma' \leq n\gamma$

Donc, on peut écrire :

$$\frac{1}{n}\gamma' \cdot \frac{\rho^\circ}{N'(A)} \leq \gamma \cdot \frac{N(A)}{\rho_\circ} \cdot \frac{\rho^\circ \cdot \rho_\circ}{N(A) \cdot N'(A)} \leq n\gamma' \frac{\rho^\circ}{N'(A)}$$

et, par suite,

$$\frac{1}{n} C'(A) \leq C(A) \cdot \frac{\rho_\circ \cdot \rho^\circ}{N(A) \cdot N'(A)} \leq n\, C'(A)$$

enfin :

$$\frac{1}{n} C'(A) \frac{N(A) \cdot N'(A)}{\rho_\circ \rho^\circ} \leq C(A) \leq n\, C'(A) \cdot \frac{N(A) \cdot N'(A)}{\rho_\circ \cdot \rho^\circ}$$

D'où les relations :

$$\frac{1}{n} C(A) \frac{\rho_\circ \cdot \rho^\circ}{N(A) \cdot N'(A)} \leq C'(A) \leq n\, C(A) \cdot \frac{\rho_\circ \cdot \rho^\circ}{N(A) \cdot N'(A)}$$

On peut remarquerque :

$$\frac{\rho_\circ}{\rho^\circ} \leq \frac{\rho_\circ \cdot \rho^\circ}{N(A) \cdot N'(A)} \leq \frac{\rho^\circ}{\rho_\circ}$$

Il résulte de ces relations que les deux conditionnements C(A) et C'(A) sont pratiquement d'utilisation équivalente avec, pour le calcul, une supériorité pour C(A) qui, d'après ce qui précède, se calcule plus simplement. C'est la raison pour laquelle cette fonction sera surtout utilisée dans la suite.

9°) Relations entre C(A) et les nombres de VON NEUMANN et de TURING.

I) Pour A, symétrique, définie positive, posons :

$$\mathfrak{L}(A) = \frac{\lambda_M}{\lambda_m}$$, nombre de VON NEUMANN,

(λ_M, λ_m : plus grande et plus petite des valeurs propres de A)

Comparons le à γ et C(A).

— a) Puisque : $\Phi(Ax) = \sum_{i=1}^{i=n} |f_i(x)|$

on a (cf. Chapitre II § I) $\|Ax\| \leq \Phi(Ax) \leq \sqrt{n}\, \|Ax\|$

donc : $$\frac{\|Ax\|^2}{\|x\|^2} \leq \frac{\Phi^2(Ax)}{\|x\|^2} \leq n \frac{\|Ax\|^2}{\|x\|^2}$$

par suite : $\frac{x^T A^T A x}{x^T . x} \leqslant m_A^2 \leqslant \frac{\Phi^2(Ax)}{\|x\|^2} \leqslant M_A^2 \leqslant n . \frac{x^T A^T . A x}{x^T . x}$

et, dans le cas où A est symétrique : $\lambda_m^2 \leqslant m_A^2 \leqslant M_A^2 \leqslant n . \lambda_M^2$

Donc $\frac{1}{n} \frac{\lambda_m^2}{\lambda_M^2} \leqslant \frac{m_A^2}{M_A^2}$ ou $\boxed{\frac{1}{\sqrt{n}} \leqslant P . \gamma}$ (I)

l'égalité est réalisée pour A = I

et, par suite $\frac{1}{P(A)} \leqslant \sqrt{n} . C(A) . \frac{P_o}{N(A)}$

b) Je dis que : $P . \gamma \leqslant \sqrt{n}$

Soit x le vecteur propre correspondant à la plus petite valeur propre, de composantes x_i (i=1,.....,n)

y " " " à la plus grande valeur propre, de composantes y_i (i=1,.....,m)

et ces deux vecteurs de module euclidien = 1

$$m_A \leqslant \lambda_m . \sum_i |x_i| \quad , \quad M_A \geqslant \lambda_M \sum_i |y_i|$$

$$\frac{m_A}{M_A} \leqslant \frac{\lambda_m}{\lambda_M} . \frac{\sum |x_i|}{\sum |y_i|}$$

Il est évident que :

$$1 \leqslant \sum_i |x_i| \leqslant \sqrt{n} \quad , \text{ si } \|x\| = 1$$

donc : $\frac{m_A}{M_A} \leqslant \frac{\lambda_m}{\lambda_M} . \sqrt{n}$, $\boxed{P . \gamma \leqslant \sqrt{n}}$ (II)

On aurait pu obtenir (I) et (II) des inégalités générales établies au chapitre III, Th. IV.

Si l'on considère :

$$A = \begin{vmatrix} n^2 & 0 & 0 & \cdots & 0 \\ 0 & (n+\varepsilon-1)\eta & -\eta & \cdots & -\eta \\ 0 & -\eta & (n+\varepsilon-1)\eta & & \vdots \\ \vdots & \vdots & & & -\eta \\ 0 & -\eta & -\eta & \cdots & (n+\varepsilon-1)\eta \end{vmatrix}$$

on peut choisir ε et η de sorte que $P(A) . \gamma(A)$ soit aussi près que l'on veut de $\sqrt{n}$

II) Soit le premier nombre de TURING

$$T_1(A) = n \cdot \max_{ij} |a_{ij}| \cdot \max_{ij} |\alpha_{ij}|$$

α_{ij} terme général de A^{-1}

Il est à remarquer d'abord que $T_1(A) \geqslant 1$

car $\sum_k a_{ik} \cdot \alpha_{ki} = 1$, implique

$$1 \leqslant n \cdot \max|a_{ik}| \cdot \max|\alpha_{ki}|$$

et à plus forte raison , $T_1(A) \geqslant 1$

- a) $$\frac{1}{\gamma} = \max_i \| A^{-1} e_i \| \cdot \max_j \| A^T E_j \|$$
$$\leqslant \sqrt{n} \max_{ij} |\alpha_{ij}| \sqrt{n} \cdot n \max_{ij} |\alpha_{ij}| \leqslant n T_1$$

(III) $$\boxed{\frac{1}{n} \leqslant T_1 \cdot \gamma}$$

- b) d'autre part

$$\max_i \| A^{-1} e_i \| \geqslant \max_{ij} |\alpha_{ij}|$$
$$\max_j \| A^T E_j \| \geqslant \max_{ij} |\alpha_{ij}|$$

donc :

$$\frac{1}{\gamma} \geqslant \frac{T_1}{n} \qquad \text{(IV)} \quad \boxed{T_1 \cdot \gamma \leqslant n}$$

Remarquons que des exemples simples prouvent que les bornes sont atteintes.

III) Soit enfin le 2ème nombre de TURING

$$T_2(A) = \frac{1}{n} N(A) \cdot N(A^{-1})$$

- a) Puisque

$$\max_i \| A^{-1} e_i \| \geqslant \frac{1}{\sqrt{n}} N(A^{-1})$$
$$\max_j \| A^T E_j \| \geqslant N(A)$$

on a :

(V) $$\boxed{T_2 \cdot \gamma \leqslant \frac{1}{\sqrt{n}}}$$

- b) Enfin

$$\max_i \| A^{-1} e_i \| \leqslant N(A^{-1})$$
$$\max_j \| A^T E_j \| \leqslant \sqrt{n} \cdot N(A)$$

donc : (VI) $$\boxed{T_2 \gamma \geqslant \frac{1}{n\sqrt{n}}}$$

En résumé, voici les inégalités obtenues :

$$\frac{1}{\sqrt{n}} \leqslant \rho \cdot \gamma \leqslant \sqrt{n} \qquad (I), (II)$$

$$\frac{1}{n} \leqslant T_1 \cdot \gamma \leqslant n \qquad (III), (IV)$$

$$\frac{1}{n\sqrt{n}} \leqslant T_2 \cdot \gamma \leqslant \frac{1}{\sqrt{n}} \qquad (V), (VI),$$

les inégalités pouvant être atteintes.

On obtient les relations avec C(A) en multipliant par $\frac{N(A)}{\rho_0}$ puisque $C(A) = \gamma(A) \cdot \frac{N(A)}{\rho_0}$

10°) OPérations laissant C(A) invariant .

Théorème VII Le conditionnement général γ_{ij} , relatif aux deux normes φ_i et φ_j , d'une matrice A ne change pas si A est remplacée par $\lambda . A$, pour λ scalaire non nul. Il en est de même pour C(A).

En effet si $m = \min_{x \neq 0} \left(\frac{\varphi_i(Ax)}{\varphi_j(x)} \right)$ et $M = \max_{x \neq 0} \left(\frac{\varphi_i(Ax)}{\varphi_j(x)} \right)$

on aura: $m' = \min_{x \neq 0} \left(\frac{\varphi_i(\lambda A x)}{\varphi_j(x)} \right) = |\lambda| m$, $M' = |\lambda| . M$

donc $\gamma'_{ij} = \frac{m'}{M'} = \frac{m}{M} = \gamma_{ij}$

De plus, dans le cas du conditionnement normalisé C(A)

$$\frac{\mu_0(A)}{N(A)} = \frac{\mu_0(\lambda A)}{N(\lambda A)} = \frac{|\lambda| \mu_0(A)}{|\lambda| N(A)}$$

donc :

$$C(A) = C(\lambda . A)$$

Théorème VIII Le conditionnement général γ_{ij} , relatif aux deux normes φ_i et φ_j , d'une matrice A ne change pas si A est remplacée par $Q_i A Q_j$ avec:

Q_i appartenant au groupe associé à la norme de vecteur φ_i

Q_j appartenant au groupe associé à la norme de vecteur φ_j

En effet si $r_{ij} = \frac{\varphi_i(Ax)}{\varphi_j(x)} = \frac{\varphi_i(Q_i A x)}{\varphi_j(x)}$

et puisque Q_j appartient au groupe associé à φ_j (cf chapitre II)

on aura $r_{ij} = \frac{\varphi_i(Q_i A Q_j y)}{\varphi_j(Q_j y)} = \frac{\varphi_i(Q_i A Q_j y)}{\varphi_j(y)}$

donc $r_{ij}(A, x) = r_{ij}(Q_i A Q_j, y)$

et par suite le rapport : $\gamma_{ij}(A) = \gamma_{ij}(Q_i A Q_j)$

Corollaire Le conditionnement normalisé C(A) est invariant si A est remplacée par J A Q

Si J est soit un produit de forme $(I - 2E_{ii}) . (I - 2E_{ee}) \cdots$

soit un matrice d'échange de deux lignes $V = (I - E_{ii} - E_{jj} + E_{ij} + E_{ji})$

(Produit de matrices du 2° degré bien particulières)

et si Q est une matrice orthonormale ($Q^T Q = I$)

En effet il est clair que $I - 2E_{ii} = J$ n'est autre que A où la i° ligne a un signe contraire . Donc quelque soit x,

$$\Phi(J . Ax) = \Phi(Ax)$$

et par suite J est du groupe associé à la norme $\Phi(x)$ de vecteurs

De même $V = I - E_{ii} - E_{jj} + E_{ij} + E_{ji}$ appartient à ce groupe .

Il est aussi évident que $\mu_0(JA) = \mu_0(A)$, $N(JA) = N(A)$.

D'autre part, le rapport $\gamma = \frac{m_A}{M_A}$ d'après le théorème précédent ne change pas si A est remplacée par J A Q.

Enfin, si ℓ_i est la longueur d'une ligne de A

on a $\ell_i^2 = A_{i.} \cdot A_{i.}^T$

puis $\ell_i^2 = A_{i.} \cdot Q \cdot Q^T \cdot A_{i.}^T = (A.Q)_{i.} \cdot (A.Q)_{i.}^T$

enfin : $\ell_i^2 = \ell_i'^2 \rightarrow \mu_0(A) = \mu_0(A.Q)$

Comme

$$N^2(A) = \text{trace}\,(A.A^T) = \text{trace}\,(A.Q.Q^T A^T) = N^2(AQ)$$

par suite le rapport $\frac{\mu_0(A)}{N(A)} = \frac{\mu_0(AQ)}{N(AQ)}$

Application : Signalons comme conséquence de ce corollaire une application que nous étudierons plus en détails dans le chapitre consacré à l'étude de la méthode d'orthogonalisation.

Dans la partie théorique relative à cette méthode on peut prouver le résultat suivant: Etant donnée une matrice A non singulière il existe toujours une matrice triangulaire inférieure et unitaire, T telle que T A = Q, avec Q orthogonale en ligne

soit $Q = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix}$ les q_i etant les lignes

de cette matrice, si je pose $d_{ii} = \| q_i \| = \sqrt{q_i \cdot q_i^T}$

et $D = \begin{vmatrix} d_{11} & & & \\ & d_{22} & & \\ & & \ddots & \\ & & & d_{nn} \end{vmatrix}$, $Q = D.Q'$

je vois que Q' est orthonormale

Donc $T.A = D.Q'$

et $A = T^{-1} D.Q' = T_1.Q'$

Théorème IX Le conditionnement normalisé $C(A)$ d'une matrice A (non singulière) est celui de la matrice triangulaire $C(T_1)$, qui se calcule aisément dans la méthode d'orthogonalisation.

11°) Conditionnement d'un produit de deux matrices.

Soient A et B deux matrices non singulières,

formons $$\frac{\Phi(A.Bx)}{\|x\|} = \frac{\Phi(AX)}{\|X\|} \cdot \frac{\|Bx\|}{\|x\|}$$

si $X = Bx$.

on sait que (thèorème V du Chapitre III) $\lambda_m^{\frac{1}{2}} \leq \frac{\|Bx\|}{\|x\|} \leq \lambda_M^{\frac{1}{2}}$

si λ_M, λ_m désignent les plus grandes et plus petites valeurs propres de la matrice $B.B^T$.

D'où $$m_A \cdot \lambda_m^{\frac{1}{2}} \leq m_{AB} \leq M_{AB} \leq M_A \cdot \lambda_M^{\frac{1}{2}}$$

et $$\frac{m_A}{M_A}\left(\frac{\lambda_m}{\lambda_M}\right)^{\frac{1}{2}} \leq \frac{m_{AB}}{M_{AB}} \quad , \quad \gamma_{12}(A).\gamma_{22}(B) \leq \gamma_{12}(A,B)$$

enfin $$C_{AB} \geq \left(\frac{\lambda_m}{\lambda_M}\right)^{\frac{1}{2}} . C_A . \frac{\rho_0(A)}{N(A)} . \frac{N(AB)}{\rho_0(AB)}$$

si l'on se rappelle que (Théorème IV du chapitre III)

$$\frac{1}{\sqrt{n}}\gamma_{22}(B) \leq \gamma_{12}(B) \leq \sqrt{n}\,\gamma_{22}(B)$$

on voit que :

(*) $$C_{AB} \geq \frac{C_A}{\sqrt{n}} \frac{m_B}{M_B} \frac{\rho_0(A)}{N(A)} . \frac{N(AB)}{\rho_0(AB)}$$

donc :

Theorème X Entre les conditionnements C_{AB} d'un produit A.B de deux matrices A et B non singulières on a la relation:

$$1 \geq C(AB) \geq \frac{C_A . C_B}{\sqrt{n}} \frac{\rho_0(A).\rho_0(B)}{\rho_0(A.B)} . \frac{N(A.B)}{N(A).N(B)} \qquad (1)$$

En particulier, pour que le produit de deux matrices soit orthogonal en lignes il suffit que:

$$\sqrt{n} = C_A . C_B . \frac{\rho_0(A).\rho_0(B)}{\rho_0(AB)} . \frac{N(A.B)}{N(A).N(B)}$$

Corollaire Entre les conditionnements de A et de son inverse A^{-1}, il est la relation:

(*) $$1 \geq C_A . C_{A^{-1}} \frac{\rho_0(A)}{N(A)} . \frac{\rho_0(A^{-1})}{N(A^{-1})}$$

En effet il suffit de remarquer que $\frac{N(I)}{\rho_0(I)} = \sqrt{n}$ et de faire $AB = I$ dans la formule (1).

(*) On montre aisément, sur des exemples simples, qu'il ne peut y avoir d'inégalité de sens inverse à celle-ci.

12°) Exemple de calcul d'un conditionnement normalisé, Matrice de HILBERT H_n

Définition: On appelle matrice de HILBERT d'ordre n la matrice symétrique de type (n,n) dont le terme général est

$$h^{(n)}_{ij} = \frac{1}{i+j-1} \qquad \text{pour} \quad i = 1,2 \ldots n \; ; \; j = 1,2 \ldots n$$

c'est à dire que

$$H_n = \begin{vmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & & & & \vdots \\ \frac{1}{4} & \vdots & & & & \\ \vdots & & & & & \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & & & \frac{1}{2n-1} \end{vmatrix}$$

Cette matrice a été très fréquemment étudiée cf. [19] [20] [21] [22] et aussi [15]. La raison en est que l'on sait formellement en déterminer l'inverse T_n, d'autre part l'inversion numérique de cette matrice, même pour de très faibles valeurs de n est très difficile. A ce sujet on se rapportera aux expériences de calcul de H_6^{-1}, que nous expliquerons dans le chapitre relatif à la résolution des systèmes linéaires .

On sait que si $T_n = H_n^{-1} = (t_{ij})$

$$t_{ij} = (-1)^{i+j} \frac{(n+j-1)!\,(n+i-1)!}{(i+j-1)\left[(i-1)!\,(j-1)!\right]^2 (n-i)!\,(n-j)!}$$

Donc :

$$(1) \qquad \left| \frac{t_{i,j+1}}{t_{ij}} \right| = \left(1 - \frac{1}{i+j}\right) \cdot \left(1 + \frac{n^2 - 2j^2}{j^2}\right)$$

Nous nous proposons dévaluer $C(H_n)$ pour des valeurs de n assez grandes .

- I°) Détermination de $\rho_0(H_n)$.

On a évidemment $\rho_0^2(H_n) = \frac{1}{n^2} + \frac{1}{(n+1)^2} + \cdots + \frac{1}{(2n-1)^2}$

en écrivant : $\frac{1}{n^2} = \frac{1}{n-1} - \frac{1}{n} - \frac{1}{n^2(n-1)}$

on voit que : $\rho_0^2(H_n) = \frac{1}{n-1} - \frac{1}{(2n-1)} - \sum_{i=n}^{i=2n-1} \cdot \frac{1}{i^2(i-1)}$

La somme du 2ème membre est inférieure à $\frac{n}{n^2(n-1)} = \frac{1}{n(n-1)}$

On peut donc dire que si n augmente : $\rho_0^2(H_n) \sim \frac{1}{n-1} - \frac{1}{2n-1} \sim \frac{1}{2n}$

2°) Détermination de $N(H_n)$

Ecrivons : $N^2(H_n) = 1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{3}\right)^2 + \cdots\cdots + \left(\frac{1}{n}\right)^2$

$+ \left(\frac{1}{2}\right)^2 + \left(\frac{1}{3}\right)^2 + \cdots\cdots + \left(\frac{1}{n}\right)^2 + \left(\frac{1}{n+1}\right)^2$

$+ \left(\frac{1}{3}\right)^2 + \cdots\cdots + \left(\frac{1}{n}\right)^2 + \left(\frac{1}{n+1}\right)^2 + \left(\frac{1}{n+2}\right)^2$

. . . . . .

$+ \left(\frac{1}{n}\right)^2 + \left(\frac{1}{n+1}\right)^2 + \cdots\cdots + \left(\frac{1}{2n-1}\right)^2$

Donc : $N^2(H_n) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots\cdots + \frac{1}{n} + \varphi(n)$

Regardons : $\varphi(n) = \frac{n-1}{(n+1)^2} + \frac{n-2}{(n+2)^2} + \cdots\cdots + \frac{n-(n-1)}{(2n-1)^2}$

en écrivant sous la forme $\varphi(n) = \frac{1}{n}\left[\frac{1-\frac{1}{n}}{\left(1+\frac{1}{n}\right)^2} + \frac{1-\frac{2}{n}}{\left(1+\frac{2}{n}\right)^2} + \cdots\cdots + \frac{1-\frac{n-1}{n}}{\left(1+\frac{n-1}{n}\right)^2}\right]$

On voit que si $n \to \infty$ cela a pour limite $\int_0^1 \frac{1-x}{(1+x)^2}\,dx = 1 - \log 2$

par suite, puisque $1 + \frac{1}{2} + \frac{1}{3} \cdots + \frac{1}{n} \sim \log n + C$

on peut écrire:

$$N^2(H_n) \sim 1 - \log 2 + C + \log n = A + \log n \quad , \quad A = 0{,}8841$$

3°)Détermination de $M_{H_n} = \underset{E}{\text{Max}} \|A^T E\|$, ou ici puisque H_n est symétrique $= \underset{E}{\text{Max}} \|AE\|$

$$M_{H_n}^2 = \underset{E}{\text{Max}} \left[\left(\varepsilon_1 + \frac{\varepsilon_2}{2} + \frac{\varepsilon_3}{3} + \cdots\cdots + \frac{\varepsilon_n}{n}\right)^2 + \left(\frac{\varepsilon_1}{2} + \frac{\varepsilon_2}{3} + \cdots + \frac{\varepsilon_n}{n+1}\right)^2 + \cdots\cdots + \left(\frac{\varepsilon_1}{n} + \frac{\varepsilon_2}{n+1} + \cdots\cdots + \frac{\varepsilon_n}{2n-1}\right)^2\right]$$

$$= \left[\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)^2 + \left(\frac{1}{2} + \frac{1}{3} + \cdots\cdots + \frac{1}{n+1}\right)^2 + \cdots\cdots + \left(\frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{2n-1}\right)^2\right]$$

En remarquant que $(\beta \neq 1)$

$$\frac{1}{\beta} + \frac{1}{\beta+1} + \cdots\cdots + \frac{1}{\beta+n-1} \sim \log(\beta+n-1) - \log(\beta-1) = \log\left(1 + \frac{n}{\beta-1}\right)$$

On peut écrire:

$$\left(\frac{1}{2} + \cdots + \frac{1}{n+1}\right)^2 + \cdots\cdots + \left(\frac{1}{n} + \cdots + \frac{1}{2n-1}\right)^2 \sim \left(\log\left(1+\frac{n}{1}\right)\right)^2 + \left(\log\left(1+\frac{n}{2}\right)\right)^2 + \cdots + \left(\log\left(1+\frac{n}{n-1}\right)\right)^2$$

Il suffit de remarquer que l'intégrale $\int_0^1 \left[\log\left(1+\frac{1}{x}\right)\right]^2 dx$

a un sens , Soit k sa valeur

pour pouvoir écrire

$$M_{H_n}^2 \simeq n.k + (C + \log n)^2$$

4°) Détermination de $m_{H_n} = 1/\operatorname{Max}_i \| A^{-1} e_i \|$

De la formule (1) on peut déduire avec TODD [15]

$$\operatorname{Max}_i \| T_n e_i \| \# n^{\frac{1}{2}} \max_{ij} (|t_{ij}|) \sim A_1 n^{-\frac{1}{2}} e^{nB} \quad , \quad B = 3,525$$

D'après ce qui précède on voit que une valeur asymptotique de $C(H_n)$

est $C(H_n) \simeq \alpha . n^{\frac{1}{2}} (\log n)^{\frac{1}{2}} e^{-nB}$

Cela montre la rapidité avec laquelle $C(H_n) \to 0$ et explique le mauvais conditionnement de H_n

13°) Cas d'une matrice d'interpolation d'opérateur différentiel

Nous allons retrouver le même genre de difficulté dans le cas d'un autre problème numérique. Soit à résoudre numériquement le problème: $y'' = F(x,y)$.

Résoudre l'équation différentielle $y'' = F(x,y)$, $0 \leqslant x \leqslant 1$

avec $y(0) = y_0$ et $y(1) = y_{n+1}$. (problème de conditions aux limites)

Il est classique de savoir que si l'on remplace l'équation proposée par le système linéaire

$$y_{i-1} - 2y_i + y_{i+1} = h^2 . F(x_i, y_i) \qquad (i = 1, \ldots n)$$

$$x_i = \frac{i}{n} \quad , \quad h = \frac{1}{n}$$

celui-ci peut s'écrire :

$$(1) \qquad A_n . Y = b$$

Avec

$$A_n = \begin{vmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{vmatrix}, \quad Y = \begin{vmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{vmatrix}, \quad b = \begin{vmatrix} y_0 - h^2 F_1 \\ -h^2 F_2 \\ \vdots \\ y_{n+1} - h^2 F_n \end{vmatrix}, \quad F_n = F(x_n, y_n)$$

Il n'est pas difficile de déterminer A_n^{-1} (par exple en triangularisant) On trouve ainsi :

$$A_n^{-1} = \frac{1}{n+1} \begin{vmatrix} 1.n & 1.(n-1) & 1.(n-2) & \cdots & 1.1 \\ 1.(n-1) & 2.(n-1) & 2.(n-2) & \cdots & 2.1 \\ 1.(n-2) & 2.(n-2) & 3.(n-2) & \cdots & 3.1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1.1 & 2.1 & 3.1 & & n.1 \end{vmatrix}$$

De cette expression on tire.

I°) $N^2(A_n) = 4+1+(n-2)(1+4+1)+4+1 = 2(3n-1)$

$$\mu_c^2(A) = 4+1 = 5$$

2°) $M_A = \max_E \|A^T E\| = \max_E \|AE\|$ si $E = \begin{vmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{vmatrix}$

$$\|AE\|^2 = (2\varepsilon_1 - \varepsilon_2)^2 + (-\varepsilon_1 + 2\varepsilon_2 - \varepsilon_3)^2 + (\varepsilon_2 + 2\varepsilon_3 - \varepsilon_4)^2 + \cdots + (-\varepsilon_{n-1} + 2\varepsilon_n)^2$$

Il est évident que le maximum est obtenu pour une suite de signes $\varepsilon_1, \varepsilon_2, \cdots \varepsilon_n$, qui alternent régulièrement.

et alors,

$$M_A^2 = 3^2 + (n-2)4^2 + 3^2 = 2(8n-7)$$

Il nous reste à étudier $m_A = \dfrac{1}{\max_i \|A^{-1} e_i\|}$

On peut écrire:

$$A_n^{-1} e_i = \frac{1}{n+1} \begin{vmatrix} 1 \cdot (n-i+1) & \\ 2 \cdot (n-i+1) & \\ \vdots & \\ i \cdot (n-i+1) & \leftarrow i^{\circ} \\ i \cdot (n-i) & \leftarrow (i+1)^{\circ} \\ \vdots & \\ i \cdot 1 & \end{vmatrix}$$

Donc :

$$\| A_n^{-1} e_i \|^2 = \frac{1}{(n+1)^2} \left[(n-i+1)^2 (1+2^2+\cdots+i^2) + i^2 (1+2^2+\cdots (n-i)^2) \right]$$

$$= \frac{1}{(n+1)^2} \left[(n-i+1)^2 \frac{(i+1)i(2i+1)}{6} + i^2 \frac{(n-i)(n-i+1)(2n-2i+1)}{6} \right]$$

$$= \frac{i(n-i+1)}{6(n+1)^2} \left[2i(n^2+n+1) + (n+1)(i-2i^2) \right]$$

Le 2° facteur est maximum pour $i \sim \frac{n}{2} + \frac{1}{2(n+1)} \sim \frac{n}{2}$

le I° facteur est maximum pour $i = \frac{n+1}{2} \sim \frac{n}{2}$

par suite $\max_i \| A_n^{-1} e_i \|^2 \sim \frac{n^2}{4.6(n+1)^2} \cdot \frac{n^3}{2} \sim \frac{n^3}{48}$

Donc Pour $C(A_n)$: $C(A_n) \sim \frac{1}{\sqrt{2(8n-7)}} \cdot \sqrt{\frac{48}{n^3}} \cdot \sqrt{\frac{2}{5}(3n-1)}$

$$C(A_n) \sim \frac{1,8}{\sqrt{n^3}}$$

On peut enconlure que l'inversion numérique du système $A_n y = b$ est d'autant plus difficile que n est grand .Or cette invesion est engénéral utilisée si lès F_i varient peu,oubien dans certaines méthodes itératives .On voit que les opérateurs aux différences fournissent des systèmes linéaires de plus en plus mal conditionnés et cela au fur et à mesure que l'on augmente l'ordre,c'està dire si l'on désire diminuer l'erreur de mèthode . Ce que l'on veut gagner d'un côté est perdu de l'autre. Cette décroissance est cependant beaucoup plus lente que dans des cas analogues à celui du paragraphe précédent.

CHAPITRE V

Les erreurs dans la résolution de systèmes linéaires par élimination

I°) Les principes de ces calculs d'erreurs.

La résolution de systèmes du premier degré est de toute première importance en calcul numérique. La plupart des procédés numériques ne consistent-ils pas à se ramener à la solution de systèmes linéaires ? Un tel problème ne comporte aucune erreur de méthode, seules des erreurs dues aux opérations élémentaires (sommes, différences, produits et quotients) peuvent intervenir et la répercussion, l'accumulation de telles erreurs sont d'une très grande importance.

Lorsque les moyens de calculs étaient limités aux machines de bureaux, le contrôle, à chaque opération, par le calculateur, des erreurs d'arrondis se faisait très facilement. Celui-ci se rendait compte à chaque instant de la valeur de ses résultats intermédiaires, pouvait par exemple, décider en cours de calcul, de changer le nombre de chiffes significatifs conservés par lui, pour remedier à une certaine "dérive" de son calcul global. Il faut ajouter qu'à ce moment des résolutions de systèmes linéaires d'ordre 20 étaient considérés comme des curiosités et bien rares sont les calculateurs en ayant résolus.

A l'heure actuelle, les calculateurs électroniques ont permis, de par leur extême rapidité, d'envisager et d'exécuter des résolutions de systèmes d'ordre extrèmement élevé (10.000 par exemple). Ces grands systèmes sont résolus à l'aide de programmes généraux pour ces calculateurs qui ne donnent pas de renseignements sur la valeur des résultats fournis par la machine. On peut penser combien, des résultats dépendant d'un aussi grand nombre d'opérations élémentaires peuvent être sensibles aux erreurs faites dans chacune de ces opérations.

Les calculs d'erreurs qui vont suivre sont basés, d'une part sur des hypothèses de travail relatives aux erreurs dans les opérations élémentaires, d'autre part, sur une méthode de détermination de la matrice d'erreur.

— I°) Il y a deux façons d'utiliser une machine àcalculer à programme relativement aux opérations élémentaires. L'une est dite "calcul en opérations à virgule fixe", l'autre est dite "calcul en opérations en virgule flottante " Nous ne pouvons pas entrer dans les détails mais l'on peut justifier les règles suivantes:

Règle I Si un calcul se fait en virgule fixe, tout se passe comme si il n'y avait pas d'erreur dans une somme algébrique, les erreurs dans les produits sont constantes = e , les erreurs dans les quotients sont aussi constantes = e'

Si a et b sont deux nombres

$\oplus, \ominus, \odot, \oslash$ désignent les opérations-machine correspondant à $+, -, \cdot, :$.

On écrira:

$$a \oplus b = a + b$$
$$a \ominus b = a - b$$
$$a \odot b = a \cdot b + e$$
$$a \oslash b = a : b + e'$$

Pour e et e' on prendra

$$e = k \cdot \beta^{-n}$$
$$e' = k' \cdot \beta^{-n'}$$

si β est la base de numération en laquelle la machine calcule n' et n désignant les nombres de chiffres conservés après la virgule dans ce calcul pour les divisions et pour les multiplications .Ce sont des constantes tout au long du calcul, c'est la raison de la dénomination"en fixe " de ce genre de calcul.

Règle II Si un calcul se fait en virgule flottante , tout se passe comme si Dans toute opération il y avait une erreur relative constante.

Si a et b sont deux nombres "flottants"(c'est à dire sur lequels la machine peut effectuer ces opérations flottantes).

$\boxplus, \boxminus, \boxdot, \boxslash$ désignent les opérations -machine correspondant à $+, -, \cdot, :$.

On écrira:

$$a \boxplus b = (1+\varepsilon)(a+b)$$
$$a \boxminus b = (1+\varepsilon)(a-b)$$
$$a \boxdot b = (1+\varepsilon')(a \cdot b)$$
$$a \boxdiv b = (1+\varepsilon')(a : b)$$

Pour ε et ε' on prendra

$$\varepsilon = k_1 \cdot \beta^{-N}$$
$$\varepsilon' = k'_1 \cdot \beta^{-N}$$

si β est la base denumération en laquelle la machine calcule et N est le nombre de chiffres de mantisse (ou capacité de larepresentation flottante)utilisés.

Nous répétons que les règles I et II sont des hypothèses de travail bien suffisantes d'ailleurs pour dire quelque chose de valable sur des programmes généraux. Il est d'ailleurs inextricable de vouloir utiliser des règles plus fines :les efforts de VON NEUMANN et GOLDSTINE [16] le prouvent suffisament. Nous reparlerons plus loin d'un prodédé (Le principe de simulation d'opérations) capable de doser la sensibilité d'un calcul particulier aux erreurs d'arrondis.

Voici maintenant la méthode de détermination de la matrice d'erreur :

Soit à résoudre le système linéaire:

$$A x = y$$

par exemple par la régle d'élimination de GAUSS.

On a déja dit un mot du procédé au chapitre I ,où l'on a mis en évidence le rôle important joué par des matrices du 2° degré.

Si M est une matrice (n,n) n'ayant que des termes nuls au dessous de sa diagonale dans les colonnes $1, 2, \ldots, k$, Il existe une matrice du 2° degré $f(M)$,telle que

$$M_1 = f(M).M$$

n'ait que des termes nuls au dessous de sa diagonale dans les colonnes $1, 2, \ldots k, k+1$.

On peut remarquer que si le calcul effectif donne $\overline{M}_1$ au lieu de M on peut déterminer δM telle que l'on ait exactement:

$$\overline{M}_1 = f(M+\delta M).(M+\delta M)$$

De proche en proche on montre que :

la solution approchée, obtenue en faisant des erreurs dans les opérations élémentaires, du système linéaire $A x = y$, peut être considérée comme la solution exacte d'un système $(A+\delta A).x = y+\delta y$

On donnera les expressions de δA et δy. Nous dirons que δA est la "matrice d'erreur" et δy le "second membre d'erreur"

2°) Calculs en fixe. Triangularisation. et résolution d'un système

I) Notations Nous raisonnerons sur l'algorithme de GAUSS général, c'est à dire sans tenir compte du cas de pivots nuls.

On doit multiplier A par $J_1, J_2, \dots\dots J_{n-1}$
de sorte que $J_{n-1}.J_{n-2}.\dots\dots J_2.J_1.A = \mathcal{C}$
soit une matrice triangulaire inférieure.

Alors le système $Ax = y$ à résoudre est remplacé par

$$J_{n-1}.J_{n-2}.\dots\dots J_2.J_1.Ax = J_{n-1}.J_{n-2}\dots J_2.J_1.y$$

ou $$\mathcal{C}x = \mathcal{Y}$$

Pour trouver $\mathcal{C}$ on pose $A^{(k+1)} = J_k.A^{(k)}$, avec

$$A^{(1)} = A \qquad (a^{(1)}_{ij} = a_{ij} \;;\; i,j = 1,\dots n)$$

alors si $A^{(k)} = (a^{(k)}_{ij})$

La matrice du 2° degré J_k est :

$$J_k = \begin{vmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ & & 1 & & & \\ & & & 1 & & \\ & 0 & & & \ddots & \\ & & & & & 1 \\ k^\circ\to & & & -\rho^{(k)}_{k+1} & 1 & \\ & 0 & & -\rho^{(k)}_{k+2} & & 1 \\ & & & \vdots & & \ddots \\ & & & -\rho^{(k)}_{n} & & 1 \end{vmatrix}$$

avec : $$\rho^{(k)}_i = \frac{a^{(k)}_{ik}}{a^{(k)}_{kk}} \qquad (i = k+1, \dots n)$$

Si bien que les formules de calcul sont:

$$A^{(k+1)} \begin{cases} a^{(k+1)}_{ij} = a^{(k)}_{ij} & , \quad i \le k \;,\; j < k \\ a^{(k+1)}_{ik} = 0 & , \quad i > k \\ a^{(k+1)}_{ij} = a^{(k)}_{ij} - \rho^{(k)}_i . a^{(k)}_{kj} & , \quad \begin{cases} i = k+1, \dots n \\ j = k+1, \dots n \end{cases} \end{cases}$$

et $y^{(k+1)}$ $\begin{cases} y_i^{(k+1)} = y_i^{(k)} & , \ i \leq k \\ y_i^{(k+1)} = y_i^{(k)} - \rho_i^{(k)} y_k^{(k)} & ; \ i = k+1, \ldots n \end{cases}$

2°) Calcul réel . Dans le calcul réel, l'existance d'erreur dans les opérations élémentaires conduit à une autre suite de matrices:

$\bar{A}^{(1)} = A$, $\bar{A}^{(2)}$, $\bar{A}^{(3)}$, ... $\bar{A}^{(n)} = \bar{\mathcal{C}}$

et une autre suite de 2° membres:

$\bar{y}^{(1)} = y$, $\bar{y}^{(2)}$, $\bar{y}^{(3)}$, ... $\bar{y}^{(n)} = \bar{y}$

$\bar{A}^{(1)}$ est identique à $A^{(1)}$, $\bar{y}^{(1)}$ à $y^{(1)}$, mais $\bar{\mathcal{C}}$ et $\bar{y}$ diffèrent de $\mathcal{C}$ et y . Cela provient de deux raisons :

- a) les matrices J_k sont mal déterminées et remplacées par $\bar{J}_k$ (car il y a des erreurs dans la détermination des nombres $\rho_i^{(k)}$) . Les termes $\rho_i^{(k)}$ de J_k sont remplacés par $\bar{\rho}_i^{(k)} = \rho_i^{(k)} + \varepsilon_i^k$, $(i = k+1, \ldots n)$ ε_i^k désignant l'erreur de division $a_{ik}^{(k)} : a_{kk}^{(k)}$.

- b) Le produit $J_k . A^{(k)}$ est remplacé par $\bar{J}_k \odot \bar{A}^{(k)}$, qui lui-même n'est réalisé qu'avec des erreurs de calcul on désignera par $-\eta_{ij}^{(k)}$ l'erreur dans les produits $\bar{\rho}_i^{(k)} . \bar{a}_{kj}^{(k)}$ $(i, j = k+1, \ldots n)$.

Donc les formules donnant les termes de $\bar{A}^{(k+1)}$ sont telles que

$$\bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - \bar{\rho}_i^{(k)} \bar{a}_{kj}^{(k)} + \eta_{ij}^{(k)}$$

De même pour les 2° membres, si l'erreur dans $\bar{\rho}_i^{(k)} \bar{y}_k^{(k)}$ est $-\mu_i^{(k)}$

$$\bar{y}_i^{(k+1)} = \bar{y}_i^{(k)} - \bar{\rho}_i^{(k)} \bar{y}_k^{(k)} + \mu_i^{(k)}$$

Conclusions : Les formules qui permettent d'obtenir la suite des $\bar{A}^{(k)}$, et des $\bar{y}^{(k)}$ sont :

$$\bar{A}^{(1)} = A \quad ; \quad \bar{y}^{(1)} = y^{(1)}$$

G.F.

$$\bar{\rho}_i^{(k)} = \left(\bar{a}_{ik}^{(k)} : \bar{a}_{kk}^{(k)} \right) + \varepsilon_i^k \qquad (i = k+1, \ldots n)$$

$$\begin{cases} \bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} & , \ i \leq k \ , \ j < k \\ \bar{a}_{ik}^{(k+1)} = 0 & , \ i > k \qquad (j = k) \\ \bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} + \eta_{ij}^{(k)} - \bar{\rho}_i^{(k)} \bar{a}_{kj}^{(k)} & (i, j = k+1, \ldots n) \end{cases}$$

$$\begin{cases} \bar{y}_i^{(k+1)} = \bar{y}_i^{(k)} & , \ i \leq k \\ \bar{y}_i^{(k+1)} = \bar{y}_i^{(k)} + \mu_i^{(k)} - \bar{\rho}_i^{(k)} \bar{y}_k^{(k)} & (i = k+1, \ldots n) \end{cases}$$

Si l'on remarque que les premières relations peuvent être écrites:

$$\bar{\rho}_i^{(k)} = \left(\bar{a}_{ik}^{(k)} + \varepsilon_i^{(k)}\,\bar{a}_{kk}^{(k)}\right) : a_{kk}^{(k)} \qquad (i = k+1, \ldots, n)$$

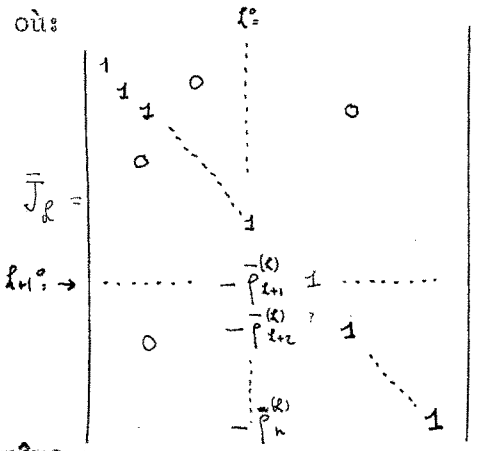
D'après ces formules on voit qu'on peut écrire :

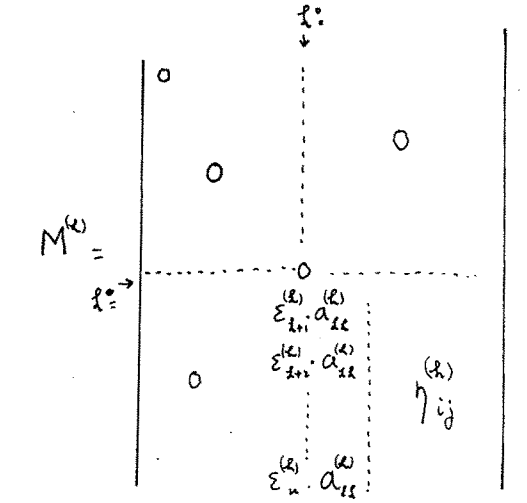
$$\bar{A}^{(k+1)} = \bar{\bar{J}}_k \cdot \bar{\bar{A}}^{(k)}$$

avec :

$$(1) \quad \bar{\bar{A}}^{(k)} = \bar{A}^{(k)} + M^{(k)}$$

où:





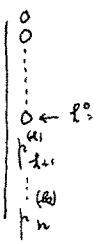
De même :

$$\bar{y}^{(k+1)} = \bar{\bar{J}}_k \cdot \bar{\bar{y}}^{(k)}$$

avec

$$\bar{\bar{y}}^{(k)} = \bar{y}^{(k)} + z^{(k)}$$

où:



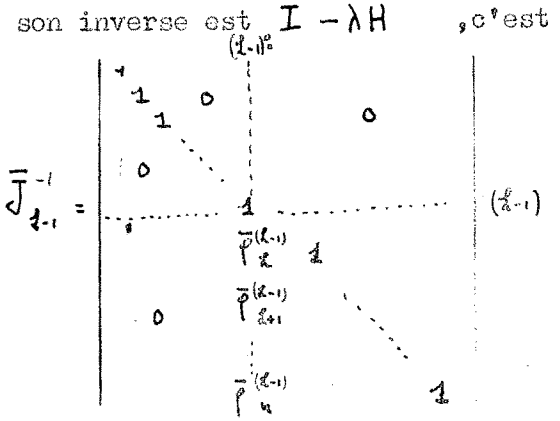
3°) Détermination de $\bar{\mathcal{C}}$.

De (1) on tire : $\bar{A}^{(k+1)} = \bar{\bar{J}}_k \left(\bar{A}^{(k)} + M^{(k)}\right) = \bar{\bar{J}}_k \left(\bar{\bar{J}}_{k-1} \left(\bar{A}^{(k-1)} + M^{(k-1)}\right) + M^{(k)}\right)$

$$= \bar{\bar{J}}_k \cdot \bar{\bar{J}}_{k-1} \left(\bar{A}^{(k-1)} + M^{(k-1)} + \bar{\bar{J}}_{k-1}^{-1} \cdot M^{(k)}\right)$$

Or $\bar{\bar{J}}_{k-1}$ est du 2° degré et de la forme $I + \lambda H$ avec $H^2 = 0$

donc son inverse est $I - \lambda H$,c'est à dire :



Or on a vu que la matrice $M^{(k)}$ a ses k premières lignes nulles, le produit $\bar{J}_{k-1}^{-1} \cdot M^{(k)}$ vu la forme de $\bar{J}_{k-1}^{-1}$ est donc identique à $M^{(k)}$.

Donc:

$$\bar{A}^{(k+1)} = \bar{J}_k \cdot \bar{J}_{k-1} \left(\bar{A}^{(k-1)} + M^{(k-1)} + M^{(k)} \right)$$

De même on prouve que:

$$\bar{A}^{(k+1)} = \bar{J}_k \cdot \bar{J}_{k-1} \cdots\cdots \bar{J}_1 \cdot \left(\bar{A}^{(1)} + M^{(1)} + \cdots\cdots + M^{(k-1)} + M^{(k)} \right)$$

Il en résulte

$$\bar{A}^{(n)} = \bar{\mathcal{C}} = \bar{J}_{k-1} \cdot \bar{J}_{k-2} \cdots\cdots \bar{J}_1 \left(A + \sum_{i=1}^{i=n-1} M^{(i)} \right)$$

D'où le thèorème:

<u>Théorème I</u> : Dans la triangularisation de la matrice A , si l'on commet dans les calculs les erreurs $\varepsilon_i^{(k)}$ et $\eta_{ij}^{(k)}$, On peut considérer la matrice triangulaire obtenue $\bar{\mathcal{C}}$ comme résultant de la triangularisation <u>sans erreur</u> de la matrice $A + \delta A$ avec $\delta A = \sum_{i=1}^{i=n-1} M^{(i)}$
les $M^{(i)}$ ayant les valeurs indiquées plus haut.

De même on voit que : $\bar{y}^{(k+1)} = \bar{J}_k \cdot \bar{J}_{k-1} \left(\bar{y}^{(k-1)} + z^{(k-1)} + z^{(k)} \right)$

et $\bar{y}^{(n)} = \bar{\bar{y}} = \bar{J}_{n-1} \cdots\cdots \bar{J}_1 \left(y + \delta y \right)$, $\delta y = \sum_{i=1}^{i=n-1} z^{(i)}$

4°) <u>Structures de la matrice δA et du vecteur δy .</u>

- a) Pour δA , il est facile de voir, si l'on pose : $a_{kk}^{(k)} = \pi_k$. $(k = 1 \cdots\cdots n)$

$$M^{(k)} = \pi_k \cdot E_k + H_k$$

avec :

$$E_k = \begin{vmatrix} & & & & \\ & & & & \\ & & & & \\ & & \varepsilon_{k+1}^{(k)} & & \\ & & \varepsilon_{k+2}^{(k)} & & \\ & & \vdots & & \\ & & \varepsilon_n^{(k)} & & \end{vmatrix} \leftarrow k+1^{e} \qquad , \qquad H_k = \begin{vmatrix} & & \downarrow (k+1)^{e} & & \\ & & & & \\ & & & & \\ & & \eta_{ij}^{(k)} & & \end{vmatrix} \leftarrow (k+1)^{e}$$

que: $\delta A = \sum_{k=1}^{k=n-1} \pi_k \cdot E_k + H_k$

et

$$\delta A = \begin{vmatrix} & & & & \\ \varepsilon_2^1 & & & & \\ \varepsilon_3^1 & \varepsilon_3^2 & & & \\ \varepsilon_4^1 & \varepsilon_4^2 & \varepsilon_4^3 & & \\ \vdots & & & \ddots & \\ \varepsilon_n^1 & \varepsilon_n^2 & \varepsilon_n^3 & \cdots\cdots & \end{vmatrix} \cdot \begin{vmatrix} \pi_1 & & & & \\ & \pi_2 & & & \\ & & \pi_3 & & \\ & & & \ddots & \\ & & & & \pi_n \end{vmatrix} + \begin{vmatrix} & & \downarrow j & & \\ & \eta_{22}' & \eta_{23}' & \cdots\cdots & \eta_{2n}' \\ & \eta_{32}' & & & \\ i \rightarrow & \sum_{k=1}^{k=p} \eta_{ij}^{k} & & & \\ & \eta_{n2}' & & & \end{vmatrix} \quad , \quad p = \mathrm{Min}(i,j) - 1$$

Par suite, si les erreurs de division sont de l'ordre de e' et celles de multiplication de l'ordre de e, cette matrice est de la forme:

$$\delta A \# e' \cdot \begin{vmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 1 & 1 & 1 & \ddots \\ \vdots & & & \\ 1 & 1 & 1 & \cdots \quad 0 \end{vmatrix} \cdot \begin{vmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \pi_n \end{vmatrix} + e \cdot \begin{vmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \vdots & 3 & 3 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & 3 \cdots & n-1 \end{vmatrix}$$

b) Pour δy, la valeur exacte est:

$$\delta y = \begin{vmatrix} 0 \\ \mu_2' \\ \mu_3' + \mu_3^2 \\ \vdots \\ \sum_{k=1}^{k=j-1} \mu_j^{(k)} \quad \leftarrow j \\ \vdots \\ \mu_n' + \mu_n^2 + \cdots + \mu_n^{n-1} \end{vmatrix}$$

et la valeur, en ordre de grandeur:

$$\delta y \# e \cdot \begin{vmatrix} 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \end{vmatrix} \quad *$$

5°) Retour-arrière . Soit alors à résoudre le système triangulaire:

$$\bar{\mathcal{C}}\, x = \bar{y}$$

avec:

$$\bar{\mathcal{C}} = \begin{vmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ & \alpha_{22} & \alpha_{23} & \cdots \\ & & \alpha_{33} & \\ 0 & & & \ddots \\ & & & \alpha_{nn} \end{vmatrix} \quad , \quad \bar{y} = \begin{vmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{vmatrix}$$

Les formules sont:

$$\xi_n = \frac{\gamma_n}{\alpha_{nn}} \quad ,$$

$$\vdots$$

$$\xi_k = \frac{1}{\alpha_{kk}} \left[\gamma_k - \left(\alpha_{k,k+1} \xi_{k+1} + \alpha_{k,k+2} \xi_{k+2} + \cdots + \alpha_{k,n} \xi_n \right) \right] ; \left(k = n-1, \cdots 1 \right)$$

Plus précisément, pour calculer ξ_k, $(k = n, n-1, \cdots 1)$ on "boucle" les différentes valeurs de S_k^h par:

$$S_k^{h+1} = 0 \quad , \quad S_k^i = S_k^{i+1} + \alpha_{k,i} \cdot \xi_i \qquad (i = n, \cdots, k+1)$$

$(\text{pour}, \ k = n, \cdots 1)$

* En supposant un nombre de chiffres différent dans la manipulation des seconds membres.

et alors $$\bar{\xi}_k = \frac{1}{\alpha_{kk}} \left[\gamma_k - S_k^{k+1} \right]$$

Mais le calcul effectif conduit non pas à des ξ_k mais à des $\bar{\xi}_k$

En effet les produits $\alpha_{ki} \cdot \bar{\xi}_i$ se font avec des erreurs $-\rho_{ki}$, si bien que l'on a, non pas des S_k^i, mais des $\bar{S}_k^i$.

Ceux-ci sont obtenus de sorte que l'on puisse écrire exactement:

$$\bar{S}_k^{k+1} = 0 \qquad \bar{S}_k^i = \bar{S}_k^{i+1} + \alpha_{k,i} \cdot \bar{\xi}_i - \rho_{ki}$$

et par suite:

$$\bar{\xi}_k = \frac{1}{\alpha_{kk}} \left[\gamma_k - \bar{S}_k^{k+1} \right] + d_k$$

d_k étant l'erreur de division.

d'où : $$\bar{\xi}_k = \frac{1}{\alpha_{kk}} \left[\gamma_k + \alpha_{kk} \cdot d_k + \rho_{k,k+1} + \rho_{k,k+2} + \cdots + \rho_{k,n} - \left(\alpha_{k,k+1} \bar{\xi}_{k+1} + \cdots + \alpha_{k,n} \bar{\xi}_n \right) \right]$$

Mais cette dernière formule prouve le théorème:

Théorème II Dans la résolution d'unsystème triangulaire $\mathcal{C} x = \bar{y}$ si l'on commet les erreurs d_k , ρ_{ki} , on peut considérer la solution obtenue comme la solution exacte du système triangulaire $\mathcal{C} . x = \bar{y} + \delta_1 y$ de même matrice de I° membre, avec :

$$\delta_1 \bar{y} = \begin{vmatrix} \pi_1 d_1 \\ \pi_2 d_2 \\ \vdots \\ \pi_n d_n \end{vmatrix} + \begin{vmatrix} \rho_{12} + \rho_{13} + \cdots + \rho_{1n} \\ \rho_{23} + \rho_{24} + \cdots + \rho_{2n} \\ \vdots \\ \rho_{k-2,n-1} + \rho_{n-2,n} \\ \rho_{n-1,n} \\ 0 \end{vmatrix} \# e'_1 \begin{vmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{vmatrix} + e_1 \begin{vmatrix} n-1 \\ n-2 \\ \vdots \\ 2 \\ 1 \\ 0 \end{vmatrix} \quad *$$

6°) Système complet .

D'après ce qui précède on peut enoncer le thèorème :

Théorème III La solution approchée du système $Ax = y$,et obtenue par un calcul en virgule fixe où les erreurs sont $\varepsilon_i^{(k)}$, $\eta_{ij}^{(k)}$, $\rho_i^{(k)}$ pour la triangularisation et d_k , ρ_{ki} pour le retour-arrière et par la méthode de GAUSS, peut être considérée comme la solution exacte du système $(A + \delta A) x = y + \delta y + \delta_1 \bar{y}$, $\delta A, \delta y$ et $\delta_1 \bar{y}$ ayant les valeurs exactes précisées plus haut ,et les valeurs, enordre de grandeur,

$$\delta A \# \begin{vmatrix} 0 & 0 & \cdots & 0 \\ \pi_1 & 0 & \cdots & 0 \\ \pi_1 & \pi_2 & 0 & \vdots \\ \pi_1 & \pi_2 & \pi_3 & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \pi_1 & \pi_2 & \pi_3 & \cdots & 0 \end{vmatrix} e' + e \begin{vmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & 2 \\ \vdots & 1 & 2 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & 3 & \cdots & n-1 \end{vmatrix}, \quad \delta y + \delta_1 \bar{y} \# e' \begin{vmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{vmatrix} + e \begin{vmatrix} n-1 \\ n-1 \\ \vdots \\ n-1 \end{vmatrix} \quad *$$

* Cf note page précédente

3°) Calculs en fixe. Inversion d'une matrice par la méthode de GAUSS.

Nous pouvons avec les formules qui précédent, obtenir, dés à présent, un résultat relatif à l'inversion numérique des matrices par la méthode de GAUSS.

On sait que l'inverse d'une matrice se détermine numériquement en résolvant les n systèmes linéaires

$$(1) : A x = e_1$$
$$(2) : A x = e_2$$
$$\vdots$$
$$(n) : A x = e_n$$

les 2° membres etant les vecteurs unités : $e_i = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ 1 \leftarrow i \\ 0 \\ \vdots \end{vmatrix}$

Or ces systèmes ont la même matrice de premier membre A .

D'autre part il est clair que :

Pour le système (1), les $\mu_i^{(1)}$ de la formule donnant δy sont nuls

Pour le système (2), les $\mu_i^{(1)}$, $\mu_i^{(2)}$ sont nuls,

etc...

Pour le système (n), les $\mu_i^{(1)}, \mu_i^{(2)}, \ldots\ldots \mu_i^{(n-1)}$ sont nuls.

Si bien que :

La solution approchée du système (1), est lasolution exacte d'un système de la forme:

$$(1)' : (A+\delta A).x = e_1 + e' \begin{vmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{vmatrix} + e \cdot \begin{vmatrix} n-1 \\ n-2 \\ n-2 \\ \vdots \\ n-2 \end{vmatrix}$$

La solution approchée du système (2) est la solution exacte d'un système de la forme:

$$(2)' : (A+\delta A).x = e_2 + e' \begin{vmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{vmatrix} + e \begin{vmatrix} n-1 \\ n-2 \\ n-3 \\ \vdots \\ n-3 \end{vmatrix}$$

etc.....

La solution approchée du système (n) est la solution exacte d'un système de la forme :

$$(n)' : (A+\delta A).x = e_n + e' \begin{vmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{vmatrix} + e \begin{vmatrix} n-1 \\ n-2 \\ n-3 \\ \vdots \\ 2 \\ 1 \\ 0 \end{vmatrix}$$

on en conclut le théorème:

Théorème IV — Inverser une matrice A par la méthode de GAUSS avec des erreurs de calcul c'est résoudre exactement : $(A+\delta A)X = I+\delta I$, au lieu de $A.X = I$

Avec

$$\delta A \# e' \cdot \begin{vmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ \eta_1 & 0 & 0 & 0 & \cdots & 0 \\ \eta_1 & \eta_2 & 0 & & & \\ \eta_1 & \eta_2 & & \ddots & & \\ \vdots & \vdots & & & & \\ \eta_1 & \eta_2 & & & \eta_{n-1} & 0 \end{vmatrix} + e \cdot \begin{vmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \vdots & 3 & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & \cdots & n-1 \end{vmatrix}$$

et

$$\delta I \# e' \begin{vmatrix} \eta_1 & \eta_1 & \cdots & \eta_1 \\ \eta_2 & \eta_2 & \cdots & \eta_2 \\ \eta_3 & \eta_3 & \cdots & \eta_3 \\ \vdots & \vdots & & \\ \eta_n & \eta_n & \cdots & \eta_n \end{vmatrix} + e \begin{vmatrix} n-1 & n-1 & \cdots & n-1 \\ n-2 & n-2 & \cdots & n-2 \\ n-2 & n-3 & \cdots & n-3 \\ n-2 & n-3 & & \vdots \\ \vdots & \vdots & & \\ n-2 & n-3 & \cdots 1 & 0 \; 0 \end{vmatrix}$$

- Nous allons procéder à une étude tout-à-fait analogue dans le cas de calculs conduits en point décimal flottant.

4°) Calculs en virgule flottante. Triangularisation et résolution d'un système .

- 1°) Les notations utilisées seront les plus voisines possibles de celles du § 2° , dans le calcul réel , l'existence d'erreurs dans les opérations élémentaires conduit à une suite de matrices,

$$\bar{A}^{(1)} = A, \; \bar{A}^{(2)}, \; \ldots\ldots\ldots \; \bar{A}^{(n)} = \bar{C}$$

et une suite de seconds membres:

$$\bar{y}^{(1)} = y, \; \bar{y}^{(2)}, \; \ldots\ldots\ldots \; \bar{y}^{(n)} = \bar{y}$$

$\bar{A}^{(1)}$ est identique à A , $\bar{y}^{(1)}$ à y , mais $\bar{C}$ et $\bar{y}$ diffèrent de C et y . Cela provient de deux raisons:

- a) Les matrices J_k sont mal déterminées et remplacées par $\bar{J}_k$ (car il y a des erreurs dans la détermination des nombres $\rho_i^{(k)}$) . Les termes $\rho_i^{(k)}$ de J_k sont remplacés par $\bar{\rho}_i^{(k)} = \rho_i^{(k)} + \varepsilon_i^{(k)} \cdot \rho_i^{(k)}$

$\varepsilon_i^{(k)}$ désignant l'erreur relative de division .

d'où: $$\bar{\rho}_i^{(k)} = \left(\bar{a}_{ik}^{(k)} + \varepsilon_i^{(k)} \bar{a}_{ik}^{(k)} \right) : \bar{a}_{kk}^{(k)}$$

b) Le produit $J_k . A^{(k)}$ est remplacé par $\bar{J}_k \boxdot \bar{A}^{(k)}$, qui lui-même n'est réalisé qu'avec des erreurs de calcul, on désignera dans ce cas par $\eta_{ij}^{(k)}$ l'erreur relative faite dans le calcul de $\bar{a}_{ij}^{(k+1)}$ donc on peut écrire:

$$\bar{a}_{ij}^{(k+1)} = \bar{a}_{ij}^{(k)} - \bar{\rho}_i^{(k)} \cdot \bar{a}_{kj}^{(k)} + \eta_{ij}^{(k)} \cdot \left(\bar{a}_{ij}^{(k+1)} \right)$$

De même , pour les secoonds membres , on désignera par $\mu_i^{(k)}$ l'erreur relative dans le calcul de $\bar{y}_i^{(k+1)}$. Si bien que :

$$\bar{y}_i^{(k+1)} = \bar{y}_i^{(k)} - \bar{\rho}_i^{(k)} \cdot \bar{y}_k^{(k)} + \mu_i^{(k)} \cdot \bar{y}_i^{(k+1)}$$

Conclusions : Les formules qui permettent, dans ce cas, d'obtenir la suite des $\bar{A}^{(k)}$, et des $\bar{y}^{(k)}$ sont:

$$\bar{A}^{(1)} = A \quad , \quad \bar{y}^{(1)} = y$$

G. Fe

$$\left[\begin{array}{l}
\bar{\rho}_i^{(k)} = \left(\bar{a}_{ik}^{(k)} + \varepsilon_i^{(k)} \cdot \bar{a}_{ik}^{(k)} \right) : a_{kk}^{(k)} \qquad , \quad (i = k+1, \ldots n) \\
\\
\left\{\begin{array}{l}
\bar{a}_{ij}^{(k+1)} = a_{ij}^{(k)} \qquad , \quad i \leq k \ , \ j < k \\
\bar{a}_{ik}^{(k+1)} = 0 \qquad , \quad i > k \ , \ (j = k) \\
\bar{a}_{ij}^{(k+1)} = \left(\bar{a}_{ij}^{(k)} - \bar{\rho}_i^{(k)} \cdot \bar{a}_{kj}^{(k)} \right)\left(1 + \eta_{ij}^{(k)}\right) \# \ \bar{a}_{ij}^{(k)} + \eta_{ij}^{(k)} \cdot \left(\bar{a}_{ij}^{(k+1)} \right) - \bar{\rho}_i^{(k)} \cdot \bar{a}_{kj}^{(k)} \qquad , \quad (i,j = k+1, \ldots n)
\end{array}\right. \\
\\
\left\{\begin{array}{l}
\bar{y}_i^{(k+1)} = \bar{y}_i^{(k)} \qquad , \quad i \leq k \\
\bar{y}_i^{(k+1)} = \left(\bar{y}_i^{(k)} - \bar{\rho}_i^{(k)} \bar{y}_k^{(k)} \right)\left(1 + \mu_i^{(k)}\right) \# \ \bar{y}_i^{(k)} + \mu_i^{(k)} \ \bar{y}_i^{(k+1)} - \bar{\rho}_i^{(k)} \bar{y}_k^{(k)} \qquad , \quad (i = k+1, \ldots n)
\end{array}\right.
\end{array}\right.$$

Comme nous l'avons fait nous écrivons ces formules :

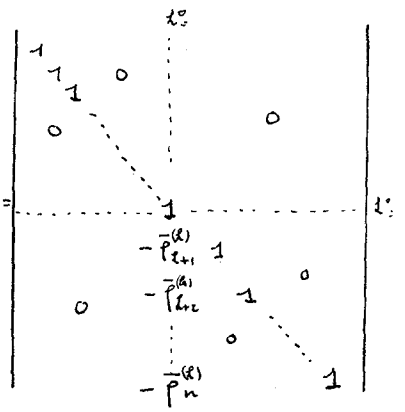
$$\bar{A}^{(k+1)} = \bar{J}_k \cdot \bar{\bar{A}}^{(k)}$$

avec:

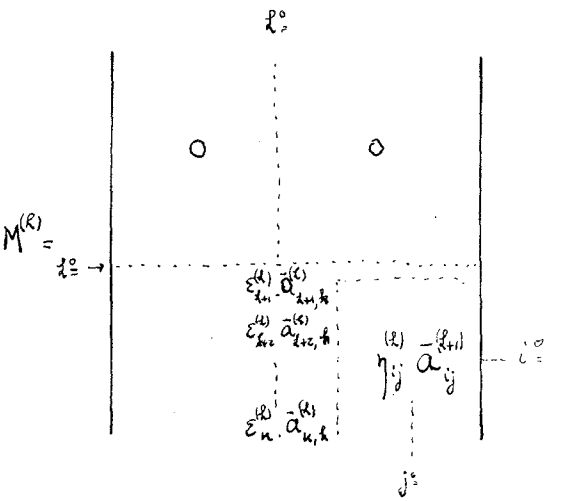
$$\bar{\bar{A}}^{(k)} = \bar{A}^{(k)} + M^{(k)}$$

où:

$\bar{J}_k =$



$M^{(k)} =$

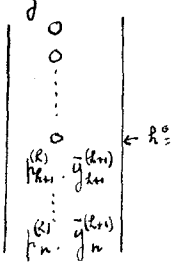


de même: $\bar{y}^{(k+1)} = \bar{J}_k \cdot \bar{\bar{y}}^{(k)}$

avec: $\bar{\bar{y}}^{(k)} = \bar{y}^{(k)} + z^{(k)}$

où:

$z^{(k)} =$



2°) Détermination de $\bar{\mathcal{C}}$.

De même que nous l'avons déja fait, on tire

$$\bar{A}^{(k+1)} = \bar{J}_k \left(\bar{A}^{(k)} + M^{(k)} \right) = \bar{J}_k \left[\bar{J}_{k-1} \left(\bar{A}^{(k-1)} + M^{(k-1)} \right) + M^{(k)} \right]$$

Par un raisonnement identique on prouve que cela peut être écrit:

$$\bar{A}^{(k+1)} = J_k . \bar{J}_{k-1} \left[\bar{A}^{(k-1)} + M^{(k-1)} + M^{(k)} \right]$$

et par suite :

$$\bar{A}^{(n)} = \bar{\mathcal{C}} = \bar{J}_{n-1} . \bar{J}_{n-2} \ldots\ldots \bar{J}_1 \left(A + \sum_{i=1}^{i=n-1} M^{(i)} \right)$$

Donc,

Théorème V Dans la triangularisation de la matrice A , si l'on commet dans les calculs élémentaires les erreurs relatives $\varepsilon_i^{(k)}$ et $\eta_{ij}^{(k)}$, on peut considérer la matrice triangulaire obtenue $\bar{\mathcal{C}}$ comme résultant de la triangularisation sans erreur de la matrice $A + \delta A$ avec $\delta A = \sum_{i=1}^{i=n-1} M^{(i)}$ et les $M^{(i)}$, les valeurs indiquées plus haut.

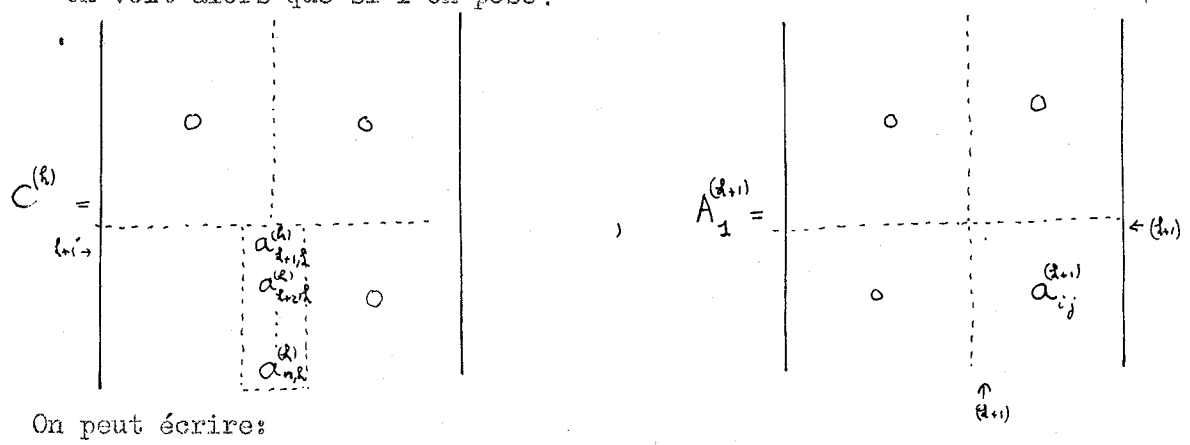
De même on voit que: $\bar{y}^{(k+1)} = \bar{J}_k . \bar{J}_{k-1} \left(\bar{y}^{(k)} + \zeta^{(k-1)} + \zeta^{(k)} \right)$

et $\bar{y}^{(n)} = \bar{\mathcal{Y}} = \bar{J}_{n-1} . \bar{J}_{n-2} \ldots\ldots \bar{J}_1 . (y + \delta y)$, $\delta y = \sum_{i=1}^{i=n-1} \zeta^{(i)}$

3°) Stuctures de la matrice δA et du vecteur δy .

Devant la complexité des formes obtenues nous ferons ici une hypothèse simplificatrice pour obtenir en ordre de grandeur les valeurs de δA et δy . Nous admettrons que les erreurs relatives $\varepsilon_i^{(k)}$, $\eta_{ij}^{(k)}$ et $\rho_i^{(k)}$ sont constantes (pour $\varepsilon_i^{(k)}$ cela résulte de la règle II du § I) et de l'ordre de $\varepsilon = k . \beta^{-N}$ (N est la capacité de la représentation flottante utilisée, β la base de numération) . On fait cette hypothèse dans ce §.

On voit alors que si l'on pose:



On peut écrire:

$$M^{(k)} = \varepsilon \left(C^{(k)} + A_1^{(k+1)} \right)$$

D'où $\delta A = \sum_{i=1}^{i=n-1} M^{(i)} = \varepsilon \sum_{i=1}^{i=n-1} C^{(i)} + \varepsilon \sum_{i=1}^{i=n-1} A_1^{(i+1)} = \varepsilon . T + \varepsilon . R$

Donc :

La matrice d'erreur δA , à ces approximations est la somme de $\varepsilon . T + \varepsilon . R$

T : matrice triangulaire à diagonale nulle des "colonnes éliminées "

R : matrice somme des matrices (rendues à l'ordre n) des systèmes intermédiaires.

Pour δy on posera de même: $\bar{y}_1^{(i+1)} = \begin{vmatrix} 0 \\ \cdots \\ 0 \\ \bar{y}_i^{(i+1)} \\ \bar{y}_{i+1} \\ \cdots \\ \bar{y}_n^{(i+1)} \end{vmatrix}$

et $\delta y = \varepsilon . \sum_{i=1}^{i=n-1} \bar{y}_1^{(i)}$

4°) Retour-arrière.

Si l'on doit alors résoudre le système: $\bar{C} x = \bar{y}$,

(mêmes notations que précédemment)

Le calcul conduit comme on le voit immédiatement à prendre :

$$\bar{S}_k^i = \bar{S}_k^{i+1} + \alpha_{k,i} . \bar{\xi}_i - \rho_{ki} . \bar{S}_k^i$$

si ρ_{ki} est l'erreur relative dans cette opération.

Et par suite

$$\bar{\xi}_k = \frac{1}{\alpha_{kk}} \left[\gamma_k - \bar{S}_k^{k+1} \right] + d_k . \bar{\xi}_k$$

d_k étant l'erreur relative dans la division.

d'où:

$$\bar{\xi}_k = \frac{1}{\alpha_{kk}} \left[\gamma_k + \alpha_{kk} . d_k \bar{\xi}_k + \rho_{k,k+1} . \bar{S}_k^{k+1} \cdots \rho_{k,n} \bar{S}_k^n - (\alpha_{k,k+1} \bar{\xi}_{k+1} + \cdots + \alpha_{k,n} \bar{\xi}_n) \right]$$

Théorème VI Dans la résolution d'un système triangulaire $\bar{C} x = \bar{y}$, si l'on commet les erreurs relatives ρ_{ki}, d_k , on peut considérer la solution obtenue comme la solution exacte du système triangulaire $\bar{C} x = \bar{y} + \delta_1 y$ demême matrice de I° membre, avec

$$\delta_1 y = \begin{vmatrix} \alpha_{11} d_1 \bar{\xi}_1 \\ \vdots \\ \alpha_{ii} d_i \bar{\xi}_i \\ \vdots \\ - \end{vmatrix} + \begin{vmatrix} \rho_{12} \bar{S}_1^2 + \rho_{13} \bar{S}_1^3 + \cdots + \rho_{1n} \bar{S}_1^n \\ \vdots \\ \rho \cdots \bar{S}_n^n \end{vmatrix}$$

5°) Système complet.

Il résulte de tout cela que l'on peut énoncer, comme plus haut :

Théorème VII : La solution approchée du système $Ax = y$ et obtenue par un calcul en virgule flottante où les erreurs sont, pour la triangularisation, $\varepsilon_i^{(i)}$, $\eta_{ij}^{(k)}$, $\rho_i^{(k)}$, et , d_i, $\rho_{ij}^{(k)}$ pour le retour-arrière, et par la méthode de GAUSS, peut être considérée comme la solution exacte du système : $(A + \delta A)\, x = y + \delta y$

$$\delta A = \varepsilon . T + \varepsilon . R$$

T matrice des colonnes éliminées,

R somme des matrices des systèmes intermédiaires

$$\delta y = \varepsilon . \sum \bar{y}_i^{(k)}$$

somme des seconds membres sous-diagonaux

Obtention de bornes pour l'erreur

Soit le système $A x = y$,résolu par la méthode de GAUSS,
Le calcul conduit à un vecteur solution x^* .
Celui-ci satisfait par suite du théorème III à l'égalité :

$$(A + \Delta A).x^* = y + \Delta y$$

où l'on a posé: $\delta A = \Delta A$

$$\delta y + \delta_1 y = \Delta y$$

D'où l'on tire :

$$(A + \Delta A).x^* = A x + \Delta y$$

puis ,

$$A.(x^* - x) = \Delta y - \Delta A.x^*$$

D'où la formule fondamentale :

$$(F) \qquad \boxed{x^* - x = A^{-1}(\Delta y - \Delta A\, x^*)}$$

D'autre part, soient φ_i et φ_j deux normes de vecteurs

on a : $\varphi_i(A(x^* - x)) = \varphi_i(\Delta y - \Delta A\, x^*)$

Donc $$\frac{\varphi_i(A(x^* - x))}{\varphi_j(x^* - x)} = \frac{\varphi_i(\Delta y - \Delta A x^*)}{\varphi_j(x^* - x)}$$

, si $m_{ij}(A)$ $M_{ij}(A)$ ont les significations du chapitre III.

$$m_{ij}(A) \leq \frac{\varphi_i(\Delta y - \Delta A x^*)}{\varphi_j(x^* - x)} \leq M_{ij}(A)$$

Il en résulte : $$\frac{\varphi_i(\Delta y - \Delta A x^*)}{M_{ij}(A)} \leqslant \varphi_j(x^* - x) \leqslant \frac{\varphi_i(\Delta y - \Delta A \cdot x^*)}{m_{ij}(A)}$$

soit: $$1/S_{ij}(A) \cdot \varphi_i(\Delta y - \Delta A x^*) \leqslant \varphi_j(x^* - x) \leqslant S_{ji}(A^{-1}) \cdot \varphi_i(\Delta y - \Delta A x^*)$$

Et si l'on veut se contenter d'une majoration supérieure de l'erreur:

$$\varphi_j(x^* - x) \leqslant S_{ji}(A^{-1}) \cdot \left[\varphi_i(\Delta y) + \varphi_i(\Delta A \cdot x^*)\right]$$

ou en introduisant le conditionnement général $\gamma_{ij}(A)$:

$$\frac{\varphi_i(\Delta y - \Delta A x^*)}{S_{ij}(A)} \leqslant \varphi_j(x^* - x) \leqslant \frac{1}{\gamma_{ij}(A)} \cdot \frac{\varphi_i(\Delta y - \Delta A x^*)}{S_{ij}(A)}$$

D'où le théorème:

Théorème VIII — Si φ_i et φ_j sont deux normes de vecteurs, S_{ij} la norme de matrice associée à ces deux normes, dans cet ordre, et γ_{ij} le conditionnement général relatif à celles-ci, on a pour la norme φ_j de l'erreur commise dans la résolution d'un système linéaire:

$$(1) \qquad \boxed{\frac{\varphi_i(\Delta y - \Delta A x^*)}{S_{ij}(A)} \leqslant \varphi_j(x^* - x) \leqslant \frac{1}{\gamma_{ij}(A)} \cdot \frac{\varphi_i(\Delta y - \Delta A x^*)}{S_{ij}(A)}}$$

Voici ce que donne cette formule dans les cas usuels:

I°) si $i = j = 2$, on a $$\frac{\|\Delta y - \Delta A x^*\|}{\sqrt{\lambda_n}} \leqslant \|x^* - x\|$$

puis:

$$(2) \qquad \|x^* - x\| \leqslant \frac{\|\Delta y - \Delta A x^*\|}{\sqrt{\lambda_1}}$$

2°) Si $i = 1$ et $j = 2$ on trouve:

$$(2') \qquad \frac{1}{M_A} \Phi(\Delta y - \Delta A x^*) \leqslant \|x^* - x\| \leqslant \frac{1}{m_A} \Phi(\Delta y - \Delta A x^*)$$

En introduisant le conditionnement normalisé $C(A)$,

$$(3) \qquad \frac{1}{M_A} \Phi(\Delta y - \Delta A x^*) \leqslant \|x^* - x\| \leqslant \frac{1}{C(A)} \frac{N(A)}{\rho_0(A)} \cdot \frac{1}{M_A} \cdot \Phi(\Delta y - \Delta A x^*)$$

Tout ce que nous venons de voir est valable quel que soit la façon dont les calculs sont conduits, je vais alors supposer qu'ils sont tout d'abord exécutés en fixe.

Utilisons tout d'abord, la formule (2)

On peut écrire:

$$\|x^* - x\| \leqslant \frac{1}{\sqrt{\lambda_1}} \|\Delta y\| + \frac{1}{\sqrt{\lambda_1}} \|\Delta A \cdot A^{-1} y\|$$

$x^* \# A^{-1} y$

Il est rappelé que dans ces relations λ_1 désigne la plus petite valeur propre de la matrice $A^T.A$.

Or il est immédiat de voir que:

$$\frac{\|\Delta A.A^{-1}y\|}{\|y\|} = \frac{\|\Delta A.X\|}{\|X\|} \cdot \frac{\|X\|}{\|AX\|} \quad , \quad y = AX$$

puis,

$$\|\Delta A.A^{-1}y\| \leq N(\Delta A) \cdot \frac{1}{\sqrt{\lambda_1}} \|y\|$$

donc:

$$\|x^* - x\| \leq \frac{1}{\sqrt{\lambda_1}} \|\Delta y\| + \frac{1}{\lambda_1} N(\Delta A) . \|y\|$$

Mais ici, d'après les formules du théorème III de ce chapitre,

$$N(\delta A) = N(\Delta A) \leq e' . N \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \eta_1 & 0 & & \vdots \\ \eta_1 & \eta_2 & & \vdots \\ \vdots & \vdots & & \vdots \\ \eta_1 & \eta_2 & & 0 \end{pmatrix} + e . N \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 1 \\ \vdots & 1 & \cdots & 2 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & n-1 \end{pmatrix}$$

Donc

$$N(\Delta A) \leq e' . R_1 + e . R_2$$

$$R_1^2 = (n-1)\eta_1^2 + (n-2)\eta_2^2 + \cdots + \eta_{n-1}^2$$

$$R_2^2 = (2n-3) + (2n-5).2^2 + \cdots + (n-1)^2 = \sum_{i=1}^{i=n-1} [2n-(2i+1)]i^2 = (2n-1)\sum_{i=1}^{i=n-1} i^2 - 2\sum_{i=1}^{i=n-1} i^3$$

$$\simeq \frac{1}{6} n(n-1)(n^2-n+1) \simeq \frac{n^4}{6}$$

Puis;

$$\|\Delta y\| \leq e' . R'_1 + e.(n-1)\sqrt{n}$$

$$R'^2_1 = \eta_1^2 + \eta_2^2 + \cdots + \eta_n^2$$

Enfin:

$$(4) \qquad \|x^* - x\| \leq \underbrace{\frac{\|y\|}{\lambda_1} R_1 . e'}_{①} + \underbrace{\frac{1}{\sqrt{\lambda_1}} . R'_1 . e'_1}_{②} + \underbrace{\frac{\|y\|}{\lambda_1} . R_2 . e}_{③} + \underbrace{\frac{(n-1)\sqrt{n}}{\sqrt{\lambda_1}} . e_1}_{④}$$

où les termes (1) et (2) sont relatifs aux divisions dans la triangularisation, et dans le retour-arrière

(3) et (4) sont relatifs aux multiplications dans la triangularisation et le retour-arrière

Intéressons-nous tout d'abord aux erreurs dues aux produits. La chose la la plus remarquable est de voir la croissance en ordre de n^2 due à R_2, autrement dit si n est grand, afin de pouvoir obtenir

une erreur sur la solution due aux multiplications dans la triangularisation inférieure à β^{-m} donné, il suffit de prendre e de sorte que :

$$\frac{1}{\sqrt{6}} \cdot e \cdot \frac{\|y\|}{\lambda_1} n^2 \leq \beta^{-m}$$

Or, on adit que pourdes calculs en fixe, $e = k \cdot \beta^{-p}$

si l'on fait $k = \frac{1}{2}$

on peut énoncer :

Théorème IX : Pour pouvoir résoudre un système linéaire en virgule fixe et pour un ordre n assez élevé afin que ce soit le 3ème terme de la formule (4) qui prédomine, et pour assurer une erreur sur la solution inférieure à β^{-m}, il suffit de conduire les calculs avec p chiffres après la virgule ; p satisfaisant à :

$$p \geq m - \log_\beta \left(\frac{4,9}{n^2} \frac{\lambda_1}{\|y\|} \right) \qquad (2\sqrt{6} \# 4,9)$$

Comme suite à cela, étudions la partie relative aux erreurs de division. C'est à dire le I° terme du second membre de la formule (4)

Deux cas sont à examiner :

I°) $\sqrt{\lambda_1}$ est plus grand que $\|y\|$, oubien l'ordre n est peu élevé de sorte que R'_1 et R_1 soient du même ordre de grandeur.

Dans ces cas les erreurs systèmatiques dûes aux divisions ont une borne supérieure dont le facteur prépondérant est

$$R'_1 = \left(\sum_{i=1}^{i=n} \pi_i^2 \right)^{\frac{1}{2}} \# \quad R_1 \quad \text{pour } e_1 \text{ et } e'_1 \text{ de même ordre.}$$

Comment affaiblir le plus possible ce facteur ?

Si l'on dispose d'un système $Ax = y$ à résoudre, un certain nombre d'opérations peuvent être faites, sans aucune erreur, qui substituent au système à résoudre, un autre dont la solution est celle du système proposé à l'ordre près des composantes: ce sont des échanges de lignes ou de colonnes dans le système donné ou dans les systèmes intermèdiaires.

Par exemple, l'échange de deux lignes dans le système proposé donne, après élimination de la première inconnue un système (d'ordre $n-1$) qui diffère de celui obtenu à partir du système initial, etc... Cela provient de la dissymétrie de la méthode de GAUSS. Il a toujours été important pour le calculateur d'avoir une règle de conduite, pour organiser son calcul afin d'obtenir la meilleure solution.

Nous pouvons remarquer qu'il y a n^2 façon de réorganiser, en ce sens, le système donné, puis $(n-1)^2$ façons de réorganiser le système d'ordre $(n-1)$ après elimination de la première inconnue, etc .... . Donc on peut dire que il y a en tout: $(n!)^2$ façonsde résoudre numériquement un système donné par la méthode de GAUSS et, en ce sens.

Dans chaque résolution, les "pivots " changent et l'on peut se demander comment les choisir de sorte que la solution obtenue soit la meilleure. La formule (4) permet d'essayer de répondre à cette question: Il suffit de s'assurer que

$$R'^2_1 = \pi_1^2 + \pi_2^2 + \cdots + \pi_n^2$$

est minimum.

Or d'après laproposition II du chapitre I, il est clair que le produitdes valeurs absolues des pivots $|\pi_1| . |\pi_2| \cdots |\pi_n|$ est égal à la valeur absolue du déterminant $|\Delta|$ de la matrice A du premier membre du système:donc ce produit est un nombre qui est toujours le même quelle que soit le procédé de résolution adopté.

Mais le minimun d'une somme de nombres positifs dont le produit est constant : $\pi_1^2 + \pi_2^2 + \cdots + \pi_n^2$

avec: $\pi_1^2 . \pi_2^2 \cdots \pi_n^2 = \Delta^2$

a lieu si lés nombres sont égaux, donc on peut dire que le facteur R'_1 est le plus petit possible si les pivots sont les plus voisins entre -eux . D'ou la règle:

Règle I Dans le cas où $\sqrt{\lambda_1}$ est plus grand que $\|y\|$, oubien que l'ordre n du système est assez faible pour que l'on puisse considérer R'_1 et R_1 du même ordre de grandeur , le choix le plus avantageux des pivots est celui qui les assure tous le plus près de l'égalité entre-eux et forcément alors peu différents de $\pm\sqrt[n]{|\Delta|}$ (en val. absolue) .[23]

2°) $\sqrt{\lambda_1}$ est plus petit que $\|y\|$ et l'ordre est assez élevé pour que le facteur qui prédomine soit R_1 .

Alors le meilleur choix est obtenu en essayant d'amoindrir

$$Z = (n-1)\pi_1^2 + (n-2)\pi_2^2 + \cdots + \pi_{n-1}^2$$

Soit $\mathcal{F}$ l'ensemble fini des nombres pouvant (convenir) comme valeur absolue de dernier pivot dans une résolution de la famille de celles envisagées . C'est à dire l'ensemble des nombres de la forme

$$\left|\frac{\Delta}{M}\right|$$

Où M est un mineur d'ordre $n-1$ de la matrice A .

Si un choix de dernier pivot est imposé les

$$|\pi_1|, |\pi_2|, \ldots\ldots, |\pi_{k-1}|$$

peuvent varier encore de sorte que

$$|\pi_1| \cdot |\pi_2| \cdot \ldots\ldots \cdot |\pi_{k-1}| = |M|$$

Parmi ceux-ci, ceux (qui) minimisent Z sont tels que :

$$\sqrt{k-1}\,|\pi_1| = \sqrt{k-2}\,|\pi_2| = \ldots\ldots = |\pi_{k-1}| \quad (1)$$

et dans ce cas Z_M vaudrait $(k-1)\,\pi_{k-1}^2$

et d'après (1) la valeur de ces pivots serait telle que

$$|M| = \frac{|\pi_{k-1}|^{k-1}}{\sqrt{(k-1)!}}$$

ou :

$$\pi_{k-1} = \left[(k-1)!\right]^{\frac{1}{2(k-1)}} \cdot |M|^{\frac{1}{k-1}}$$

Donc

$$Z_M = (k-1)\left[(k-1)!\right]^{\frac{1}{k-1}} \cdot |M|^{\frac{2}{k-1}}$$

Il y a donc à procéder ainsi:

règle II Dans le cas où $\sqrt{\lambda_1}$ est plus petit que $\|y\|$ et l'ordre assez élevé pour que le facteur qui prédomine dans la partie de la formule (4) relative aux erreurs de divisions, le choix le plus avantageux est obtenu de la façon suivante:

1° On "organise" la matrice de sorte que le mineur : qui tombe "en tête " de A (c'est à dire obtenu en supprimant la dernière ligne et la dernière colonne de A) soit le plus petit possible.

2° Le pivot n° I doit être pris le plus près possible de

$$\left[(k-1)!\right]^{\frac{1}{2(k-1)}} \cdot |M|^{\frac{1}{k-1}} \cdot (k-1)^{-\frac{1}{2}}$$

Le pivot n°2 doit être pris le plus près possible de

$$\left[(k-1)!\right]^{\frac{1}{2(k-1)}} \cdot |M|^{\frac{1}{k-1}} \cdot (k-2)^{-\frac{1}{2}}$$

etc.....

Remarquons, qu'au lieu d'utiliser la formule (4) faisant intervenir une majoration de l'erreur, on peut, d'après la formule fondamentale

$$(F) : \quad x^* - x = A^{-1} \cdot \left[\Delta y - \Delta A \cdot x^*\right]$$

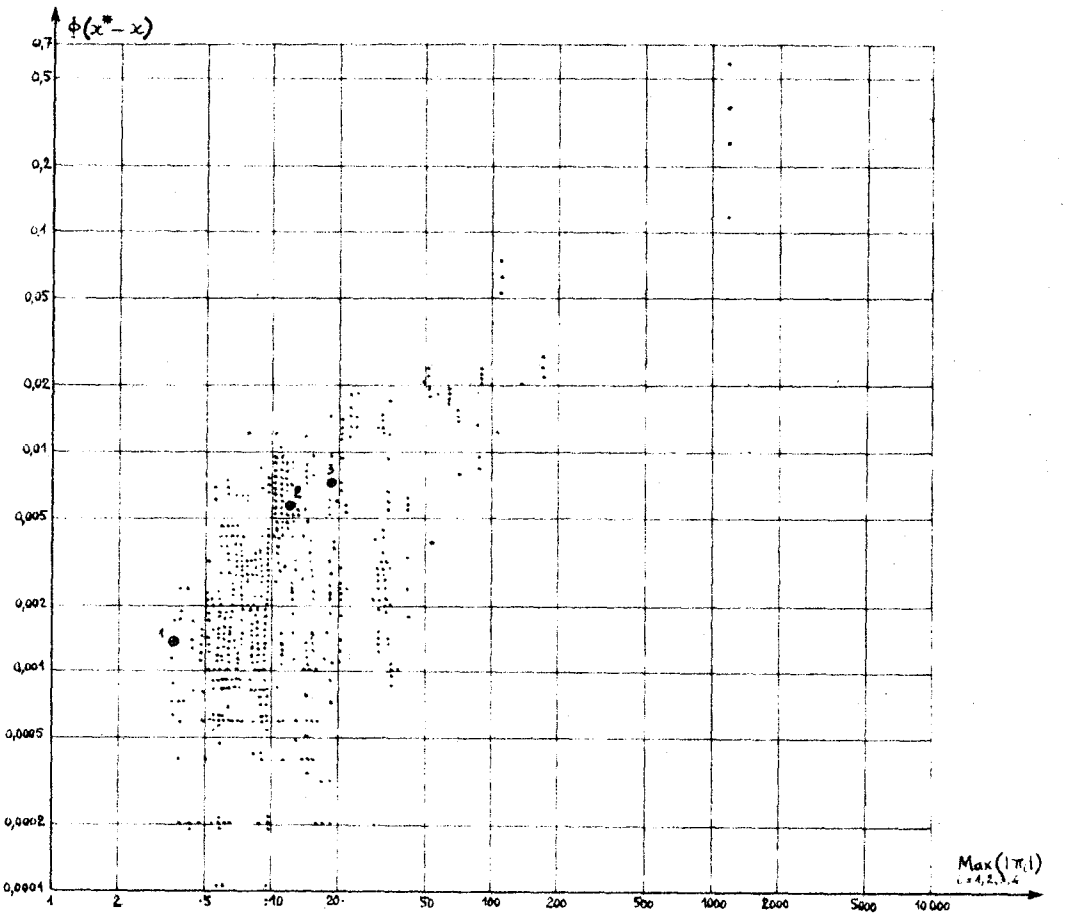
et d'après les expressions de ΔA et Δy trouvées au théorème III, montrer que l'affaiblissement de Δy et ΔA est réalisé si l'on évite les pivots par trop grands (ou par trop petits à cause de la constance du produit).

6°) Expériences de calcul en fixe

Afin de pouvoir juger de la validité des règles ci-dessus, nous avons fait les calculs suivants.

On a résolu le système A x = y , avec :

$$A = \begin{vmatrix} 0,053 & 2,0135 & 1,0172 & 1,4521 \\ 1,2142 & 1,5676 & 5,0431 & 1,5834 \\ 2,0534 & 7,9415 & 6,7123 & 2,1421 \\ 3,9835 & 9,1121 & 11,3451 & 10,4520 \end{vmatrix} \quad \text{et } y = \begin{vmatrix} 12,94 \\ 25,8123 \\ 46,6417 \\ 98,051 \end{vmatrix}$$



Pour $A.A^T$ on a : $\lambda_1 = 0,167$; $\lambda_4 = 469,5$; $\sqrt{\lambda_1} = 0,409$; $\|y\| \# 112,7$

On est donc dans le cas où la règle I s'applique (n =4 est faible)

Les calculs ont été exécutes en virgule fixe ,à quatre décimales après celle-ci ,sur la machine électronique" Gamma E.T." de l' Université de Grenoble pour laquelle nous avons réalisé un programme capable de faire les $(4!)^2 = 576$ résolutions de ce système .
Il faut environ trois heures-machine pour terminer ce calcul.
Le graphique de la page ci-contre résume les résultats obtenus.
Le système a été formé de sorte que la solution exacte soit la colonne des nombres I,2,3,4 .

Nous avons porté en abscisse le maximum de la valeur absolue des pivots relatif à une résolution,et en ordonnée,pour cette même résolution,la valeur de $\Phi(x' - x)$,norme de l'erreur.

Le point marqué I de ce graphique est relatif à la résolution donnant les pivots les plus près possibles de $\sqrt{|\det A|}$, c'est la résolution suivant la Règle I ci-dessus.

Le point marqué 2 est relatif à la règle habituelle du "pivot maximal" c'est à dire que les systèmes successifs sont toujours réorganisés de sorte que le plus grand élément ,en valeur absolue, du système partiel tombe en tête.

Le point 3 est relatif à une résolution suivant la règle contraire à la précédente:placer en tête le plus petit des éléments relatifs,en valeur absolue, des systèmes intermédiaires.

Enfin le graphique montre combien les variations de l'erreur sont grandes et qu'il faut éviter à tout prix d'utiliser de trop grands pivots (et par contre coup,à cause du produit constant, de trop petits) .

Il faut signaler que ce calcul n'a pu se faire sur la machine électronique qu'en utilisant ce que nous appelons des "simulations d'opérations "

En effet ,la programmation en fixesur un calculateur est très difficile si l'on ne connait à l'avance la position de la virgule dans tous les résultats intermédiaires ,ce qui est le cas dans un tel calcul. Par contre ,la programmation en virgule flottante est très aisée .Or dans la plus part des machines l'ordre de commande de l'opération flottante est un ordre d'appel d'un sous-programme , Il suffit donc d'écrire un programme en virgule flottante,de substituer aux sous-programmes d'opérations flottantes dessous-programmes d'opérations déterminées (en fixe dans notre cas,mais pouvant ,pour l'étude d'autres questions être des opérations d'autre nature: virgule fixe à capacité variable, avec ou sans arrondi ,virgule flottante à capacité variable,avec ou sans arrondi) pour obtenir ce quel'on désire.

7°) Calculs en virgule flottante , Evaluation de l'erreur .

Nous chercherons ici une majoration convenable pour une erreur relative dans la solution.

Avec les mêmes notations qu'au § 5 , le Théorème VIII est encore valable .

On peut écrire,

$$(1) \qquad \varphi_j(x^* - x) \leq \frac{1}{\gamma_{ij}(A)} \cdot \frac{\varphi_i(\Delta y - \Delta A\, x^*)}{S_{ij}(A)}$$

ou

$$(2) \qquad \| x^* - x \| \leq \frac{1}{\sqrt{\lambda_1}} \| \Delta y - \Delta A\, x^* \|$$

ou

$$(2)' \qquad \| x^* - x \| \leq \frac{1}{m_A} \Phi(\Delta y - \Delta A\, x^*)$$

Mais ici, il est plus difficile d'obtenir des ordres de grandeurs corrects pour ΔA et Δy .

Pour $\Delta A = \delta A$, sa structure est déterminée par ce que l'on a dit au §4, 3°.

$$\boxed{\Delta A = \varepsilon_1 (T + R)}$$

Pour δy , d'après toujours ce même § , on peut remarquer (mêmes notations)

$$A_1^{(k+1)} \cdot x^* \ \# \ \bar{y}_1^{(k+1)}$$

donc :

$$\delta y = \varepsilon \cdot \sum_{i=1}^{i=n-1} \bar{y}_1^{(i)} \ \# \ \varepsilon \cdot R \cdot x^*$$

Pour $\delta_1 y$ D'après §4, 4°; on peut écrire

$$\delta_1 y = \varepsilon \cdot \begin{vmatrix} \alpha_{11}\bar{\xi}_1 \\ \alpha_{21}\bar{\xi}_2 \\ \vdots \\ \alpha_{n1}\bar{\xi}_n \end{vmatrix} + \varepsilon \cdot \begin{vmatrix} \bar{S}_1^2 + \bar{S}_1^3 + \cdots + \bar{S}_1^n \\ \bar{S}_2^3 + \bar{S}_2^4 + \cdots + \bar{S}_2^n \\ \vdots \\ \bar{S}_{n-1}^n \\ 0 \end{vmatrix}$$

Or, si :

$$\bar{C}_1 = \begin{vmatrix} \alpha_{11} & \alpha_{12} & 2\alpha_{13} & 3\alpha_{14} & \cdots & (n-1)\alpha_{1n} \\ 0 & \alpha_{22} & \alpha_{23} & 2\alpha_{24} & \cdots & (n-2)\alpha_{2n} \\ 0 & \cdots & \alpha_{33} & \cdots & & \vdots \\ \vdots & & & & \alpha_{n-1,n-1} & \alpha_{n-1,n} \\ 0 & \cdots & & & 0 & \alpha_{nn} \end{vmatrix}$$

Donc :

$$\delta_1 y = \varepsilon \bar{C}_1 \cdot x^*$$

Il en résulte que en ordre de grandeur on peut dire que:

$$\boxed{\Delta y = \delta y + \delta_1 y = \varepsilon (R + \bar{C}_1)\, x^*}$$

De la formule(1) on tire donc :

$$\frac{\varphi_i(x^*-x)}{\varphi_i(x^*)} \leqslant \frac{1}{\gamma_{ij}(A)} \cdot \varepsilon \cdot \frac{1}{S_{ij}(A)} \cdot \left(\frac{\varphi_i((R+\bar{C}_1)x^*)}{\varphi_i(x^*)} + \frac{\varphi_i((R+T)x^*)}{\varphi_j(x^*)} \right)$$

puis,

$$\frac{\varphi_i(x^*-x)}{\varphi_i(x^*)} \leqslant \frac{\varepsilon}{\gamma_{ij}(A)} \cdot \frac{1}{S_{ij}(A)} \left(S_{ij}(R+\bar{C}_1) + S_{ij}(R+T) \right)$$

enfin,

$$\frac{\varphi_i(x^*-x)}{\varphi_i(x^*)} \leqslant \frac{\varepsilon}{\gamma_{ij}(A)} \cdot \frac{2S_{ij}(R) + S_{ij}(T) + S_{ij}(\bar{C}_1)}{S_{ij}(A)}$$

qui est la majoration en erreur relative cherchée.

Dans les cas particuler qui nous intéressent,

$$\frac{\|x^*-x\|}{\|x^*\|} \leqslant \frac{\varepsilon}{m_A} \cdot \left(2M_R + M_T + M_{\bar{C}_1} \right)$$

et

$$\frac{\|x^*-x\|}{\|x^*\|} \leqslant \frac{\varepsilon}{\sqrt{\Lambda_1}} \left(2\sqrt{\Lambda_R} + \sqrt{\Lambda_T} + \sqrt{\Lambda_{\bar{C}_1}} \right)$$

Si Λ_R, Λ_T, $\Lambda_{\bar{C}_1}$ désignent les plus grandes valeurs propres des matrices $R^T.R$, $T^T.T$, $\bar{C}_1^T.\bar{C}_1$.

En particulier, soit $\mathcal{M}$ le plus grand des nombres (en valeur absolue)qui peuvent être obtenus dans le calcul, c'est à dire pouvant être un élément des matrices $A^{(k)}$.

Puisque

$$M_R \leqslant n\sqrt{n}.(n-1)\mathcal{M}$$

et

$$M_T \leqslant n\sqrt{n}\,\mathcal{M}$$

et

$$M_{\bar{C}_1} \leqslant n\sqrt{n}\,(n-1)\mathcal{M}$$

on peut dire que: $$\frac{\|x^*-x\|}{\|x^*\|} \leqslant \frac{\varepsilon}{m_A} \cdot \mathcal{M} \cdot \left(2n-2+1+n-1 \right) n^{\frac{3}{2}}$$

Donc : $$\frac{\|x^*-x\|}{\|x^*\|} \leqslant (3n-2)n^{\frac{3}{2}} \cdot \frac{\varepsilon}{m_A} \cdot \mathcal{M}$$

Thèorème X Pour pouvoir résoudre en virgule ~~fixe~~ flottante un système linéaire, afin d'assurer au résultat une précision relative inférieure à β^{-p} il suffit de conduire les calculs avec une virgule flottante ayant une capacité de mantisse N satisfaisant à:

$$(3n-2).n^{\frac{3}{2}}\ \beta^{-N} \cdot \frac{\mathcal{M}}{m_A} \leqslant \beta^{-p}$$

$$\boxed{N \geqslant p - \log_\beta \left(\frac{m_A}{\mathcal{M}} \cdot \frac{1}{(3n-2).n^{\frac{3}{2}}} \right)}$$

On peut remarquer que dans cette formule intervient seulement le nombre m_A / m_b qui est presque de même nature que les conditionnements généraux que nous avons définis. On peut appeler ce nombre " conditionnement global "pour le système à résoudre

3°) Expériences de calcul en virgule flottante.

On peut se demander si les évaluations d'erreurs qui sont données dans ce qui précède ne sont pas trop peu précises en ce sens que les inégalités qu'elles font intervenir ne sont pas assez "sérrées".

Il est donc indispensable de faire des expériences de calcul pour savoir ce qu'il en est.

Nous avons considéré le système suivant: $A x = y$,

avec

$$A = \begin{vmatrix} 6! & 0 & \cdots & 0 \\ 0 & 7! & & \vdots \\ \vdots & & 8! & \\ & & & 9! \\ & & & 10! \\ 0 & \cdots & & 11! \end{vmatrix} \cdot \begin{vmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{3} & \cdots & & & \frac{1}{7} \\ \frac{1}{3} & & & & & \\ \vdots & & & & & \vdots \\ \frac{1}{6} & \cdots & & & & \frac{1}{11} \end{vmatrix}, \quad y = \begin{vmatrix} 6! \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{vmatrix}$$

La solution exacte de ce système n'est autre que la première colonne :

$$x = \begin{vmatrix} 36 \\ -630 \\ 3.360 \\ -7.560 \\ 7.560 \\ -2.772 \end{vmatrix}$$

de l'inverse de la matrice H_6 de HILBERT;

Dans ce cas, on trouve:

$m_A \# 2$

$m_b \# 3 \cdot 10^6$

$(3n-2) \cdot n^{\frac{3}{2}} \# 2,3 \cdot 10^2$

Si bien que la formule du théorème X s'écrit:

$$N \geq t - \log_{10}\left(\frac{2}{3} \cdot 10^{-8} \cdot \frac{1}{2,3}\right) \# t + 8 + 0,65$$

On voit d'après cela que le nombre de chiffres significatifs à la mantisse du vecteur solution doit être à peu près:

$$P \# N - 8,65$$

Pas de chiffre significatif

H_6

Valeur exacte	3	4	5	6	7	8	9 PDF BULL	9 PDF avec arrondi
36	22,9	12,38	15,456	49,327 7	36,526 11	36,067 416	36,009 469 4	35,996 275 0
− 630	−205	− 60,26	− 82,547	− 998,547	− 644,152 3	− 631,850 95	− 630,253 404	− 629,901 619
3 360	477	− 43,12	− 167,93	5 808,43	3 451,857	3 372,223 6	3 361,643 28	3 359,366 98
−7 560	−252	480,8	1 289,0	−13 852,7	−7 791,998	−7 591,270 2	−7 564,149 10	−7 558,411 51
7 560	−219	− 662,7	−1 935,7	14 447,7	7 810,528	7 594,100 6	7 564,479 93	7 558,293 47
−2 772	174	272,6	886,1	− 5 469,26	−2 869,055	−2 785,313 9	−2 773,735 48	− 2 771,341 69
Max. erreur relative sur une compos?	?	?	?	1	$\# \frac{1}{30}$	$\frac{3}{750} \# \frac{4}{1000}$	$\frac{4}{7\,560} \# \frac{5}{10\,000}$	$\frac{2}{7\,560} \# \frac{2}{10\,000}$

Tableau des résultats des différentes résolutions au sujet du $\frac{2}{2}$ 8

Ce qui revient à dire que pour résoudre le système avec quelque chance d'avoir des résultats avec un chiffre décimal exact, il faut atteindre les mantisses de capacité 8 au moins. Latable ci-dessus indique des résultats moins pessimistes. On constate, sur cette table, que c'est à partir de 7 chiffres significatifs (en décimal) de mantisse que l'on a une erreur relative, sur la solution, de l'ordre de 1/10. En fait, le tableau montre que l'on a $P \neq N-6$, au lieu de la formule précédente. Cela prouve que notre formule n'est pas trop "lâche". Il faut signaler, au point de vue de la réalisation de ces expériences, que c'est encore le principe de simulation d'opérations qui nous a servi pour l'exécution en machine de ces calculs. On peut noter le gain assez faible qu'apporte l'utilisation d'une virgule flottante à 9 chiffres de mantisse et avec arrondis sur l'utilisation sans arrondi, mais avec toujours 9 chiffres de mantisse.

9°) Evolution des conditionnements de systèmes successifs d'une élimination.

Dans ce dernier paragraphe, nous nous proposons d'étudier ce que l'on peut dire des conditionnements des matrices obtenues comme premiers membres des systèmes provenant d'une élimination.

Soit à résoudre le système $Ax = y$

Eliminons la première inconnue x_1 ; on peut considérer le système $A' \cdot x' = y'$ où la matrice est la partie carrée

de la matrice $A^{(2)}$, c'est à dire que:

$$a'_{ij} = a_{ij} - \frac{a_{i1} \cdot a_{1j}}{a_{11}} \qquad (i,j = 2, \ldots n)$$

et l'ordre en étant n-I,

$$x' = \begin{vmatrix} x_2 \\ \vdots \\ x_n \end{vmatrix} \quad , \quad y' = \begin{vmatrix} y_2^{(2)} \\ y_3^{(2)} \\ \vdots \\ y_n^{(2)} \end{vmatrix}$$

On va déterminer les nombres $m_{A'}, M_{A'}$ pour A' d'ordre n-I

Posons :

$$\Phi_A(x_1, x_2, \dots, x_n) = \sum_{i=1}^{i=n} |f_i(x_1, \dots x_n)|$$

si

$$f_i(x_1, \dots, x_n) = \sum_{j=1}^{j=n} a_{ij} x_j$$

puis

$$\Phi_{A'}(x_2, \dots x_n) = \sum_{i=1}^{i=n} \left| f_i - \frac{a_{i1}}{a_{11}} f_1 \right| \qquad (1)$$

On a donc;

$$\Phi_{A'}(x_2, \dots x_n) \equiv \Phi_A\left(-\frac{1}{a_{11}}(a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n), x_2, x_3, \dots, x_n\right)$$

et par suite

$$\frac{\Phi_{A'}(x_2 \dots x_n)}{\sqrt{x_2^2 + \dots + x_n^2}} \geqslant \frac{\Phi_A\left(-\frac{1}{a_{11}}(\dots), x_2, x_3 \dots, x_n\right)}{\sqrt{\frac{1}{a_{11}^2}(\dots)^2 + x_2^2 + \dots + x_n^2}} \geqslant m_A$$

D'où :

$$m_A \leqslant m_{A'} \qquad (I)$$

D'autre part, d'après (1)

$$\Phi_{A'}(x_2, \dots x_n) \leqslant \Phi_A(x_1, \dots x_n) + |f_1| \cdot \left[\frac{|a_{21}| + \dots + |a_{n1}|}{|a_{11}|} - 1\right]$$

donc:

$$\frac{\Phi_{A'}(x_2, \dots x_n)}{\sqrt{x_2^2 + \dots + x_n^2}} \leqslant \frac{\Phi_A(x_1, \dots x_n)}{\sqrt{x_1^2 + \dots + x_n^2}} + \frac{|f_1|}{\sqrt{x_2^2 + \dots x_n^2}} \cdot \left[\frac{|a_{21}| + \dots + |a_{n1}|}{|a_{11}|} - 1\right]$$

et si l'on fait $x_1 = 0$ dans cette inégalité valable quels que soient les $x_1, x_2, \dots, x_n$,

$$\frac{\Phi_{A'}(x_2, \dots x_n)}{\sqrt{x_2^2 + \dots + x_n^2}} \leqslant \frac{\Phi_A(0, x_2, \dots x_n)}{\sqrt{0 + x_2^2 + \dots + x_n^2}} + \frac{|a_{12}x_2 + \dots + a_{1n}x_n|}{\sqrt{x_2^2 + \dots + x_n^2}} \left[\frac{|a_{21}| + \dots + |a_{n1}|}{|a_{11}|} - 1\right]$$

I) si $\left[\frac{|a_{21}| + \dots + |a_{n1}|}{|a_{11}|} - 1\right] < 0$, il est clair qu'on peut choisir les $x_2, \dots, x_n$ de sorte que $|a_{12}x_2 + \dots + a_{1n}x_n| = 0$ et par conséquent:

$$\frac{\Phi_{A'}}{\|x'\|} \leqslant \frac{\Phi_A(0, x_2, \dots x_n)}{\sqrt{0 + x_2^2 + \dots + x_n^2}} \leqslant M_A$$

d'où:

$$M_{A'} \leqslant M_A \qquad (II)$$

2) Si $\left[\frac{|a_{21}|+\cdots+|a_{n1}|}{|a_{11}|}-1\right]\geqslant 0$, puisque (Inégalité de SCHWARZ)

$$\frac{|a_{12}x_2+\cdots+a_{1n}x_n|}{\sqrt{x_2^2+\cdots+x_n^2}}\leqslant\sqrt{a_{12}^2+\cdots+a_{1n}^2}$$

on a donc ,

$$\frac{\Phi_{A'}}{\|x'\|}\leqslant\frac{\Phi_A(0,x_2,\ldots x_n)}{\sqrt{x_2^2+\cdots x_n^2}}+\sqrt{a_{12}^2+\cdots+a_{1n}^2}\left[\frac{|a_{21}|+\cdots+|a_{n1}|}{|a_{11}|}-1\right]$$

or si dans $\frac{\Phi_A(x)}{\|x\|}$ on fait $x_1=0$, $x_i=a_{1i}$ $(i=2,\ldots n)$

cela donne: $\sqrt{a_{12}^2+\cdots+a_{1n}^2}\leqslant M_A$

et par suite: $\frac{\Phi_{A'}}{\|x'\|}\leqslant M_A+M_A\left[\frac{|a_{21}|+\cdots+|a_{n1}|}{|a_{11}|}-1\right]$

d'où $$M_{A'}\leqslant M_A\cdot\frac{|a_{21}|+\cdots+|a_{n1}|}{|a_{11}|}\qquad(\text{II})'$$

Mais les deux cas peuvent se résumer dans cette même formule:

$$(\text{III})\qquad M_{A'}\leqslant M_A\cdot K\qquad,\ \frac{1}{K}\frac{m_A}{M_A}\leqslant\frac{m_{A'}}{M_{A'}}$$

où $K=\text{Max}\left\{1,\frac{|a_{21}|+\cdots+|a_{n1}|}{|a_{11}|}\right\}$

D'où le théorème:

Théorème XII: Pour être sûr que l'élimination de x_1 ,conduise à un système de conditionnement général $\frac{m_{A'}}{M_{A'}}$ supérieur à celui $\frac{m_A}{M_A}$ du système initial, il suffit que l'on ait:

$$|a_{11}|\geqslant|a_{21}|+|a_{31}|+\cdots+|a_{n1}|$$

CHAPITRE VI

Les erreurs dans la résolution des systèmes linéaires par orthogonalisation.

1° Solution d'un système linéaire par orthogonalisation en ligne :

Soit $A.x = y$ (1) : un système linéaire

$A = (a_{ij})\ (i,j = 1, \dots n)$ matrice non singulière, et à éléments réels.

On désigne toujours par a_k (un seul indice), la k° ligne de A,

$a_k = A_{k.} = (a_{k1}, a_{k2}, \dots, a_{kn})$

$y_i\ ,(i = 1, \dots n)$ sont les composantes du vecteur du 2° membre.

La méthode consiste à remplacer le système donné par

$$H.x = Y \qquad (2)$$

avec $H = A^{(n)} = S_{n-1}.S_{n-2} \dots S_1.A$; $Y = S_{n-1}.S_{n-2} \dots S_1.y$

H orthogonale en lignes. (Cf. Chapitre I)

La solution du système $H.x = Y$ est alors très facile, en effet l'orthogonalité en lignes de H se traduit, sous forme matricielle, par:

$$H.H^T = D$$

où D est une matrice diagonale, dont les éléments diagonaux sont les carrés des longueurs des lignes de H.

Si l'on pose

$$x = H^T.z$$

le système (1) est remplacé par:

$$\begin{cases} HH^T z = Y \\ x = H^T z \end{cases}$$

puis :

$$\begin{cases} z = D^{-1} Y \\ x = H^T z \end{cases}$$

Enfin:

$$(3) \qquad x = H^T.D^{-1} Y$$

Pour écrire les formules de calcul nous poserons:

$$H = (h_{ij}) \quad , \quad (i,j = 1,2 \ldots\ldots n)$$

puis, $h_i = (h_{i1}, h_{i2}, \ldots\ldots h_{in})$

et

$$Y = (y_i) \quad , \quad (i = 1, \ldots\ldots n)$$

alors on a :

$$(I)\quad \begin{cases} h_1 = a_1 \\ h_2 = a_2 + \lambda_{21} \cdot h_1 \\ h_3 = a_3 + \lambda_{31} \cdot h_1 + \lambda_{32} \cdot h_2 \\ \vdots \\ h_n = a_n + \lambda_{n1} \cdot h_1 + \ldots\ldots + \lambda_{n,n-1} \cdot h_{n-1} \end{cases}$$

$$(II)\quad \begin{cases} Y_1 = y_1 \\ Y_2 = y_2 + \lambda_{21} Y_1 \\ Y_3 = y_3 + \lambda_{31} Y_1 + \lambda_{32} Y_2 \\ \vdots \\ Y_n = y_n + \lambda_{n1} Y_1 + \ldots\ldots + \lambda_{n,n-1} Y_{n-1} \end{cases}$$

$$(III)\quad \lambda_{ij} = - \frac{a_i \cdot h_j^T}{h_j \cdot h_j^T} \qquad \begin{matrix} (i>j \; ; \; j = 1, \ldots\ldots n-1) \\ (\qquad\quad i = 2, \ldots\ldots n) \end{matrix}$$

D'après ces formules on voit que si l'on pose

$$\Lambda = \begin{vmatrix} 0 & 0 & 0 & \ldots\ldots & 0 \\ \lambda_{21} & 0 & \ldots\ldots & & 0 \\ \lambda_{31} & \lambda_{32} & 0 & & \vdots \\ \vdots & & \ddots & & \\ \lambda_{n1} & \lambda_{n2} & \ldots\ldots & \lambda_{n,n-1} & 0 \end{vmatrix}$$

Les formules ci-dessus permettent d'écrire:

$$(I)' \quad H = A + \Lambda \cdot H$$

$$(II)' \quad Y = y + \Lambda Y$$

et si $T = (I - \Lambda)^{-1}$, triangulaire inférieure et unitaire,

$$(I)'' \quad H = T \cdot A$$

$$(II)'' \quad Y = T \cdot y$$

La matrice T n'est autre que le produit :

$$T = S_{n-1} \cdot S_{n-2} \ldots\ldots S_2 \cdot S_1 \cdot A$$

Comme nous l'avons déja fait nous allons étudier les erreurs de calcul pour cette méthode en séparant le cas où le calcul est conduit en virgule fixe de celui conduit en virgule flottante.

2° Etude des erreurs en calcul en virgule fixe .

Nous désignerons par ε_{ij} l'erreur de calcul d'un λ_{ij} :

$$\bar{\lambda}_{ij} \pm = -\frac{a_i \cdot \bar{h}_j^T}{\bar{h}_j \cdot \bar{h}_j^T} + \varepsilon_{ij}$$

et par suite la détermination de H et Y conduit à $\bar{H}$, $\bar{Y}$ tels

$$(I)_e \begin{cases} \bar{h}_1 = a_1 \\ \bar{h}_2 = a_2 + \bar{\lambda}_{21} \odot \bar{h}_1 \\ \bar{h}_3 = a_3 + \bar{\lambda}_{31} \odot \bar{h}_1 + \bar{\lambda}_{32} \odot \bar{h}_2 \\ \vdots \\ \bar{h}_n = a_n + \bar{\lambda}_{n1} \odot \bar{h}_1 + \cdots\cdots + \bar{\lambda}_{n\,n-1} \odot \bar{h}_{n-1} \end{cases} \quad (II)_e \begin{cases} \bar{Y}_1 = y_1 \\ \bar{Y}_2 = y_2 + \bar{\lambda}_{21} \odot \bar{Y}_1 \\ \vdots \\ \bar{Y}_n = \bar{y}_n + \bar{\lambda}_{n1} \odot \bar{Y}_1 + \cdots + \bar{\lambda}_{n\,n-1} \odot \bar{Y}_{n-1} \end{cases}$$

si nous supposons (calcul en fixe) que nous ne faisons pas d'erreur dans les additions algébriques .

Posons $\quad \bar{\lambda}_{ij} \odot \bar{h}_j = \bar{\lambda}_{ij} \cdot \bar{h}_j + \eta_{ij}$

$$\bar{\lambda}_{ij} \odot \bar{Y}_j = \bar{\lambda}_{ij} \cdot \bar{Y}_j + \rho_{ij}$$

où η_{ij} désigne une ligne dont les valeurs sont les erreurs de multiplication dans l'évaluation du produit:

$$\bar{\lambda}_{ij} \cdot \bar{h}_j$$

Les formules de calcul réel s'écrivent donc:

$$(I)'_e \begin{cases} \bar{h}_1 = a_1 \\ \bar{h}_2 = a_2 + (\lambda_{21} + \varepsilon_{21}) \cdot \bar{h}_1 + \eta_{21} \\ \bar{h}_3 = a_3 + (\lambda_{31} + \varepsilon_{31}) \cdot \bar{h}_1 + (\lambda_{32} + \varepsilon_{32}) \cdot \bar{h}_2 + \eta_{31} + \eta_{32} \\ \vdots \\ \bar{h}_n = a_n + (\lambda_{n1} + \varepsilon_{n1}) \cdot \bar{h}_1 + (\lambda_{n2} + \varepsilon_{n2}) \cdot \bar{h}_2 + \cdots\cdots + (\lambda_{n\,n-1} + \varepsilon_{n\,n-1}) \cdot \bar{h}_{n-1} + \eta_{n1} + \eta_{n2} \cdots\cdots + \eta_{n\,n-1} \end{cases}$$

$$(II)'_e \begin{cases} \bar{Y}_1 = y_1 \\ \bar{Y}_2 = y_2 + (\lambda_{21} + \varepsilon_{21}) \cdot \bar{Y}_1 + \rho_{21} \\ \vdots \\ \bar{Y}_n = y_n + (\lambda_{n1} + \varepsilon_{n1}) \cdot \bar{Y}_1 + (\lambda_{n2} + \varepsilon_{n2}) \cdot \bar{Y}_2 + \cdots\cdots + (\lambda_{n\,n-1} + \varepsilon_{n\,n-1}) \cdot \bar{Y}_{n-1} + \rho_{n1} + \rho_{n2} + \cdots\cdots + \rho_{n\,n-1} \end{cases}$$

Par suite les $\bar{H}$ et $\bar{Y}$ auquels on arrive satisfont aux équations:

$$(I)''_e \qquad \bar{H} = A + (\Lambda + E) \cdot \bar{H} + K$$

$$(II)''_e \qquad \bar{Y} = y + (\Lambda + E) \cdot \bar{Y} + \rho$$

Si l'on pose :

$$E = \begin{vmatrix} 0 & 0 & 0 & \cdots & 0 \\ \varepsilon_{21} & 0 & 0 & \cdots & 0 \\ \varepsilon_{31} & \varepsilon_{32} & 0 & \cdots & 0 \\ \vdots & & & & \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{n,n-1} & 0 \end{vmatrix}, \quad K = \begin{vmatrix} 0 \\ \eta_{21} \\ \eta_{31} + \eta_{32} \\ \vdots \\ \eta_{n1} + \eta_{n2} + \cdots + \eta_{n,n-1} \end{vmatrix}, \quad \mathcal{E} = \begin{vmatrix} 0 \\ \rho_{21} \\ \rho_{31} + \rho_{32} \\ \vdots \\ \rho_{n1} + \rho_{n2} \cdots + \rho_{n,n-1} \end{vmatrix}$$

Pour la matrice K il faut noter que chaque symbole représente une ligne.

On peut remarquer qu'en ordre de grandeur, ces matrices peuvent être écrites

$$E \# e' \begin{vmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & & \vdots \\ 1 & 1 & 1 & 0 & \\ \vdots & & & & \\ 1 & 1 & \cdots & 1 & 0 \end{vmatrix}, \quad K \# e \begin{vmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & 1 & \cdots & 1 \\ 2 & 2 & 2 & \cdots & & 2 \\ \vdots & & & & & \\ n-1 & n-1 & \cdots & & & n-1 \end{vmatrix}, \quad \mathcal{E} \# e \begin{vmatrix} 0 \\ 1 \\ 2 \\ \vdots \\ n-1 \end{vmatrix}$$

Si comme plus haut, e' et e désignent les erreurs en division et en multiplication.

Donc la solution de $A x = y$, rigoureusement la même que celle du système $H x = Y$, est remplacée par celle de : $\bar{H} x = \bar{Y}$

Posons encore $\bar{H} = H + \delta H$ et $\bar{Y} = Y + \delta Y$.

Il en résulte que :

$$\delta H = (\Lambda + E).(H + \delta H) - \Lambda . H + K$$

$$\delta Y = (\Lambda + E).(Y + \delta Y) - \Lambda . Y + \mathcal{E}$$

d'où en négligeant les termes de $E.\delta H$ et ceux de $E.\delta Y$,

$$\boxed{\begin{aligned} \delta H &= E.H + \Lambda.\delta H + K \\ \delta Y &= E.Y + \Lambda.\delta Y + \mathcal{E} \end{aligned}}$$

et puisque $T = (I - \Lambda)^{-1}$, on en déduit le thèorème:

Théorème I Si T est la matrice triangulaire inférieure unitaire qui dans la procédé d'orthogonalisation permet d'écrire $H = T.A$ avec H orthogonale en lignes, le calcul numérique, en virgule fixe, de cette matrice conduit à une matrice $\bar{H}$ et le système à résoudre $H.x = Y$ équivalent à $Ax = y$ est remplacé par $(H + \delta H)\, x = Y + \delta Y$ avec,

$$\boxed{\begin{aligned} \delta H &= T.(E.H + K) \\ \delta Y &= T.(E.Y + \mathcal{E}) \end{aligned}}$$

Il est bien clair que la matrice $\bar{H}$ n'est pas orthogonale en ligne, et cependant nous allons traiter le système comme si il était à matrice de premier membre orthogonale en lignes . Cela conduit aux opérations

I°) Déterminer les éléments diagonaux de $(H+\delta H)\cdot(H+\delta H)^T = (H+\delta H)(H^T+(\delta H)^T)$

Je pose $\bar{D}$ pour la matrice obtenue en ne prenant que les éléments diagonaux de ce produit, comme on peut écrire :

$$Z = H.H^T + \delta H.H^T + H.(\delta H)^T$$

(négligeant les éléments de $\delta H.\delta H$)

il en résulte quesi l'on pose encore : $\bar{D} = D + \delta D$

$$\delta D = diag\left[\delta H.H^T + H.(\delta H)^T\right]$$

2) Cela fait il reste pour déterminer la solution la formule:

$$x^* = \bar{H}^T \odot \left(\bar{D}^{-1} \odot \bar{Y}\right)$$

- a) Je vais poser $\bar{D}^{-1}\odot\bar{Y} = \bar{D}^{-1}.\bar{Y} + Q$, avec

$$Q = \begin{vmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{vmatrix} \quad , \quad Q \# e'. \begin{vmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{vmatrix}$$

Si q_i désignent les erreurs de division dans le calcul de $\bar{D}^{-1}.\bar{Y}$

En négligeant les termes du 2° ordre par rapport à δD ,

$$\bar{D}^{-1} = (D+\delta D)^{-1} = (I + D^{-1}\delta D)^{-1}.D^{-1} = (I - D^{-1}\delta D).D^{-1}$$

$$= D^{-1} - D^{-1}.\delta D.D^{-1}$$

d'où :

$$\bar{D}^{-1}\odot\bar{Y} = (D^{-1} - D^{-1}.\delta D.D^{-1})(Y+\delta Y) + Q$$

enfin

$$\bar{D}^{-1}\odot\bar{Y} = D^{-1}.Y + D^{-1}\delta Y - D^{-1}.\delta D.D^{-1}Y + Q$$

- b) Il reste à calculer le produit $\bar{H}^T\odot(\bar{D}^{-1}\odot\bar{Y})$

Là encore on posera $\bar{H}^T\odot(\bar{D}^{-1}\odot\bar{Y}) = \bar{H}^T.(\bar{D}^{-1}\odot\bar{Y}) + R$

la matrice étant d'ordre n, le vecteur d'erreur R aura pour composantes les erreurs sur les composantes du produit et:

$$R = \begin{vmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{vmatrix} \quad , \quad R \# e. \begin{vmatrix} n \\ n \\ \vdots \\ n \end{vmatrix}$$

Il en résulte que :

$$x^* = x + \delta x = \left(H^T + (\delta H)^T\right) \cdot \left(D^{-1} Y + D^{-1} \delta Y - D^{-1} \delta D \cdot D^{-1} Y + Q\right) + R$$

Donc:

Théorème II La méthode de résolution d'un système d'équations linéaires par le procédé d'orthogonalisation en lignes, conduit à une solution $x^* = x + \delta x$ avec

$$(I) \quad \boxed{\delta x = (\delta H)^T D^{-1} Y - H^T D^{-1} \delta D \cdot D^{-1} Y + H^T D^{-1} \delta Y + H^T Q + R}$$

les matrices et les vecteurs $\delta H, \delta D, \delta Y, Q, R$ ayant été précédemment définis, les calculs étant conduits en virgule fixe.

Cette formule permet d'écrire:

$$\delta x = \delta\left(H^T \cdot D^{-1} Y\right) + H^T Q + R$$

Or:

$$H^T D^{-1} = H^{-1}$$

donc:

$$\delta x = H^{-1} \delta Y - H^{-1} \cdot \delta H \cdot H^{-1} Y + H^T Q + R$$

Mais si l'on revient aux expréssions indiquées plus haut pour δH et δY cela se réduit à la formule très simple:

$$\delta x = H^{-1}\left[T(EY + f) - T(EH + K)H^{-1} Y\right] + H^T Q + R$$

$$(II) \quad \boxed{\delta x = A^{-1} \cdot \left[f - K x^*\right] + H^T Q + R}$$

x^* désigne la solution obtenue.

Comme utilisation de cette dernière formule nous l'écrirons en ordre de grandeur avec les expressions données pour les vecteurs et matrices K, f, Q, R .

$$\delta x = e \cdot A^{-1} \cdot \begin{bmatrix} 0 \\ (1 - \xi_1 - \xi_2 \cdots - \xi_n) \\ 2(1 - \xi_1 - \xi_2 \cdots - \xi_n) \\ \\ (k-1)(1 - \xi_1 \cdots - \xi_n) \end{bmatrix} + e' \cdot H^T \cdot \begin{vmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{vmatrix} + ne \begin{vmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{vmatrix}$$

Si $\xi_1, \xi_2, \ldots \xi_n$, sont les composantes de la solution: x^*

Le fait le plus remarquable à signaler est que ,à cet ordre d'appoximation et en calcul en fixe, la matrice E ne figure plus dans l'expression de l'erreur,cela d'après la formule II .

Comme application posons-nous la question suivante:
Etant donné un système linéaire, $A x = y$, résolvons-le une premiére fois en fixe,par orthogonalisation,cela conduit à une solution x_1^* ,puis échangeons les positions de deux équations du système cela donne une autre solution etc.. .Il est clair que je peux ainsi résoudre $n!$ systèmes linéaires qui ne diffèrent entre-eux qu'à une permutation près des équations,donc qui ont parfaitement le même vecteur solution x.

Or l'expérience prouve que la valeur du vecteur solution-approchée x^* n'est pas la même et cela à cause des erreurs de calcul . Quelle est donc la solution la meilleure ?

D'après la dernière formule écrite on voit que la norme (φ_2 , par exemple) du premier vecteur donnant δx ne va pas beaucoup varier pour n peu élevé, par contre pour le second $H^T . Q$,les choses sont différentes . il est clair que l'on aura toujours le plus grand intéret à choisir la résolution qui conduit à affaiblir le plus possible une norme de H^T .

Or

$$H . H^T = D$$

Puis comme : $H = T . A$, $Det(H) = \Delta = Det(A)$, $Det(D) = \Delta^2$

On en déduit que : $N^2(H^T) = Trace(H . H^T) = Trace(D)$

Si l'on pose $\ell_i^2 = h_i . h_i^T$, $\ell_i > 0$

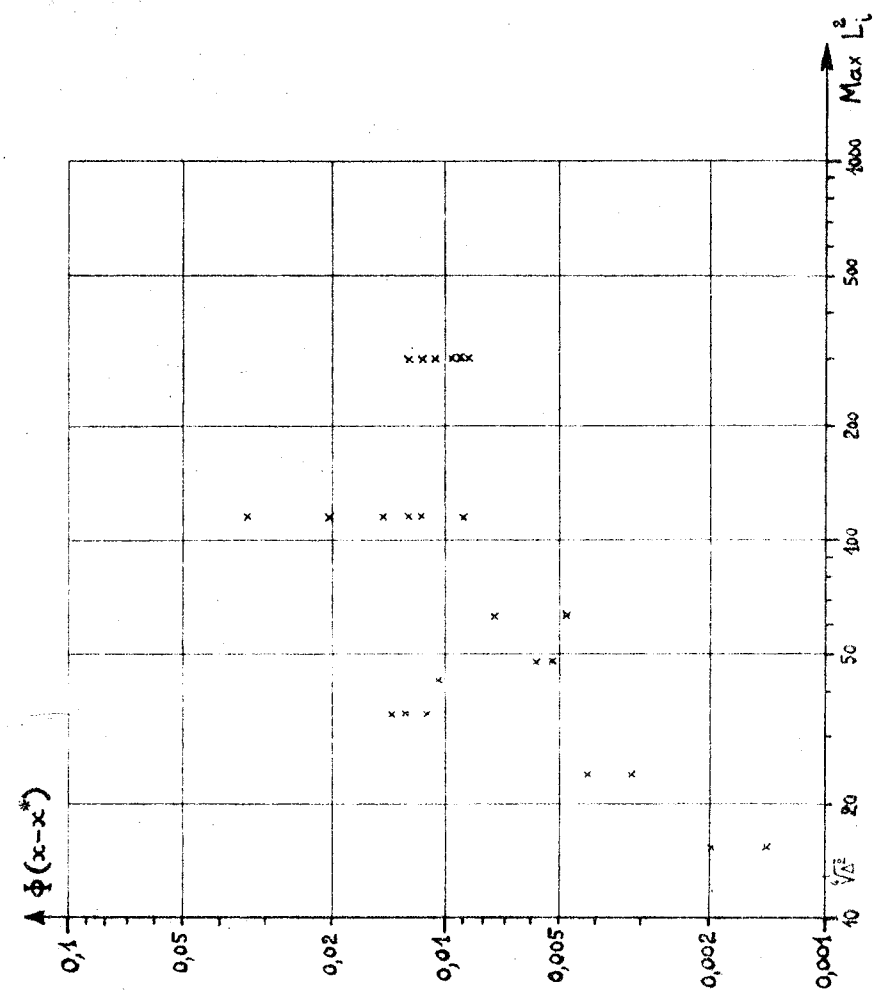
$$N^2(H^T) = N^2(H) = \ell_1^2 + \ell_2^2 + \dots + \ell_n^2$$

Donc puisque le produit $\ell_1^2 . \ell_2^2 \dots \ell_n^2 = Det(D) = \Delta^2$ est constant dans toute nos résolutions , $N(H^T)$ sera minimum si les longueurs des lignes de H sont toutes presque égales entre-elles.

D'où :

Règle Afin de minimiser l'erreur de calcul en virgule fixe dans la solution d'un système linéaire par la méthode d'orthogonalisation ,il suffit d'organiser le système de sorte que les différentes longueurs des lignes de H soient les plus près possible de l'égalité entre-elles et par suite près de : $\sqrt[n]{\Delta}$

Afin de vérifier cette règle nous avons fait résoudre au calculateur



Resolutions d'un même système linéaire
par la méthode d'orthonormalisation en lignes.

électronique les 24 systèmes linéaires obtenus par toutes les permutations des équations àpartir du système déja utilisé, d'ordre 4 au chapitre V, § 6 qui nous a servi à verifier la règle du choix des pivots dans une résolution par élimination. Les résultats sont indiqués sur la figure ci-contre . On a porté en abscisse sur ce graphique $\underset{i=1,2,3,4}{\text{Max}}\,(\varepsilon_i^2)$ pour une résolution et en ordonnée, la valeur de la norme $\bar{\Phi}(x^*-x)$ de l'erreur correspondant à cette résolution.

3°) Etude des erreurs en calcul en virgule flottante .

De même qu'au § 3, nous sommes conduits à trouver une matrice $\bar{H}$ et un 2° membre $\bar{Y}$ définis par les relations (mêmes notations)

$$(I)'_f \quad \begin{cases} \bar{h}_1 = a_1 \\ \bar{h}_2 = a_2 + (\lambda_{21}+\varepsilon_{21})\cdot\bar{h}_1 + \xi_2\cdot\bar{h}_2 \\ \vdots \\ \bar{h}_n = a_n + (\lambda_{n1}+\varepsilon_{n1})\cdot\bar{h}_1 + (\lambda_{n2}+\varepsilon_{n2})\cdot\bar{h}_2 + \cdots\cdots + (\lambda_{nn-1}+\varepsilon_{nn-1})\cdot\bar{h}_{n-1} + \xi_n\cdot\bar{h}_n \end{cases}$$

$$(II)'_f \quad \begin{cases} \bar{Y}_1 = a_1 \\ \bar{Y}_2 = a_2 + (\lambda_{21}+\varepsilon_{21})\cdot\bar{Y}_1 + \pi_2\cdot\bar{Y}_2 \\ \vdots \\ \bar{Y}_n = a_n + (\lambda_{n1}+\varepsilon_{n1})\cdot\bar{Y}_1 + \cdots\cdots + (\lambda_{nn-1}+\varepsilon_{nn-1})\cdot\bar{Y}_{n-1} + \pi_n\cdot\bar{Y}_n \end{cases}$$

Cette fois ξ_i et π_i désignent les erreurs relatives faites sur les calculs des $\bar{h}_i$ et $\bar{Y}_i$.

On peut écrire ces formules sous forme matricielle :

$$(I)''_f \qquad \bar{H} = A + (\Lambda+E)\cdot\bar{H} + K_1\cdot\bar{H}$$

$$(II)''_f \qquad \bar{Y} = y + (\Lambda+E)\cdot\bar{Y} + P_1\cdot\bar{Y}$$

Où l'on a posé:

$$K_1 = \begin{vmatrix} 0 & & & \\ \xi_2 & 0 & & \\ & \xi_3 & \ddots & \\ & & \ddots & 0 \\ & & & \xi_n \end{vmatrix} \quad , \quad P_1 = \begin{vmatrix} 0 & & & \\ \pi_2 & 0 & & \\ & \pi_3 & \ddots & \\ & & \ddots & 0 \\ & & & \pi_n \end{vmatrix}$$

Il est à remarquer que les formules $(I)'_f$ ne sont qu'approchées car le fait d'écrire que l'erreur absolue dans le calcul de la ligne $\bar{h}_i$ est proportionnelle à cette ligne, elle-même, est assez schématique. Malgré cela, nous nous contenterons de cette hypothèse simplificatrice et nous poursuivrons dans ce cas l'étude des erreurs. Il est à noter que dans l'étude faite en virgule fixe les formules $(I)'_e$ sont exactes.

On notera aussi les valeurs en ordre de grandeur des matrices:

$$K_1 = \varepsilon \begin{vmatrix} 0 & & & \\ 1 & 1 & & 0 \\ & & \ddots & \\ 0 & & & 1 \end{vmatrix} \quad , \quad f_1 = \varepsilon \begin{vmatrix} 0 & & & \\ 1 & 1 & & 0 \\ & & \ddots & \\ 0 & & & 1 \end{vmatrix}$$

Cela posé, si $\bar{H} = H + \delta H$ et $\bar{Y} = Y + \delta Y$

On trouve de la même façon qu'au § 3, δH, δY:

D'où:

Théorème III Si T est la matrice triangulaire inférieure et unitaire qui dans le procédé d'orthogonalisation permet d'écrire $H = T.A$, avec H orthogonale en lignes, le calcul numérique en virgule flottante, de cette matrice conduit à une matrice $\bar{H}$ et le système à résoudre $Ax = y$ qui a la même solution que $Hx = Y$ est remplacé par $(H+\delta H)\cdot x = Y + \delta Y$, avec:

$$\delta H = T.(E + K_1).H$$

$$\delta Y = T.(\epsilon + f_1).Y$$

Si l'on continue à suivre lemême raisonnement qu'au §3, nous voyons que : $x^* = \bar{H}^T \boxdot \left(\bar{D}^{-1} \boxdot \bar{Y}\right)$

- I°) On peut encore écrire

$$\bar{D}^{-1} \boxdot \bar{Y} = \bar{D}^{-1}.\bar{Y} + Q_1.(\bar{D}^{-1}.\bar{Y})$$

où

$$Q_1 = \begin{vmatrix} q'_1 & & & \\ & q'_2 & & 0 \\ & & \ddots & \\ 0 & & & q'_k \end{vmatrix} \quad \# \quad \varepsilon . I$$

- 2°) Puis pour calcule le produit $\bar{H}^T \boxdot \left(\bar{D}^{-1} \boxdot \bar{Y}\right) = x^*$

On aura:

$$\bar{H}^T \boxdot \left(\bar{D}^{-1} \boxdot \bar{Y}\right) = \bar{H}^T.\left(\bar{D}^{-1} \boxdot \bar{Y}\right) + R_1.\left[\bar{H}^T.(\bar{D}^{-1} \boxdot \bar{Y})\right]$$

Avec :

$$R_1 = \begin{vmatrix} r'_1 & & & \\ & r'_2 & & \\ & & \ddots & \\ & & & r'_n \end{vmatrix} \# \varepsilon . I \quad , \quad R_1 . \left[\bar{H}^T . (\bar{D}^{-1} \boxdot \bar{Y}) \right] = R_1 . x^*$$

Enfin, après le même calcul que celui déjà fait au paragraphe 3, on trouve :

$$\delta x = \delta \left(H^T . D^{-1} Y \right) + H^T . Q_1 . D^{-1} Y + R_1 . x^*$$

D'où :

Théorème IV : La méthode de résolution d'un système d'équations linéaires par le procédé d'orthogonalisation en lignes, en virgule flottante, conduit à une solution

$x^* = x + \delta x$ avec : $\boxed{\delta x = A^{-1} \left[P_1 . H - K_1 H \right] x^* + H^T Q_1 D^{-1} H x^* + R_1 . x^*}$

les matrices P_1, K_1, Q_1, R_1 , ayant été précédemment définies.

En ordre de grandeur, on peut écrire :

et, par suite : $\delta x \# 2\varepsilon A^{-1} . H x^* + 2\varepsilon x^*$

$$\delta x \# 2\varepsilon (T + I) x^* = 2\varepsilon T_1 x^* \qquad (1) \quad ,$$

$T_1 = T + I$, triangulaire .

Ces formules sont, certes, assez grossières, mais leur extrême simplicité peut justifier leur utilisation.

De la formule (1), on tire : $\dfrac{\Phi(\delta x)}{\| x^* \|} \leq 2\varepsilon . M_{T_1}$

$M_{T_1} = \max_E \left[T_1^T E \right]$, facile à calculer, et, par suite, si l'on veut obtenir une précision relative de l'ordre de β^{-p}, 2ε étant de l'ordre de β^{-k}, il suffit d'avoir : $\beta^{-k} . M_{T_1} \leq \beta^{-p}$

ou $k \geq p + \log_\beta (M_{T_1})$ d'où le résultat :

Théorème V : Pour obtenir pour la solution d'un système linéaire par la méthode d'orthogonalisation, en point décimal flottant, au moins p chiffres significatifs exacts aux nombres-solution, il suffit de conduire les calculs en point décimal flottant de capacité k telle :

$$\boxed{k \geq p + \log_\beta (M_{T_1})}$$

CHAPITRE VII

Les normes générales et les procédés itératifs de résolution des systèmes linéaires .

I°) Notations

Dans ce qui suit ,pour débuter, nous ne considérons que l'espace affine R^n . La solution d'un système linéaire : (1) $A.x = y$ sera désignée par ω ,pour la colonne,et le point de l'espace ayant ces coordonnées sera désigné par Ω .

Ce point Ω est le point commun aux hyper-plans (ou plans) dont les équations sont:

$$\begin{cases} f_1(x) \equiv A_1.x - y_1 = 0 \\ f_2(x) \equiv A_2.x - y_2 = 0 \\ \vdots \\ f_n(x) \equiv A_n.x - y_n = 0 \end{cases}$$

Si x_0 est un "point" de R^n le vecteur :

$$r(x_0) = A.x_0 - y$$

est dit "vecteur résidu" en x_0 .

Il est évident que $r(x_0) = 0$, si,et seulement si $x_0 = \omega$

Si non, peut-on obtenir facilement un vecteur (ou un point) x_1 tel que:

$$\varphi[r(x_1)] < \varphi[r(x_0)]$$

Si φ est une certaine norme sur R^n ? C'est la question fondamentale posée dans toutes les méthodes de résolution itératives .

Si $\varphi(x)$ est une certaine norme je poserai:

$$\varphi(x) = F(x_1, x_2, \ldots\ldots, x_n)$$

si x est la colonne : $x = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{vmatrix}$.

Soit x_0 un point ou $F(x_1^0, \ldots\ldots x_n^0) = \varphi(x_0)$ est différentiable on désignera par $p(x_0)$, le vecteur-colonne:

$$p(x) = \begin{vmatrix} \left(\frac{\partial F}{\partial x_1}\right)_0 \\ \left(\frac{\partial F}{\partial x_2}\right)_0 \\ \vdots \\ \left(\frac{\partial F}{\partial x_n}\right)_0 \end{vmatrix} = \text{grad}\,(\varphi(x))$$

Soit x_0 un point de R^n, et considérons le lieu $\mathcal{B}^\circ_\omega(A)$ des points x de R^n tels:

$$\varphi(Ax - y) \leq \varphi(Ax_0 - y)$$

puisque $A.\omega = y$,il est clair que c'est aussi la lieu des points tels: $\varphi(A(x-\omega)) \leq \varphi(A(x_0-\omega))$

si je pose $x' = A(x - x_0)$ je vois que ces lieux ne sont autres que les transformés des boules:

$$\varphi(x') \leq \varphi(x'_0)$$

de centre l'origine du système d'axes de R^n , dans la transformation définie par:

$$x' = A(x-\omega) \quad , \quad x = \omega + A^{-1}x'$$

c'est à dire le produit d'une translation qui les centre en ω et d'une affinité définie par A .

$\Pi^\circ_\omega(A)$ Désignera la frontière de ce lieu : $\mathcal{B}^\circ_\omega(A)$.

2°) Rapprochement de la solution .

En calcul numérique les seules normes utilisées sont les trois normes $\varphi_1(x)$, $\varphi_2(x)$, $\varphi_\infty(x)$ déjà étudiées , et le but de la méthode itérative cherchée, doit être de pouvoir obtenir x^* tel que $\varphi_1(x^*-\omega) \leq \varepsilon$,ou $\varphi_2(x^*-\omega) \leq \varepsilon$,ou $\varphi_\infty(x^*-\omega) \leq \varepsilon$, ε étant donné.

Or nous disposons pour définir ω des plans donnés :

$$f_1(x) = 0 \quad , \quad f_2(x) = 0, \quad \cdots\cdots \quad , \quad f_n(x) = 0$$

et aussi d'autant de plans que l'on veut:il suffit de remplacer une des équations ci-dessus par une combinaison linéaire convenable de celles-ci pour obtenir un système ayant la même solution . Si ζ est un vecteur colonne quelconque,

$$(1) \quad \zeta^T.(Ax - y) = 0$$

est aussi l'équation d'un plan passant par ω .

Or de nos trois normes, une seule , la norme euclidienne $\varphi_2(x)$ a cette propriété qu'il est immédiat de trouver dans un plan arbitraire passant par ω un point x_1 plus rapproché ,au sens de cette norme, de ω . Ce point n'est autre que la projection orthogonale de x_0 sur ce plan .

Soit x_0 donné, cherchons la projection orthogonale x_1 de ce point sur le plan :

$$z^T.(Ax-y)=0$$

Ce dernier est perpendiculaire au vecteur $A^T.z$ donc x_1, projection cherchée est tel qu'il existe un scalaire λ avec le quel on puisse écrire:

$$x_1 = x_0 + \lambda . A^T.z$$

Mais comme x_1 est dans le plan, c'est que λ satisfait à:

$$z^T.(Ax_0 + \lambda AA^Tz - y)=0$$

d'où :

$$\lambda = -\frac{z^T.(Ax_0-y)}{z^TAA^Tz} = -\frac{z^T.r(x_0)}{\|A^Tz\|^2}$$

et par suite :

$$(1)\qquad \boxed{x_1 = x_0 - \frac{z^T.r(x_0)}{\|A^Tz\|^2}.A^Tz}$$

3°) Décomposition d'une norme .

Définition : Nous dirons qu'une norme φ a été décomposée, si l'on a pu trouver une application $z : x \longrightarrow z(x)$ de R^n dans lui-même, telle que l'on ait quel que soit x :

$$\varphi(x) = z^T(x).x$$

Il est clair que:

- 1°) Pour $\Phi(x)=\varphi_1$ on peut écrire : $\Phi(x) = z^T(x).x$

avec :

$$z(x) = m(x) = \begin{vmatrix} sig(\xi_1) \\ sig(\xi_2) \\ \vdots \\ \vdots \\ sig(\xi_n) \end{vmatrix}$$

où la fonction $sig(\xi)$ est définie par:

$$sig(\xi) = +1 \quad , \xi \geq 0$$
$$sig(\xi) = -1 \quad , \xi < 0$$

- 2°) Pour $\mathcal{M}_0(x)=\varphi_\infty$ on peut écrire: $\mathcal{M}_0(x) = z^T(x).x$

avec :

$$z(x) = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ sig(\xi_i) \\ 0 \\ \vdots \end{vmatrix} = p(x) \quad , |\xi_i| = \max_{j=1,\dots n} |\xi_j|$$

3°) Pour $\|x\| = \varphi_2$, on peut évidemment écrire : $\|x\| = \frac{x^T}{\|x\|} \cdot x$

avec : $\zeta(x) = \frac{x}{\|x\|}$

Plus généralement si B est une matrice symétrique définie positive on sait (24) que $\sqrt{x^T B x} = \varphi(x)$ est une norme . alors si l'on remarque que:

$$(Bx)^T \cdot x = x^T B x$$

on voit que pour cette norme $\varphi(x) = \zeta^T(x) \cdot x$ avec $\zeta(x) = \frac{Bx}{\sqrt{x^T B x}}$

4°) Enfin pour une norme de HOLDER $\varphi_r(x) = \left(\sum_{i=1}^{i=n} |\xi_i|^r \right)^{\frac{1}{r}}$, $r \geq 1$

on peut écrire $\varphi_r(x) = \zeta^T(x) \cdot x$

avec

$$\zeta(x) = \frac{1}{\left(\sum_{i=1}^{i=n} |\xi_i|^r \right)^{1-\frac{1}{r}}} \cdot \begin{vmatrix} \varepsilon_{ij}(\xi_1) \cdot |\xi_1|^{r-1} \\ \varepsilon_{ij}(\xi_2) \cdot |\xi_2|^{r-1} \\ \vdots \\ \varepsilon_{ij}(\xi_n) \cdot |\xi_n|^{r-1} \end{vmatrix}$$

Conclusions: Toutes les normes de l'analyse numérique sont très facilement décomposables comme on vient de le voir .

Une remarque importante pour la suite est celle-ci :

Dans tous les cas précédents on a , pour tout x

$$\|\zeta(x)\| \leq K$$

avec $K = n$ pour le premier cas , dans les autres $(\mathfrak{M}, \|\|)$; $K = 1$

4°) Procédé itératif associé à une décomposition d'une norme .

Dès que l'application ζ , en général très simple, a été mise en évidence , nous avons un procédé qui permet de faire correspondre à un x_0 un plan passant par ω : celui qui correspond à la valeur $\zeta(r(x_0))$ de ζ .

Partant de x_0 arbitraire , on prend donc :

$$x_1 = x_0 - \frac{\varphi(r(x_0))}{\|A^T \cdot \zeta(r(x_0))\|^2} \cdot A^T \cdot \zeta(r(x_0)) .$$

cela, d'après la formule(1) du § 2 , et le choix de $\zeta(r(x))$.

Je pose pour simplifier l'écriture :

$$\nabla(x) = A^T . \zeta(r(x))$$

donc, alors :

$$x_1 = x_0 - \varphi(r(x_0)) . \frac{\nabla(x_0)}{\| \nabla(x_0) \|^2}$$

L'itération consiste à prendre

$$x_n = x_{n-1} - \varphi(r(x_{n-1})) . \frac{\nabla(x_{n-1})}{\| \nabla(x_{n-1}) \|^2}$$

Etudions la convergence du procédé.

Je remarque que l'on peut écrire ,à cause de l'orthogonalité :

$$\| \omega - x_n \|^2 = \| \omega - x_{n-1} \|^2 - \| x_n - x_{n-1} \|^2 = \| \omega - x_{n-1} \|^2 \left[1 - L(x_{n-1}) \right]$$

avec $$L(x_{n-1}) = \frac{\varphi^2(r(x_{n-1}))}{\| x_{n-1} - \omega \|^2} . \frac{1}{\| \nabla(x_{n-1}) \|^2}$$

Or $$\varphi(r(x_{n-1})) = \varphi(A(x_{n-1} - \omega))$$

ou $$L(x_{n-1}) = \frac{\varphi^2(A(x_{n-1} - \omega))}{\| x_{n-1} - \omega \|^2} . \frac{1}{\| \nabla(x_{n-1}) \|^2}$$

et si m_A désigne $\min_{X \neq 0} \frac{\varphi(AX)}{\| X \|}$, on peut dire que

$$L(x_{n-1}) \geqslant \frac{m_A^2}{\| \nabla(x_{n-1}) \|^2}$$

D'autre part, (25)

$$\| \nabla(x_{n-1}) \|^2 = \| A^T . \zeta(x_{n-1}) \|^2 \leqslant N^2(A^T) . \| \zeta(x_{n-1}) \|^2 \leqslant N^2(A) . K^2$$

en désignant toujours par $N(A) = \left(\sum_{ij} a_{ij}^2 \right)^{\frac{1}{2}}$ qui est une norme sur l'ensemble dés matrices carrées d'ordre n, et puisque sur cet emsemble toutes les normes de matrices sont équivalentes, il existe , si M_A désigne la norme $\max_{X \neq 0} \left(\frac{\varphi(AX)}{\| X \|} \right)$ (cf chap II) une constante R , indépendante de A , telle que

$$N(A) \leqslant R . M_A$$

Donc ,

$$\| \nabla(x_{n-1}) \|^2 \leqslant K^2 . R^2 . M_A^2 = R^2 . M_A^2$$

puis,

$$L(x_{n-1}) \geqslant \frac{1}{R^2} . \left(\frac{m_A}{M_A} \right)^2$$

enfin ,

$$\| \omega - x_n \| \leqslant \| \omega - x_{n-1} \| . \sqrt{1 - \frac{1}{R^2} . \left(\frac{m_A}{M_A} \right)^2}$$

et de là :

$$\| \omega - x_p \| \leqslant \| \omega - x_0 \| . \left(\sqrt{1 - \frac{1}{R^2} \left(\frac{m_A}{M_A} \right)^2} \right)^p$$

Et en introduisant le conditionnement général $\gamma(A)$ par rapport aux deux normes $\varphi(x)$ et $\|x\|$, cela peut s'écrire :

$$\|\omega - x_p\| \leq \|\omega - x_0\| \cdot \left(\sqrt{1 - \frac{\gamma^2}{R^2}}\right)^p$$

d'où le théorème :

Théorème I Etant donnée une norme φ sur R^n , le procédé itératif associé à une décomposition : $\varphi(x) = z^T(x) \cdot x$ de cette norme , avec $\|z^T(x)\|$ borné quelque soit x , est toujours convergent . Cette convergence est assurée par l'inégalité:

$$\|\omega - x_p\| \leq \|\omega - x_0\| \cdot \left(\sqrt{1 - \frac{\gamma^2}{R^2}}\right)^p$$

qui montre le rôle essentiel joué dans le procédé par le conditionnement général γ de A par rapport aux deux normes φ et $\|x\|$. [26]

De ce théorème on peut déduire autant de procédés itératifs que l'on veut , nous allons en étudier quelques-uns dans ce qui va suivre.

5°) Procédé associé à la norme $\varphi_\infty = \mathcal{M}_0(x)$.

On sait que dans ce cas , on peut écrire $\mathcal{M}_0(x) = z^T(x) \cdot x$

avec :

$$z(x) = \begin{vmatrix} 0 \\ \vdots \\ sig(\xi_i) \\ \vdots \\ 0 \end{vmatrix} \quad , \quad |\xi_i| = \max_{j=1,\dots n} |\xi_j|$$

Le plan sur lequel , pour trouver x_n , on doit projeter orthogonalement x_{n-1} n'est autre que

$$z^T \cdot (Ax - y) = 0 ,$$

c'est à dire $f_i(x) = 0$, i désignant le rang de la composante de plus grande valeur absolue de $r(x_{n-1})$:

On voit donc que, aucours de l'itération seuls les plans donnés sont utilisés, le procédé est donc celui des projections sur les plans donnés, mais au lieu de projeter sur ces plans dans un ordre imposé comme dans la méthode de KACZMARZ [27] classique (cf aussi BODEWIG [28]) à chaque étape on choisit , un peu comme dans la méthode de relaxation de SOUTHWELL [29] , le plan qui correspond au plus fort résidu en module.

La formule de calcul est :

$$x_n = x_{n-1} - \mathcal{M}_0(r(x_{n-1})) \cdot sig(r_i(x_{n-1})) \cdot \frac{A_{i.}^T}{\|A_{i.}\|^2} = x_{n-1} - r_i(x_{n-1}) \cdot \frac{A_{i.}^T}{\|A_{i.}\|^2}$$

et ici :

$$L(x_{n-1}) = \frac{\mathcal{M}_0^2(r(x_{n-1}))}{\|x_{n-1} - \omega\|^2} \cdot \frac{1}{\|A_{i.}\|^2}$$

donc si $\mu^\circ = \max_i \|A_{i.}\|$

en utilisant comme au chapitre IV , m'_A pour désigner : $\min_{X \neq 0} \frac{\mathcal{M}(AX)}{\|X\|}$

la formule déterminant le rapprochement devient:

$$\| \omega - x_p \| \leq \| \omega - x_0 \| \cdot \left(\sqrt{1 - \left(\frac{m'_A}{\mu^0} \right)^2} \right)^p$$

Enfin si le système de plans donné est orthogal le procédé doit théoriquement, s'arrêter aubout de n tours au plus . Il est à peine besoin de remarquer que dans une utilisation pratique de ce procédé il serait recommandé de transformer le système donné : $Ax = y$ en un système tel que les lignes de la matrice A soient toutes de même longueur I , avec cela les formules de calcul sont :

$$x_{p-1} \to x_p \quad \begin{cases} \xi_1^{(p)} = \xi_1^{(p-1)} - \rho_i^{(p-1)} \cdot a_{i_1} \\ \xi_2^{(p)} = \xi_2^{(p-1)} - \rho_i^{(p-1)} \cdot a_{i_2} \\ \vdots \qquad\qquad \vdots \\ \xi_n^{(p)} = \xi_n^{(p-1)} - \rho_i^{(p-1)} \cdot a_{i_n} \end{cases}$$

$$| \rho_i^{(p-1)} | = \max_{j=1,\dots n} | \rho_j^{(p-1)} |$$

$$r(x_p) = r(x_{p-1}) - \rho_i^{(p-1)} \cdot A \cdot A_i^T$$

6°) Procédé associé à la norme $\varphi_2(x) = \Phi(x)$.

On a vu que dans ce cas ,on peut écrire : $\Phi(x) = \zeta^T(x) \cdot x$

avec:

$$\zeta(x) = \begin{vmatrix} m_j(\xi_1) \\ m_j(\xi_2) \\ \vdots \\ m_j(\xi_n) \end{vmatrix} = m(x)$$

Les plans sur lesquels pour trouver x_p ,on doit projeter orthogonalement x_{p-1} sont les plans passant par ω et perpendiculaires aux vecteurs:

$$A^T \cdot E$$

si E parcours l'ensemble des vecteurs de composantes de valeur absolue I .

La formule de calcul est : $x_p = x_{p-1} - \frac{\Phi(r(x_{p-1}))}{\| A^T m(x_{p-1}) \|^2} \cdot A^T m(x_{p-1})$

étudions encore ici la convergence de ce procédé,

$$L(x_{k-1}) = \frac{\Phi^2(r(x_{k-1}))}{\|x_{k-1} - \omega\|^2} \cdot \frac{1}{\|A^T m(x_{k-1})\|^2}$$

et par suite

$$L(x_{k-1}) \geq \frac{m_A^2}{\|A^T m(x_{k-1})\|^2}$$

Or on a vu au chapitre IV que $M_A = \max_E \|AE\|$, donc

$$L(x_{k-1}) \geq \frac{m_A^2}{M_A^2} = \gamma^2 \quad , \quad \|x_k - \omega\| \leq \|x_{k-1} - \omega\| \cdot \left(\sqrt{1-\gamma^2}\right)^k$$

Nota au § précédent on aurait pu aussi bien remarquer que $\mu^0 = M'_A$, donctomber sur une formule très analogue:

$$\|x_k - \omega\| \leq \|x_0 - \omega\| \cdot \left(\sqrt{1-\gamma'^2}\right)^k$$

ou γ' est le conditionnement défini par les deux normes $\mathcal{M}_0$ et $\|\ \|$.

Or il est facile de voir que, puisque l'on sait que les conditionnemts sont a peu près équivalents, en gros les deux procédés sont de même convergence .

Les formules de calcul sont les suivantes :

$$x_k = x_{k-1} - \Phi(r(x_{k-1})) \cdot \frac{A^T E_{k-1}}{\|A^T E_{k-1}\|^2}$$

$$E_{k-1} = m(x_{k-1})$$

$$r(x_k) = r(k-1) - \Phi(r(x_{k-1})) \cdot \frac{AA^T E_{k-1}}{\|A^T E_{k-1}\|^2}$$

Cette méthode a été effectivement utilisée sur la machine électroniqque GAMMA A.E.T de l'université de Grenoble . Ont été étudiés des systèmes de 4,6,8 et I2 équations . La convergence est très faible dans le cas d'un système ayant comme matrice de premier membre la matrice H_6 de HILBERT . Cela n'a rien de surprenant d'aprés ce que nous avons dit au chapitre IV , sur les conditionnements de telles matrices . Par contre, sur un système d'ordre 8 , où la méthode d'itération classique de GAUSS-SEIDEL ne donne pas de résultats , et citée par FREEMANN [30] , à partir du vecteur initial nul , la méthode conduit au bout de 50 tours d'itération à 3 chiffres significatifs éxacts. Dans le cas d'un système d'ordre I2 , cinq chiffres ont été obtenus après 250 tours.

7°) Le procédé associé à la norme $\varphi_2(x) = \|x\|$.

Nous pouvons bien entendu, écrire $\|x\| = \frac{x^T}{\|x\|} \cdot x$

avec : $\zeta(x) = \frac{x}{\|x\|}$

Le procédé consiste à projeter orthogonalement r_{p-1} sur le plan :

$$r^T(x_{p-1}) \cdot (A.x - y) = (A x_{p-1} - y)^T \cdot (A.x - y) = 0$$

Nous obtenons ainsi la formule :

$$x_p = x_{p-1} - \frac{\|r(x_{p-1})\|^2}{\|A^T r(x_{p-1})\|^2} \cdot A^T r(x_{p-1})$$

et par suite ,

$$L(x_{p-1}) = \frac{\|r(x_{p-1})\|^2}{\|x_{p-1} - \omega\|^2} \cdot \frac{\|r(x_{p-1})\|^2}{\|A^T r(x_{p-1})\|^2}$$

Donc,

$$L(x_{p-1}) \geq \frac{\lambda_1}{\lambda_n}$$

Si λ_1 et λ_n sont les plus petite et plus grande valeur propre de $A.A^T$ (ou de $A^T.A$) .

Par suite la formule donnant une évluation de la convergence est dans ce cas encore :

$$\|x_p - \omega\| \leq \|x_0 - \omega\| \cdot \left(\sqrt{1 - \gamma^2}\right)^p$$

On peut donc affirmer:

Théorème II Dans les trois procédés associés aux trois normes $\varphi_1, \varphi_2, \varphi_\infty$, on peut dire que :

$$\|x_p - \omega\| \leq \|x_0 - \omega\| \cdot \left(\sqrt{1 - \gamma^2}\right)^p$$

si γ est le conditionnement général associé d'une des trois normes et de la norme euclidienne .

Dans ce cas les formules de calcul sont les suivantes:

$$x_p = x_{p-1} - \frac{\|r(x_{p-1})\|^2}{\|A^T r(x_{p-1})\|^2} \cdot A^T . r(x_{p-1})$$

$$r(x_p) = r(x_{p-1}) - \frac{\|r(x_{p-1})\|^2}{\|A^T r(x_{p-1})\|} \cdot A.A^T r(x_{p-1})$$

Nous allons étudier maintenant les relations entre les procédes que nous venons d'exposer et les méthodes de gradient.

3°) Méthode du gradient .

Avec les notations du § I , soit x_0 un point de R^n , par ce point il passe une seule des frontières $\Pi^0_\omega(A)$ de la région $\mathcal{B}^0_\omega(A)$, cette surface admet pour équation ;

$$\varphi(r(x)) = \varphi(r(x_0))$$

Supposons que x_0 soit tel qu'en ce point il existe un plan tangent à $\Pi^0_\omega(A)$ cela revient à dire que $\varphi(r(x))$ soit différentiable en x_0

Or $$\frac{\partial \varphi(r(x))}{\partial x_i} = \sum_{j=1}^{j=n} \frac{\partial \varphi}{\partial X_j} \frac{\partial X_j}{\partial x_i} = (A^T)_{i\cdot} \cdot p(r(x)) ; (X_j - f_j(x))$$

Donc le plan tangent est le plan perpendiculaire au vecteur :

$$p_1(x_0) = A^T \cdot p(A(x)) = A^T \cdot p(r(x_0))$$

Le plan $z^T(Ax - y) = 0$ passant par ω sera parallèle au plan tangent si:

$$p_1(x_0) = A^T \cdot z \quad , \quad z = p(r(x_0))$$

Supposons qu'il en soit ainsi et cherchons l'intersection de ce plan et de la normale en x_0 à $\Pi^0_\omega(A)$.

Ce point d'intersction x_1 est défini par :

$$x_1 = x_0 + \lambda \cdot p_1$$

(λ scalaire) et par le fait d'être dans ce plan ce qui permet d'obtenir:

$$z^T(r(x_0) + \lambda A p_1) = 0$$

d'où

$$\lambda = - \frac{z^T \cdot r(x_0)}{p_1^T \cdot A^T z}$$

et par suite :

$$x_1 = x_0 - \frac{z^T \cdot r(x_0)}{p_1^T A^T z} \cdot p_1$$

Mais si l'on remarque:

$$p_1 = A^T z = A^T \cdot p(r(x_0))$$

cela peut s'écrire:

$$\boxed{x_1 = x_0 - \frac{z^T \cdot r(x_0)}{\|p_1\|^2} \cdot p_1 = x_0 - \frac{p^T(r(x_0)) \cdot r(x_0)}{\|p_1\|^2} \cdot p_1}$$

Prenons nos différentes normes:

1°) Pour $\varphi_\infty(x) = \eta_0(x)$, on a : $p(x) = p(x)$

$$x_1 = x_0 - \eta_0(x_0) \cdot \frac{A^T p(r(x_0))}{\|A^T p(r(x_0))\|^2}$$

ce qui redonne bien la formule du §5

2°) Pour $\varphi_1(x) = \Phi(x)$, on a ; $p(x_0) = m(x)$

et la formule se réduit bien à : $x_p = x_{p-1} - \dfrac{\Phi(r(x_{p-1}))}{\| A^T m(x_{p-1}) \|^2} \cdot A^T m(x_{p-1})$

qui est celle du § 6

3°) pour $\varphi_2(x) = \|x\|$, on a $\mu(x) = \dfrac{x}{\|x\|}$

et par suite encore :

$$x_p = x_{p-1} - \frac{\| r(x_{p-1}) \|^2}{\| A^T r(x_{p-1}) \|^2} \cdot A^T r(x_{p-1})$$

On en conclut que pour les trois cas qui viennent d'être cités la méthode de la décomposition de la norme coincide avec celle du gradient .

Toutes les fois qu'une norme est dérivable dans une certaine région contenant x_0 , on peut écrire pour tout $\lambda > 0$,

$$\varphi(\lambda x_0) = \lambda \varphi(x_0)$$

la fonction $F_1(\lambda) = \varphi(\lambda x_0)$ est dérivable pour $\lambda = 1$, donc :

$$\varphi(x_0) = \left(\frac{\partial \varphi}{\partial \xi_1}\right)_0 \cdot \xi_1^0 + \cdots + \left(\frac{\partial \varphi}{\partial \xi_n}\right)_0 \xi_n^0 = \text{grad}^T \varphi \cdot x$$

cela prouve comment on peut localement découvrir une décomposition d'une norme .

9°) Généralisation de la méthode

Nous allons voir comment on peut généraliser la méthode que l'on vient d'étudier au cas où l'on ne fait pas forcément des projections orthogonales de x_{k-1} sur les plans passant par la solution et qui sont associés à ces points.

Je me restreins au cas de l'utilisation de la norme $\Phi(x)$.

Considérons les 2^k régions déterminées dans R^k par les plans

$$A_i x - y_i \equiv f_i(x) = 0$$

du système donné . Soit R_j , $(j = 1, 2, \ldots, 2^k)$, une telle région , c'est un lieu de points de R^k pour les quels les signes des $f_i(x)$ restent les mêmes .

Soient α_j , $(j = 1, 2, \ldots, 2^k)$ un ensemble de 2^k directions de R^k définies par des vecteurs α_j , unitaires , $\|\alpha_j\| = 1$ et associés bi-univoquement aux régions R_j : à α_j correspondra R_j de même indice .

D'après ce que l'on a vu plus haut , si x_0 est un point de la région R_j , le plan :

$$\pi_j : \quad m^T(x_0) \cdot (Ax - y) = 0$$

est le plan passant par ω et parallèle à la "face" du polyèdre $\prod_\omega^0(A)$, soit π_j' , qui passe par x_0 .

Au point x_0 faisons correspondre le point x_1 où la parallèle à α_j menée par x_0 coupe Π_j .(On choisit α_j non parallèle à Π_j)

on a : $x_1 = x_0 + t\alpha_j$.

et t est tel que : $m^T(x_0).\left(A(t\alpha_j + x_0) - y\right) = 0$

oubien : $t.\, m^T.A\alpha_j + m^T(x_0).(Ax_0 - y) = 0$

et si l'on remarque que : $m^T(x_0).(Ax_0 - y) = \Phi_1(x_0) = \Phi(z(x_0))$

cela donne : $t = -\Phi_1(x_0) \,/\, m^T.A.\alpha_j$:

$$x_1 = x_0 - \frac{\Phi_1(x_0)}{m^T(x_0).A\alpha_j}\cdot \alpha_j$$

formule trés simple .

La question importante est de savoir si le procédé est convergent et comment il converge dans ce cas .

Il est clair que:

$$\|\omega - x_0\|^2 = \|x_1 - x_0\|^2 + \|x_1 - \omega\|^2 - 2\,(x_1 - x_0)^T.(x_1 - \omega)$$

donc:

$$\|\omega - x_1\|^2 = \|\omega - x_0\|^2 - \|x_1 - x_0\|^2 + 2\,(x_1 - x_0)^T.(x_1 - \omega)$$

et ensuite succéssivement :

$$\|\omega - x_1\|^2 = \|\omega - x_0\|^2 - \frac{\Phi_1^2(x_0)}{(m^T.A.\alpha_j)^2} - 2\,\frac{\Phi_1(x_0)}{(m^TA\alpha_j)}\,\alpha_j^T.(x_1 - \omega)$$

$$= \|\omega - x_0\|^2 - \frac{\Phi_1^2(x_0)}{(m^TA\alpha_j)^2} - 2\,\frac{\Phi_1(x_0)}{(m^TA\alpha_j)}\,.\,\alpha_j^T\left[x_0 - \omega - \frac{\Phi_1(x_0)}{m^TA\alpha_j}\,\alpha_j\right]$$

$$= \|\omega - x_0\|^2\left[1 - \frac{\Phi_1(x_0)}{(m^TA\alpha_j)^2\,\|\omega - x_0\|^2}\left[2\,m^TA\alpha_j\,\alpha_j^T(x_0 - \omega) - \Phi_1(x)\right]\right]$$

$$= \|\omega - x_0\|^2\left[1 - L(x_0)\right]$$

avec

$$L(x_0) = \frac{\Phi_1(x_0)}{(m^TA\alpha_j)^2\,\|\omega - x_0\|^2}\left[2\,(m^TA\alpha_j).\alpha_j^T(x_0 - \omega) - \Phi_1(x_0)\right]$$

et puisque : $\Phi_1(x_0) = m^T A\,(x_0 - \omega)$

$$L(x_0) = \frac{\Phi_1(x_0)}{(m^TA\alpha_j)^2\,\|x_0 - \omega\|^2}\left[2\,(m^TA\alpha_j).\alpha_j - A^Tm\right]^T.(x_0 - \omega)$$

Remarquons que si l'on fait $\alpha_j = \frac{A^T m}{\| A^T m \|}$ dans cette formule on trouve bien $L(x_0) = \Phi_1^2(x_0) / \| A^T m \|^2 . \| x_0 - \omega \|^2$ qui correspond au procédé deja vu .

Posons

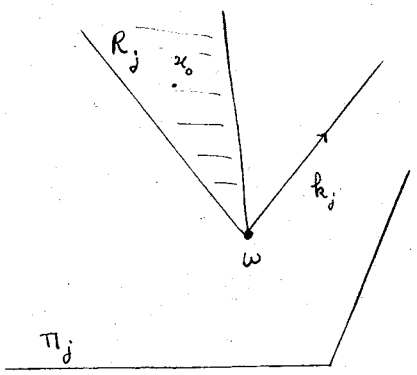
$$k_j = 2 . (m^T A \alpha_j) . \alpha_j - A^T m$$

Alors

$$L(x_0) = \frac{\Phi_1(x_0)}{(m^T A \alpha_j)^2 \| x_0 - \omega \|^2} k_j^T . (x_0 - \omega)$$

La question, pour nous permettre d'affimer la conver gence, peut se poser ainsi :

Peut-on choisir les α_j de sorte que $L(x)$ soit toujours > 0 et le plus grand possible ? (il est toujours ≤ 1)



1°) Il est clair que l'on peut toujours faire un choix des k_j de sorte que $k_j^T (x_0 - \omega) > 0$ ($x_0 \in R_j$). En effet si l'on remarque que R_j est un demi-cône convexe tout entier dans l'un des deux demi-espaces déterminés par le plan Π_j . Il suffit de choisir un k_j (convenablement orienté) de sorte que en balayant R_j $x_0 - \omega$, ne devienne jamais perpendiculaire à celui-ci . Il suffit de chbisir dans la portion du cône supplémentaire de R_j qui contient le vecteur $\frac{A^T m}{\| A^T m \|}$ que nous savons convenir .

2°) les k_j étant choisis selon la règle que nous venons de voir il nous faut déterminer les α_j cela se fait en résolvant les équations:

$$2 (m^T . A \alpha_j) . \alpha_j = k_j + A^T m$$

on tire $2 (m^T A \alpha_j)^2 = m^T A k_j + \| A^T m \|^2$

D'où $$S_j = m^T A \alpha_j = + \sqrt{\frac{\| A^T m \|^2 + m^T A k_j}{2}}$$

ayant S_j on en déduit:

$$\alpha_j = \frac{1}{2 S_j} \left[k_j + A^T m \right]$$

et pour satisfaire à la condition $\| \alpha_j \| = 1$

il suffit de prendre k_j de sorte que :

$$1 = \| \alpha_j \|^2 = \frac{\| k_j \|^2 + \| A^T m \|^2 + 2 k_j^T . A^T m}{2 \| A^T m \|^2 + 2 k_j^T A^T m}$$

$$\| k_j \| = \| A^T m \|$$

Cela fait :

$$L(x_0) = \frac{\Phi_1(x_0)}{\|x_0 - \omega\|} \cdot \frac{k_j^T (x_0 - \omega)}{\|x_0 - \omega\|} \cdot \frac{2}{m^T A k_j + \|A^T m\|^2}$$

et puisque $k_j^T \cdot (x_0 - \omega) \geqslant \|k_j\| \cdot \|x_0 - \omega\| \cdot \cos \mu_j$

μ_j étant le maximum de l'angle aigu que fait k_j avec $x_0 - \omega$ si x_0 parcourt la région R_j .

donc $$L(x_0) \geqslant \frac{\Phi_1(x_0)}{\|x_0 - \omega\|} \|k_j^T\| \cos \mu_j \cdot \frac{2}{m^T A k_j + \|A^T m\|^2} \geqslant \frac{\Phi_1(x_0)}{\|x_0 - \omega\|} \cos \mu_j$$

soit $m = \min_j (\cos \mu_j)$

il en résulte :

$$L(x_0) \geqslant$$

et puisque $\frac{\Phi_1(x)}{\|x - \omega\|} \geqslant m_A$

on a :

$$L(x) \geqslant m_A \cdot m$$

enfin :

$$\|x_1 - \omega\| \leqslant \|x_0 - \omega\| \cdot \sqrt{1 - m_A \cdot m}$$

Si l'on a déterminé les α_j par les conditions précédentes on peut donc énoncer :

Théorème III Le procédé de résolution relatif à la norme Φ converge encore si les projections orthogonales sont remplacées par des projections parallèles à des directions α_j convenablement choisies .

Ce qui fait l'intéret de ce théorème est le corollaire suivant:

Corollaire En calcul numérique le procédé relatif à la décomposition de la norme Φ est pratiquement insensible aux erreurs de détermination des vecteurs $A^T m$.

En effet si le calcul fournit, au lieu de $A^T m$, un vecteur $\overline{A^T m}$ on peut dire, vu que les erreurs sont tout de même assez faibles, que l'on à remplacé les directions par des directions α_j voisines , Donc que le procédé sera celui-ci-dessus , avec des $\alpha_j = \overline{A^T m}$ convenables; Seule la rapidité de la convergence peut en être affectée .

CHAPITRE VIII

Calcul du polynôme caractéristique, des valeurs et vecteurs propres .

1°) Réduction à une forme de FROBENIUS .

Nous utiliserons dans ce qui va suivre des matrices antiscalaires K :

$$K = X.Y^T$$

avec :

$$X = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{vmatrix} \quad , \quad Y = \begin{vmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{vmatrix}$$

nous poserons : $\lambda = Y^T.X = X^T.Y = x_1y_1 + x_2y_2 + \cdots + x_ny_n$

le polynôme minimal pour K est alors : $m(u) = u^2 - \lambda u$

Si l'on se donne donc $M = \alpha I + \beta K$ son inverse : $M^{-1} = \alpha' I + \beta' K$

avec α' et β' , déterminés par les formules du Chap.I , la transmuée d'une matrice $A = (a_{ij})$ $(i, j = 1, 2, \cdots)$ par M sera si $p = -\frac{\alpha}{\beta} \neq 0$ et $p \neq \lambda$:

$$A' = A + \frac{\lambda - p}{p^2 - \lambda p} K.A + \frac{p}{p^2 - \lambda p} A.K - \frac{1}{p^2 - \lambda p} K.A.K$$

ou :

$$(II) \qquad A' = A - \frac{1}{p} K.A + \frac{1}{p - \lambda} A.K - \frac{1}{p(p - \lambda)} K.A.K$$

Dans le cas que nous traitons :

$$AK = (AX).Y^T \quad , \quad KA = X.(Y^TA)$$

$$KAK = XY^TAXY^T = (Y^TAX).XY^T = (Y^TAX).K$$

Nous écrirons cette formule:

a) En explicitant le terme général a'_{ij} de A'

$$a'_{ij} = a_{ij} - \frac{1}{p} x_i (a_{1j} y_1 + a_{2j} y_2 + \cdots a_{nj} y_n) + \frac{1}{p-\lambda} y_j (a_{i1} x_1 + \cdots + a_{in} x_n) - \frac{1}{p(p-\lambda)} (Y^T A X) x_i y_j$$

$$a'_{ij} = a_{ij} - \frac{1}{p} x_i (Y^T . A_{.j}) + \frac{1}{p-\lambda} y_j (A_{i.} . X) - \frac{1}{p(p-\lambda)} (Y^T A X) x_i y_j$$

b) En explicitant une colonne (de rang j) de A'

$$A'_{.j} = A_{.j} - \frac{1}{p} (Y^T . A_{.j}) . X + \frac{1}{p-\lambda} y_j . (AX) - \frac{1}{p(p-\lambda)} (Y^T A X) . y_j X$$

Dautre part, on sait ([31], [32]) que si l'équation caractéristique d'une matrice admet des racines simples,

il est possible, par des opérations rationnelles dans le corps de ses coéfficients, de trouver une matrice $\mathcal{F}$ semblable à celle-ci et de la forme " de FROBENIUS " :

$$\mathcal{F} = \begin{vmatrix} 0 & 0 & \cdots & & p_1 \\ 1 & 0 & & & p_2 \\ 0 & 1 & 0 & & p_3 \\ & & 1 & 0 & \vdots \\ & & & \ddots & 0 \\ 0 & \cdots & & 1 & p_n \end{vmatrix}$$

c'est à dire qu'il existe une matrice $\mathcal{M}$ dont les éléments s'obtiennent par des opérations rationnelles à partir de ceux de A, non singulière et telle que :

$$\mathcal{F} = \mathcal{M} . A . \mathcal{M}^{-1}$$

Nous alons prouver qu'il est possible de trouver une matrice telle que $\mathcal{M}$ par un produit de matrices du 2° degré trés simples et qui permettent de programmer effectivement un tel calcul .

I) Première tansmutation

Faisons $y_1 = 0$ et laissons les x_i et $y_2, y_3 \ldots y_n$ indéterminés pour l'instant .

On aura :

$$A'_{.1} = A_{.1} - \frac{1}{p} (Y^T . A_{.1}) . X$$

Remarquons d'abord qu'il est impossible de choisir les x_i et y_i qui restent de sorte que les composantes de $A'_{.1}$ soient toutes nulles, car dans ce cas on aurait :

$$X = \frac{\rho}{(Y^T A_{.1})} A_{.1}$$

d'où :

$$\lambda = Y^T X = \frac{\rho}{(Y^T A_{.1})} (Y^T A_{.1}) = \rho$$

ce qui est incompatible avec les utilisations des formules.

Par contre, et dans le but de commencer la mise en forme de FROBENIUS, peut-on faire en sorte que : $A'_{.1} = e_2$?

(e_i désigne toujours le i^e vecteur de la base fondamentale de R^n).

Pour réaliser cela il suffit :

1°) de prendre $\rho = Y^T A_{.1}$.

2°) de prendre $x_1 = a_{11}$, $x_3 = a_{31}$, $x_4 = a_{41}$, ... , $x_n = a_{n1}$

alors

$$X = A_{.1} - a_{21} e_2 + x_2 e_2$$

puis

$$A'_{.1} = A_{.1} - X = (a_{21} - x_2) e_2 = e_2$$

si l'on fait $x_2 = a_{21} - 1$; $X = A_{.1} - e_2$

Voyons si ce choix donne $\rho \neq 0$ et $\rho \neq \lambda$:

Il est clair que laissé sur les $y_2, y_3, \ldots y_n$ (arbitraire) permet (si $A_{.1} \neq 0$ ce qui est supposé, A n'étant pas à première colonne nulle) d'obtenir $\rho = Y^T A_{.1} \neq 0$

Ensuite,

$$\lambda = Y^T X = Y^T (A_{.1} - e_2)$$

Donc $\rho = \lambda + y_2$

il faut prendre $y_2 \neq 0$, $y_2 = 1$ par exemple .

Cela nous conduit à prendre :

$$Y = \begin{vmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{vmatrix} = e_2$$

d'où $\rho = a_{21}$ (si $a_{21} \neq 0$) et $\lambda = a_{21} - 1$

Alors $A'_{.1} = e_2$.

Pour les autres colonnes de A' on aura :

$$A'_{.2} = A_{.2} - \frac{a_{22}}{a_{21}} (A_{.1} - e_2) + A(A_{.1} - e_2) - \frac{1}{a_{21}} (e_2^T A (A_{.1} - e_2)) \cdot (A_{.1} - e_2)$$

$$A'_{.j} = A_{.j} - \frac{a_{2j}}{a_{21}} (A_{.1} - e_2) \qquad (j = 3, \ldots, n)$$

la formule qui donne $A'_{.2}$ se simplifie en remarquant que : $A \cdot e_2 = A_{.2}$

et que $e_2^T . A . (A_{.1} - e_2) = e_2^T . A . A_{.1} - e_2^T A_{.2} = (A.A_{.1})_2 - a_{22}$

en désignant par $(A.A_{.1})_i$ la i^{e} composante de $(A.A_{.1})$

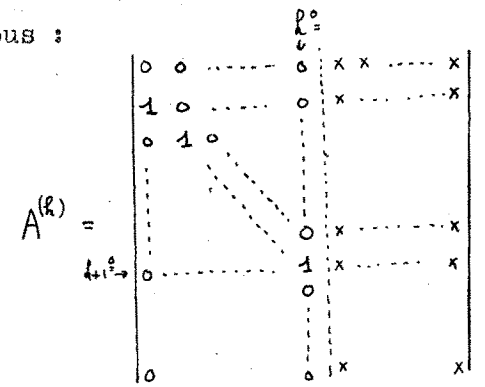
Alors, $A'_2 = A_{.2} - \frac{1}{a_{21}} \left[a_{22} + (A.A_{.1})_2 - a_{22} \right] (A_{.1} - e_2) + A.A_{.1} - A_{.2}$

cela nous fournit les formules très simples :

$$A' : \begin{cases} A'_{.1} = e_2 \\ A'_{.2} = A.A_{.1} - \dfrac{(A.A_{.1})_2}{a_{21}} . (A_{.1} - e_2) \\ A'_{.j} = A_{.j} - \dfrac{(A_{.j})_2}{a_{21}} . (A_{.1} - e_2) \end{cases}$$

2) Transmutation générale

Supposons être parvenu à transmuer A en $A^{(k)}$, de la forme ci-dessous :



$A^{(k)} =$, $k \leq n-2$

C'est à dire que les k premières colonnes de $A^{(k)}$ sont à termes nuls sauf les termes sur la première parallèle au dessous de la diagonale qui sont des I .

Nous allons montrer qu'il existe un choix des ρ, x, y tels que la transmutation qu'ils définissent assure à la transmuée de $A^{(k)}$ la forme $A^{(k+1)}$.

La formule devant donner les colonnes de $A^{(k+1)}$ est encore :

$$A^{(k+1)}_{.j} = A^{(k)}_{.j} - \frac{1}{\rho} \cdot (y^T A^{(k)}_{.j}) \cdot x + \frac{1}{\rho - \lambda} y_j A^{(k)} x - \frac{1}{\rho(\rho - \lambda)} (y^T A^{(k)} x) \cdot y_j x$$

Faisons ici $y_1 = y_2 = \cdots\cdots = y_{k+1} = 0$. Il est évident que $y^T . A^{(k)}_{.j} = 0$ pour $j = 1, 2, \ldots\ldots, k$, puisque les $A^{(k)}_{.j}$ $(j = 1, 2, \ldots\ldots k)$ n'ont que des 0 dans les lignes $k+2, k+3, \ldots\ldots, n$.

Donc : $A^{(k+1)}_{.j} = A^{(k)}_{.j}$ $\quad (j = 1, 2, \ldots\ldots, k)$

Les k premières colonnes de $A^{(k+1)}$ sont celles de $A^{(k)}$.

Prenons ensuite : $y = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ 1 \leftarrow k+2 \\ 0 \\ \vdots \end{vmatrix} = e_{k+2}$

Alors : $A^{(k+1)}_{\cdot k+1} = A^{(k)}_{\cdot k+1} - \frac{1}{p}\left(e^T_{k+2} \cdot A^{(k)}_{\cdot k+1}\right) \cdot X$

et pour faire en sorte que $A^{(k+1)}_{\cdot k+1} = e_{k+2}$

il suffit de prendre : $p = e^T_{k+2} \cdot A^{(k)}_{\cdot k+1} = a^{(k)}_{k+2,\,k+1} \quad (\neq 0)$

et $X = A^{(k)}_{\cdot k+1} - e_{k+2}$

alors

$$\lambda = Y^T X = a^{(k)}_{k+2,\,k+1} - 1 = p - 1 \neq p$$

si bien que les formules donnant $A^{(k)}_{\cdot j}$, $(j = k+1, \ldots\ldots n)$ sont :

$$A^{(k+1)}_{\cdot k+1} = e_{k+2}$$

$$A^{(k+1)}_{\cdot k+2} = A^{(k)}_{\cdot k+2} - \frac{1}{a^{(k)}_{k+2,\,k+1}}\left(e^T_{k+2} \cdot A^{(k)}_{\cdot k+2}\right)\left(A^{(k)}_{\cdot k+1} - e_{k+2}\right) + A^{(k)} \cdot \left(A^{(k)}_{\cdot k+1} - e_{k+2}\right) - \frac{1}{a^{(k)}_{k+2,\,k+1}}\left[e^T_{k+2} A^{(k)} \cdot \left(A^{(k)}_{\cdot k+1} - e_{k+2}\right)\right] \cdot \left(A^{(k)}_{\cdot k+1} - e_{k+2}\right)$$

$$A^{(k+1)}_{\cdot j} = A^{(k)}_{\cdot j} - \frac{a^{(k)}_{k+1,\,j}}{a^{(k)}_{k+2,\,k+1}}\left(A^{(k)}_{\cdot k+1} - e_{k+2}\right) \quad ; \; (j = k+3, \ldots\ldots n)$$

si l'on simplifie la 2° formule comme on l'a déja fait ,

$$A^{(k+1)}_{\cdot k+2} = A^{(k)}_{\cdot k+2} - \frac{1}{a^{(k)}_{k+2,\,k+1}}\left[a^{(k)}_{k+2,\,k+2} + \left(A^{(k)} \cdot A^{(k)}_{\cdot k+1}\right)_{k+2} - a^{(k)}_{k+2,\,k+2}\right]\left(A^{(k)}_{\cdot k+1} - e_{k+2}\right) + A^{(k)} \cdot A^{(k)}_{\cdot k+1} - A^{(k)}_{\cdot k+1}$$

Il en résulte les formules de calcul :

$$A^{(0)} = A$$

$$A^{(k)} \longrightarrow A^{(k+1)} \quad (k = 0, 1, \ldots, n-2)$$

$$\begin{cases} A^{(k+1)}_{\cdot j} = e_{j+1} & (j = 1, 2, \ldots\ldots, k+1) \\ A^{(k+1)}_{\cdot k+2} = A^{(k)} \cdot A^{(k)}_{\cdot k+1} - \dfrac{\left(A^{(k)} \cdot A^{(k)}_{\cdot k+1}\right)_{k+2}}{a^{(k)}_{k+2,\,k+1}}\left[A^{(k)}_{\cdot k+1} - e_{k+2}\right] & \\ A^{(k+1)}_{\cdot j} = A^{(k)}_{\cdot j} - \dfrac{\left(A^{(k)}_{\cdot j}\right)_{k+2}}{a^{(k)}_{k+2,\,k+1}} \cdot \left[A^{(k)}_{\cdot k+1} - e_{k+2}\right] & (j = k+3, \ldots\ldots n) \end{cases}$$

Et l'on voit que $A^{(n-1)}$ aura bien la forme de FROBENIUS $(e_2, e_3, \ldots\ldots, e_n, \Pi) = \mathcal{F}$

avec $\Pi = A^{(n-1)}_{\cdot n}$.

Donc le polynôme carractéristique de A n'est autre que :

$$F(\lambda) = (-1)^n\left[\lambda^n - \lambda^{n-1}\left(a^{(n-1)}_{n,n}\right) - \lambda^{n-2}\left(a^{(n-1)}_{n-1,n}\right) \ldots\ldots\ldots - \left(a^{(n-1)}_{1,n}\right)\right]$$

Il nous reste à étudier les différents "incidents " pouvant survenir dans l'application de l'algorithme que nous venons de définir .

Il est évident que la seule condition d'impossibilité de la transmutation $A^{(k)} \longrightarrow A^{(k+1)}$ est $a^{(k)}_{k+2,\,k+1} = 0$.

Supposons être dans ce cas . Alors deux possibilités :

a) Tous les $a^{(k)}_{j,\,k+1} = 0$, pour $j = k+2, \ldots\ldots n$. Alors $A^{(k)}$ a la forme :

$$A^{(k)} = \begin{array}{c} \\ k^\circ = \\ k+1^\circ = \\ k+2^\circ = \end{array} \begin{array}{c} \begin{array}{cccccc} & & & & k^\circ & k+1^\circ \end{array} \\ \left|\begin{array}{cccccc} 0 & 0 & \cdots & 0 & a^{(k)}_{1,k+1} & \times \cdots \times \\ 1 & 0 & \cdots & 0 & a^{(k)}_{2,k+1} & \times \cdots \times \\ 0 & 1 & 0 & & & \\ & \ddots & \ddots & & & \\ & & 1 & 0 & & \\ 0 & & & 1 & a^{(k)}_{k+1,k+1} & \times \cdots \times \\ 0 & & & & 0 & \\ \vdots & & & & 0 & B \\ 0 & & & & 0 & \end{array}\right| \\ \leftarrow n-(k+1) \rightarrow \end{array}$$

B désignant une matrice carrée d'ordre $n-(k+1)$.

Il est alors évident que si $F(\lambda)$ est le polynôme carractéristique de cherché , onpeut écrire :

$$F(\lambda) = \mathrm{Det}\begin{vmatrix} -\lambda & \cdots & a^{(k)}_{1,k+1} \\ 1 & -\lambda & a^{(k)}_{2,k+1} \\ & 1 & \ddots \\ & & a^{k}_{k+1,k+1} - \lambda \end{vmatrix} \cdot \mathrm{Det}\left(B - \lambda . I_{n-(k+1)}\right)$$

C'est à dire que $F(\lambda)$ se décompose en le produit de

a) $(-1)^{(k+1)}\left[\lambda^{k+1} - \lambda^{k} . a^{(k)}_{k+1,\,k+1} - \lambda^{k-1}\, a^{(k)}_{k,\,k+1} - \cdots\cdots - a^{(k)}_{1,\,k+1}\right]$

b) du polynôme carractéristique de B .

b) Supposons $a^{(k)}_{k+2,\,k+1} = 0$ mais, $a^{(k)}_{e,\,k+1} \neq 0$, $e > k+2$

soit $V = I - E_{ii} - E_{jj} + E_{ij} + E_{ji}$

on rappelle que V_{ij} est du 2° degré, que $V^{-1}_{ij} = V_{ij}$,

$V_{ij} . M$ est identique à M mais la i° ligne est échangée avec la j° ligne

$M . V_{ij}$ est identique à M mais la i° colonne est échangée avec la j° col.

Faisons alors $A^{(k)}_1 = A^{(k)} . V_{k+2,\,e}$: c'est une matrice identique à $A^{(k)}$ sauf échange de la $k+2^\circ$ col. et de la e° col., la $k+1^\circ$ colonne est celle de $A^{(k)}$.

Puis $A'^{(k)} = V_{k+2,\ell} \cdot A_1^{(k)}$: matrice identique à $A_1^{(k)}$ sauf échange de la $k+2$ ligne et de la ℓ° ligne .

Donc, Dans $A'^{(k)}$ le terme $a'^{(k)}_{k+2,k+1} = a^{(k)}_{\ell,k+1} \neq 0$, elle est bien de la même forme que $A^{(k)}$, enfin elle est semblable à $A^{(k)}$, $A'^{(k)} = V_{k+2,\ell} A^{(k)} V^{-1}_{k+2,\ell}$. Si l'on tombe sur un cas b) il suffit de remplacer $A^{(k)}$ par $A''^{(k)}$ et de poursuivre l'algorithme .

Ce qui précède prouve le thèorème [33]

Théorème II Etant donnée une matrice carrée d'ordre n , à éléments dans un corps $\mathcal{R}$, A, il est possible de la transmuer en une matrice de la forme de FROBENIUS générale , c'est à dire de la forme :

$$\Phi = \begin{vmatrix} \mathcal{F}_1 & & & \times \\ & \mathcal{F}_2 & & \\ & & \mathcal{F}_3 & \\ 0 & & & \\ & & & \mathcal{F}_k \end{vmatrix}$$

Les $\mathcal{F}_1, \mathcal{F}_2, \ldots \mathcal{F}_k$ étant de forme de Frobénius (simple):

$$\mathcal{F}_i = \begin{vmatrix} 0 & 0 & \cdots & \pi_1^i \\ 1 & 0 & \cdots & \pi_2^i \\ & 1 & 0 & \\ 0 & 0 & 1 & \pi_\ell^i \end{vmatrix} \quad ; \text{ ou bien : } \mathcal{F}_i = 0$$

Ces transmutations sont toutes de carractére rationnel en le corps $\mathcal{R}$ et l'on peut écrire :

$$\Phi = \mathcal{M} . A \mathcal{M}^{-1}$$

La matrice $\mathcal{M}$ étant un produit de matrices du 2° degré de la forme soit $-\rho I + X_i \cdot e_i^T$, soit V_{ij} .

De la démonstration qui précède on déduit la possibilité de toujours calculer le polynôme carxactéristique de A par cette mèthode . Il en résulte aussi un autre résultat intéressant :

Théorème III Si dans le corps $\mathcal{R}$ auquel appartiennent les éléments a_{ij} d'une matrice A le polynôme carxactéristique de celle-ci est irréductible, Il existe toujours une suite de matrices du 2° degré dont le produit transmue A en une forme simple de FROBENIUS .

C'est une simple conséquence de la démonstration précédente si l'on remarque que dans l'hypothèse du thèorème on ne peut pas tomber dans la

circonstance du cas particulier a)

La méthode que nous venons de décrire est de programmation facile, surtout sur une machine possédant un programme d'assemblage matriciel.

Dans le passage de $A^{(k)}$ vers $A^{(k+1)}$ il nous faut :

I) Calculer $A^{(k)}_{.1} \cdot A^{(k)}_{.k+1}$ ce qui vu la forme de $A^{(k)}$ exige $(n-k).n$ produits

2) Le calcul des termes des nouvelles colonnes $(k+2$ à $n)$ demande $(n-k-1).n$ produits.

3) Il y a enfin à executer $n-k-1$ divisions par le même facteur: $a^{(k)}_{k+1,k+1}$

Le "côut" total est donc de $n-k-1$ divisions et de $n(2(n-k)-1)$ multiplications dans le passage $A^{(k)} \to A^{(k+1)}$.

La transmutation complète doit exiger :

$$\sum_{k=0}^{k=n-2} (n-k-1) = 1+2+\cdots+(n-1) = \frac{n(n-1)}{2} = N_D \quad \text{divisions}$$

$$n \sum_{k=0}^{k=} [2(n-k)-1] = n.(n^2-1) = N_M \quad \text{multiplications.}$$

Ce nombre est de l'ordre de n^3.

2°) Comparaison avec la méthode de DANILEWSKI.

Le procéde que l'on vient de voir ressemble beaucoup au procédé de DANILEWSKI [34], [35], qui ne coûte que de l'ordre de n^3 multiplications aussi. Il est facile de voir que ces procédés résultent tous les deux de transmutations succéssives de la matrice donnée par des matrices du 2° degré.

En effet notre procédé consiste à prendre au premier pas:

$$A^{(1)} = \left(I - \frac{1}{a_{21}} K\right) \cdot A \cdot \left(I + K\right)$$

avec

$$K = \begin{vmatrix} 0 & a_{11} & 0 & \cdots & 0 \\ 0 & a_{21}-1 & 0 & \cdots & 0 \\ 0 & a_{31} & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & a_{n1} & 0 & \cdots & 0 \end{vmatrix}$$

comme il résulte aisément de ce que l'on a établi.

Le procédé de DANILEWSKI, utilise

$$A'^{(1)} = \left(I - \frac{1}{a_{21}} K'\right) \cdot A \cdot \left(I + \frac{1}{a_{21}} K'\right)$$

avec $K'^2 = 0$

et $K' = \begin{vmatrix} 0 & a_{11} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a_{n1} & 0 & \cdots & 0 \end{vmatrix}$

on peut remarquer que :

$$\left(1 - \frac{1}{a_{21}} K\right) = \begin{vmatrix} 1 & \cdots & 0 \\ \frac{1}{a_{21}} & & \\ & 1 & \\ 0 & \cdots & 1 \end{vmatrix} \cdot \left(1 - \frac{1}{a_{21}} K'\right)$$

- On voit que la différence des deux méthodes est très peu sensible ; elles ne sont, en fait, que des cas particuliers de la méthode de transmutation générale par des matrices du 2° degré. D'autres choix de matrices K sont possibles, et cela conduira, pour la recherche d'une forme de FROBENIUS équivalente à une matrice donnée, à des algorithmes voisins de ces deux que l'on vient de présenter.

3°) Transmutation en une forme Triple-diagonale .

Nous allons voir qu'il est possible de transmuer une matrice donnée en une matrice de forme triple-diagonale:

$$\mathcal{C}_D = \begin{vmatrix} \alpha_1 & \beta'_1 & & & \\ \beta_1 & \alpha_2 & \beta'_2 & & 0 \\ & \beta_2 & \alpha_3 & \beta'_3 & \\ & & \ddots & \ddots & \ddots \\ & 0 & & \beta_{n-1} & \alpha_n \end{vmatrix}$$

n'ayant que des termes nuls sauf dans la diagonale principale et dans les deux parallèles à celle-ci immédiatement au dessus et au dessous .

Nous utiliserons des matrices

$$M = a.I + \mathcal{C}.K \quad , \quad M^{-1} = a'.I + \mathcal{C}'.K$$

avec K , de la forme $K = X.Y^T$.

Nous allons nous astreindre à ne prendre ici que des X et Y tels :

$$X = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ x_\ell \\ \vdots \\ x_m \\ \vdots \\ 0 \end{vmatrix} \quad , \quad Y = \begin{vmatrix} 0 \\ 0 \\ \vdots \\ y_\ell \\ \vdots \\ y_m \\ \vdots \\ 0 \end{vmatrix}$$

où ℓ et m sont deux indices différents $\ell < m$, pour fixer les idées .

Alors $K = (k_{ij})$ n'a que 4 termes non nuls :

$$k_{\ell\ell} = x_\ell . y_\ell \quad , \quad k_{m\ell} = x_m . y_\ell$$

$$k_{\ell m} = x_\ell . y_m \quad , \quad k_{mm} = x_m . y_m$$

Nous avons ici : $\lambda = x_\ell y_\ell + x_m y_m$

Soit $A = (a_{ij})$ une matrice donnée, si on la transmue par M en $A' = M.A.M^{-1}$ on a, vu que :

$$a'_{ij} = a_{ij} - \frac{1}{\rho} x_i (a_{j\ell} y_\ell + \cdots + a_{jm} y_m) + \frac{1}{\rho - \lambda} y_j (a_{i\ell} x_\ell + \cdots + a_{im} x_m) - \frac{(Y^T A X)}{\rho(\rho - \lambda)} x_i y_j$$

Il en résulte que A' est identique à A, sauf pour les lignes et colonnes de rang ℓ ou m.

Si l'on pose

$$R = a_{\ell\ell} x_\ell y_\ell + a_{\ell m} x_m y_\ell + a_{m\ell} x_\ell y_m + a_{mm} x_m y_m$$

la formule ci-dessus donne :

$$a'_{ij} = a_{ij} - \frac{1}{\rho} x_i [a_{\ell j} y_\ell + a_{mj} y_m] + \frac{1}{\rho - \lambda} y_j [a_{i\ell} x_\ell + a_{im} x_m] + \frac{R x_i y_j}{\rho(\lambda - \rho)} \quad (i, j = \ell, m)$$

Soit r un indice tel que $r < \ell < m$ par exemple,

$$a'_{mr} = a_{mr} - \frac{1}{\rho} x_m [a_{\ell r} y_\ell + a_{mr} y_m]$$

et son "symétrique" :

$$a'_{rm} = a_{rm} + \frac{1}{\rho - \lambda} y_m [a_{r\ell} x_\ell + a_{rm} x_m]$$

On dispose des paramètres $\rho, x_\ell, x_m, y_\ell, y_m$, peut-on déterminer ceux-ci afin que: I°) Ces deux termes soient nuls

2°) Que les conditions de possibilité $\rho \neq 0$ et $\rho \neq \lambda$ soient assurées ?

$$\left(\lambda = x_\ell . y_\ell + x_m y_m \right).$$

Il suffit donc de choisir $\rho, x_\ell, x_m, y_\ell, y_m$, de sorte que :

$$0 = \rho . a_{mr} - x_m [a_{\ell r} y_\ell + a_{mr} y_m]$$

$$0 = (\rho - \lambda) a_{rm} + y_m [a_{r\ell} x_\ell + a_{rm} x_m]$$

de là on tire :

$$a_{mr} [\rho - x_m y_m] - x_m y_\ell a_{\ell r} = 0$$

$$a_{rm} [\rho - x_\ell y_\ell] + x_\ell y_m a_{r\ell} = 0$$

Par suite :

$$x_m [y_\ell a_{\ell r} + y_m a_{mr}] = \rho . a_{mr}$$

$$x_\ell [y_\ell a_{rm} - y_m a_{r\ell}] = \rho . a_{rm}$$

Il est facile de voir, d'après les dernières formules écrites et sous la condition que l'on n'ait pas simultanément :

$$a_{ee} = a_{re} = a_{mr} = a_{rm} = 0 \quad .$$

(Dans ce dernier cas l'on aurait déjà $a_{mr} = a_{rm} = 0$ et, par suite, il est inutile de se poser la question de trouver une matrice semblable à A où ces termes sont nuls : c'est A elle-même) que l'on peut toujours trouver y_e et y_m et prendre ρ de sorteque :

$$y_e\, a_{er} + y_m\, a_{mr} \neq 0 \quad , \quad y_e\, a_{rm} - y_m\, a_{re} \neq 0$$

Avec un tel choix pour y_e et y_m et ρ ,

$$x_m = \rho \cdot a_{mr} / (y_e\, a_{er} + y_m\, a_{mr})$$

$$x_e = \rho \cdot a_{rm} / (y_e\, a_{rm} - y_m\, a_{re})$$

Dans une programmation générale, on peut faire $y_e = y_m = \rho = 1$ et cela fournit ,dans le but d'éviter des calculs :

$$x_m = a_{mr} / (a_{er} + a_{mr}) = \xi_m$$

$$x_e = a_{rm} / (a_{rm} - a_{re}) = \xi_r$$

Si $a_{er} + a_{mr} = 0$ ou $a_{rm} - a_{re} = 0$, il est commode de choisir encore $\rho = 1$ et de prendre pour y_e et y_m des nombres tels que $y_e = \varepsilon$ et $y_m = \varepsilon'$ avec $|\varepsilon| = |\varepsilon'| = 1$, il est clair que l'on arrivera toujours à en déterminer de cette sorte et convenables .

Dans ces cas les formules de calcul sont les suivantes :

$$A \xrightarrow[(H)]{} A' \quad \begin{cases} a'_{ij} = a_{ij} & (i \neq e,m \; ; \; j \neq e,m) \\ a'_{ij} = a_{ij} - x_i \left[a_{ej}\, y_e + a_{mj}\, y_m \right] + \dfrac{1}{1 - x_m y_m - x_e y_e} \left[y_j \left(a_{ie} x_e + a_{im} x_m \right) - R \cdot x_i\, y_j \right] \\ R = a_{ee}\, x_e y_e + a_{em}\, x_m y_e + a_{me}\, x_e y_m + a_{mm}\, x_m y_m \end{cases}$$

on peut donc énoncer : [36]

Théorème III Etant donnée une matrice carrée A d'ordre n , on peut toujours trouver une matrice $M = \alpha I + \beta K$ du 2° degré qui la transmue en une matrice A' de sorte que : Si $r < \ell < m \leq n$ sont trois indices donnés

I°) La matrice A' ne diffère de A que par les lignes et colonnes de rang ℓ et m .

2°) La matrice A' ait deux termes nuls en les positions: (r, m) et (m, r) .

On désignera par $\textcircled{H}(r; \ell, m)$ une telle transmutation .

Soit alors A une matrice carrée quelconque d'ordre n,

Faisons la transmutation : $\textcircled{H}(1; 2, 3) \longrightarrow A'$ Cette matrice est identique à A sauf pour les lignes et colonnes de rang 2 ou 3 , de plus il se trouve un zéro en les positions $(1,3)$ et $(3,1)$.

Faisons ensuite, sur cette matrice A' la transmutation $\textcircled{H}(1;2,4) \rightarrow A''$, elle change seulement les lignes et colonnes de rang 2 et 4 , donc ne place pas un nombre en la place des deux zéros qui sont en $(1,3)$ et $(3,1)$. Par contre elle en place deux dans les positions $(1,4)$ et $(4,1)$.

Il est facile de comprendre que la suite des transmutations ci-après :

$$\textcircled{H}(1;2,3)\ ,\ \textcircled{H}(1;2,4)\ ,\ \textcircled{H}(1;2,5)\ ,\ \dots\dots\dots\dots\ ,\ \textcircled{H}(1;2,n)\ ,$$

$$\textcircled{H}(2;3,4)\ ;\ \textcircled{H}(2;3,5)\ ,\ \textcircled{H}(2;3,6)\ ;\ \dots\dots\dots\dots\ ,\ \textcircled{H}(2;3,n)\ ,$$

$$\vdots$$

$$\textcircled{H}(n-3;n-2,n-1)\ ,\ \textcircled{H}(n-3;n-2,n)\ ,$$

$$\textcircled{H}(n-2;n-1,n)$$

Permet toujours d'obtenir , pour celles de la première ligne , des zéros dans les positions $\left|\begin{smallmatrix}1,3\\3,1\end{smallmatrix}\right|$, $\left|\begin{smallmatrix}1,4\\4,1\end{smallmatrix}\right|$, ... , $\left|\begin{smallmatrix}1,n\\n,1\end{smallmatrix}\right|$; celles de la seconde des zéros dans les positions $\left|\begin{smallmatrix}2,4\\4,2\end{smallmatrix}\right|$ ..... $\left|\begin{smallmatrix}2,n\\n,2\end{smallmatrix}\right|$; sans perturber les zéros de la première ligne ... etc .....

En bref, au bout de ces $\frac{(n-2)(n-1)}{2}$ transmutations l'on obtient bien la forme triple-diagonale $\mathcal{C}_D$.

$$\mathcal{C}_D = \begin{vmatrix} \alpha_1 & \beta'_1 & & & \\ \beta_1 & \alpha_2 & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \beta'_{n-1} & \\ & 0 & & \beta_{n-1} & \alpha_n \end{vmatrix} .$$

De cela il résulte le théorème suivant:

Théorème IV Etant donnée une matrice A quelconque il existe toujours une suite de matrices du 2° degré dont le produit $\mathcal{M}$ est tel qu'il transmue A en une forme $\mathcal{C}_D$ triple diagonale ;

$$\mathcal{C}_D = \mathcal{M} . A \mathcal{M}^{-1}$$

Pour achever la détermination du polynôme caractéristique de A une fois $\mathcal{C}_D$ trouvée ,celui-ci est

$$F(\lambda) = \text{Dét} \begin{vmatrix} \alpha_1-\lambda & \beta'_1 & & \\ \beta_1 & \alpha_2-\lambda & \ddots & \\ & \ddots & \ddots & \beta'_{n-1} \\ & & \beta_{n-1} & \alpha_n-\lambda \end{vmatrix}$$

Pour en faire le calcul , on pose

$$f_i(\lambda) = \text{Dét} \begin{vmatrix} \alpha_1-\lambda & \beta'_1 & & \\ \beta_1 & \alpha_2-\lambda & \ddots & \\ & \ddots & \ddots & \beta'_{i-1} \\ & & \beta_{i-1} & \alpha_i-\lambda \end{vmatrix}$$

Si bien que $f_n(\lambda) = F(\lambda)$, $f_1(\lambda) = \alpha_1 - \lambda$

et on peut aussi poser $f_0(\lambda) = 1$.

Alors en developpant par rapport à la dernière ligne on trouve :

$$f_i(\lambda) = (\alpha_i - \lambda) . f_{i-1}(\lambda) - \beta_{i-1} \beta'_{i-1} f_{i-2}(\lambda)$$

Cette méthode est à rapprocher avec celle de LANCZOS ou de GIVENS , qui pour chercher une matrice équivalente à une matrice donnée et de forme triple diagonale utilisent des techniques bien différentes .GIVENS [38] [39] Transmue A par des "rotations",c'est à dire comme dans la méthode de JACOBI $A' = \Omega^T A \Omega$ est telle que Ω (*) soit orthonormale , la méthode de GIVENS est exclusivement,sous sa forme primitive, réservée au cas de matrices symétriques et dans ce cas particulièrement commode puisque la suite des polynômes $f_i(\lambda)$ ci-dessus est une suite de STURM du polynôme carractéristique.

La méthode de LANCZOS est très différente ,elle se propose de construire une base,dans laquelle la transformation ,définie dans la base fondamentale de R^n par A,soit définie par une matrice de forme triple-diagonale . On trouvera unexposé de cette méthode classique en [40] [41] . .

(*) Remarquons que les matrices des "rotations" Ω de GIVENS sont de la forme : $I + X.Y^T$.

4°) Etude des erreurs dans la méthode de réduction à une forme de FROBENIUS .

Nous allons revenir à la méthode exposée au §4 et étudier rapidement, en utilisant un mode de raisonnement analogue à celui fait pour l'étude des erreurs dans la résolution de systèmes linéaires , les erreurs de calcul pouvant arriver dans l'utilisation de ce procédé .

Pour simplifier l'écriture des formules de transmutation de $A^{(k)}$ en $A^{(k+1)}$ posons : $A^{(k)} = A$, $A^{(k+1)} = A'$, $\pi = a^{(k)}_{k+2,\,k+1}$

$$X = A^{(k)}_{.\,k+1} - e_{k+2} \ .$$

on a vu que alors :

$$A' = \left(I - \frac{1}{\pi} X.e^{T}_{k+2}\right).A.\left(I + X.e^{T}_{k+2}\right)$$

D'autre part, pour la détermination de A' il ne faut faire des calculs que pour les éléments $a^{(k+1)}_{i,j}$ de $A^{(k+1)}$ situés dans les colonnes de rang $k+2, k+3, \ldots\ldots n$.

On peut donc poser , si $\bar{A}'$ est la matrice réellement obtenue à la place de A' :

$$\bar{A}' = A' + \delta A' = A' + \begin{vmatrix} 0 \ldots\ldots 0 & \varepsilon_{1,k+2} & \ldots\ldots & \varepsilon_{1,n} \\ 0 \ldots\ldots 0 & \varepsilon_{2,k+2} & \ldots\ldots & \varepsilon_{2n} \\ \vdots & \vdots & & \vdots \\ 0 \ldots\ldots 0 & \varepsilon_{n,k+2} & \ldots\ldots & \varepsilon_{nn} \end{vmatrix}$$

les ε_{ij} étant les erreurs de calcul faites dans la détermination des termes $a^{(k+1)}_{ij}$ de $A^{(k+1)}$.

Or il est facile de déterminer δA de sorte que:

$$\bar{A}' = \left(I - \frac{1}{\pi}.X.e^{T}_{k+2}\right).(A+\delta A).\left(I + X.e^{T}_{k+2}\right)$$

c'est à dire :

$$\delta A' = \left(I - \frac{1}{\pi} X.e^{T}_{k+2}\right).\delta A.\left(I + X.e^{T}_{k+2}\right)$$

et puisque $I - \frac{1}{\pi} X.e^{T}_{k+2}$ est l'inverse de $I + X.e^{T}_{k+2}$,

$$\delta A = \left(I + X.e^{T}_{k+2}\right) \delta A' \left(I - \frac{1}{\pi} X.e^{T}_{k+2}\right)$$

Or si $\delta A'$ a la forme indiquée (ses $k+1$ premières colonnes nulles),

Il est très facile de voir que

$$\delta A = \delta A' + X e_{k+2}^T \delta A' - \frac{1}{\eta} \delta A' X e_{k+2}^T - \frac{1}{\eta} X e_{k+2}^T \delta A' X e_{k+2}^T$$

admet aussi ses $k+1$ premières colonnes nulles . Cette remarque entraine l'importante constatation : Si l'on veut soumettre $A+\delta A$ à la transmutation lui donnant la forme $(A+\delta A)^{(k+1)}$, on fera, puisque cette transmutation ne dépend que de la $k+1^{\circ}$ colonne de la matrice, qui est <u>la même</u> pour A et pour $A+\delta A$, la même transmutation, c'est à dire que :

$\bar{A}'$,obtenue par une transmutation de A , avec des erreurs de calcul peut être considérée comme obtenue par transmutation de $A+\delta A$,sans erreurs de calcul , δA ayant la forme obtenue plus haut .

Revenons aux notations générales , Partons de la matrice A , au lieu de la suite de matrices semblables :

$$A^{\circ} = A \ , \ A^{(1)} , A^{(2)} , \ldots\ldots , A^{(n-1)}$$

telles que :

$$A^{(k+1)} = M_k . A^{(k)} . M_k^{-1}$$

on obtient par le calcul effectif la suite

$$\bar{A}^{\circ} = A \ , \ \bar{A}^{(1)} , \bar{A}^{(2)} , \ldots\ldots , \bar{A}^{(n-1)}$$

telles que :

$$\bar{A}^{(k+1)} = M_k . \left(\bar{A}^{(k)} + \delta A^{(k)} \right) . M_k^{-1}$$

les $\delta A^{(k)}$ étant des matrices ayant leur $k+1$ premières colonnnes nulles et se calculant comme on l'a vu plus haut à partir des erreurs de calcul sur la détermination de A' .

Or ..... :

$$\bar{A}^{(k)} = M_{k-1} \left(\bar{A}^{(k-1)} + \delta A^{(k-1)} \right) . M_{k-1}^{-1}$$

et ..... :

$$\bar{A}^{(k+1)} = M_k . \left[M_{k-1} \left(\bar{A}^{(k-1)} + \delta A^{(k-1)} \right) . M_{k-1}^{-1} + \delta A^{(k)} \right] . M_k^{-1}$$

Considérons le produit :

$$M_k . \left[M_{k-1} \left(\bar{A}^{(k-1)} + \delta A^{(k-1)} \right) M_{k-1}^{-1} + \delta A^{(k)} \right] . M_k^{-1}$$

On peut l'écrire ,

$$M_k . M_{k-1} \left[\bar{A}^{(k-1)} + \delta A^{(k-1)} + M_{k-1}^{-1} \delta A^{(k)} M_{k-1} \right] M_{k-1}^{-1} . M_k^{-1}$$

et ensuite ,

$$\bar{A}^{(k+1)} = M_k \ldots M_0 \left[A^{\circ} + \delta A^{\circ} + M_0^{-1} \delta A' M_0 + M_0^{-1} M_1^{-1} \delta A^{2} M_1 M_0 + \ldots\ldots \right] M_0^{-1} . M_1^{-1} \ldots\ldots M_k$$

enfin ,

$$\bar{A}^{(n-1)} = \mathcal{M} \left[A + \sum_{i=0}^{i=n-2} M_0^{-1} \ldots M_{i-1}^{-1} \delta A^{(i)} M_{i-1} \ldots M_0 \right] \mathcal{M}^{-1}$$

on peut donc énoncer :

Thèorème V Si l'on transmue une matrice A donnée, carrée, d'ordre n, en une matrice de forme de FROBENIUS, selon la méthode ci-dessus exposée, on obtient une forme qui peut être considérée comme rigoureusement semblable à la matrice $A+\delta A$, avec :

$$\delta A = \sum_{i=0}^{i=k-2} M_0^{-1} \cdots M_{i-1}^{-1} \, \delta A^{(i)} M_{i-1} \cdots M_0$$

Il nous faut dire un mot de la façon dont on peut déduire de la connaissance de δA une estimation des erreurs faites sur les valeurs propres et vecteurs propres .

5°) Variations des éléments propres correspondant à la variation δA de A .

Soit A une matrice carrée, d'ordre n, ayant ses n valeurs propres $\lambda_1, \lambda_2, \ldots\ldots \lambda_n$, distinctes .

Soient, u_i et v_i les vecteurs propres à droite et à gauche pour A et correspondant à la valeur propre λ_i :

(1) $\quad A u_i = \lambda_i u_i$

(2) $\quad v_i^T A = \lambda_i v_i^T \qquad \left(A^T v_i = \lambda_i v_i \right)$

Je les supposerai "normalisés" de sorte que :

(3) $\quad u_i^T . u_i = 1$

(4) $\quad v_i^T . u_i = 1$

On désignera par T la matrice d'ordre n ayant pour colonnes les vecteurs colonnes $u_1, u_2, \ldots\ldots, u_n$. les lignes de T^{-1} sont les v_i^T.

Soit donc $A+\delta A$ une matrice légèrement variée de A, δA ayant par exemple, la valeur déterminée au § précédent.

$\lambda_i + \delta\lambda_i$ désignera la valeur propre de $A+\delta A$ voisine de λ_i de A et $u_i + \delta u_i$ et $v_i + \delta v_i$ les vecteurs propres à droite et à gauche correspondant à celle-ci.

On doit avoir :

$$\begin{cases} (1)' & \left(A+\delta A-(\lambda_i I+\delta\lambda_i I)\right)\cdot(u_i+\delta u_i)=0 \\ (2)' & \left(v_i^T+(\delta v_i)^T\right)\cdot\left(A+\delta A-(\lambda_i I+\delta\lambda_i I)\right)=0 \end{cases}$$

écrivons :

$$(1)'' \quad (\delta A-\delta\lambda_i I)\cdot u_i+(A-\lambda_i I)\cdot\delta u_i+\underline{\delta A\cdot\delta u_i-\delta\lambda_i\delta u_i}=0$$

$$(2)'' \quad v_i^T(\delta A-\delta\lambda_i I)+(\delta v_i)^T\cdot(A-\lambda_i I)+\underline{(\delta v_i)^T\cdot\delta A-\delta\lambda_i\cdot(\delta v_i)^T}=0$$

Si l'on néglige les parties soulignées comme étant d'ordre supérieur :

$$(1)''' \quad (\delta A-\delta\lambda_i I)\cdot u_i+(A-\lambda_i I)\,\delta u_i=0$$

$$(2)''' \quad v_i^T(\delta A-\delta\lambda_i I)+(\delta v_i)^T\cdot(A-\lambda_i I)=0$$

Multiplions $(1)'''$ par v_k^T (à gauche)

$$v_k^T\,\delta A\,u_i-(\delta\lambda_i)\,v_k^T\cdot u_i+v_k^T A\,\delta u_i-\lambda_i\,v_k^T\delta u_i=0$$

$$v_k^T\,\delta A\,u_i-(\delta\lambda_i)\,v_k^T\cdot u_i+(\lambda_k-\lambda_i)\cdot v_k^T\delta u_i=0$$

de même en multipliant $(2)'''$ par u_k (à droite)

$$v_i^T\,\delta A\,u_k-(\delta\lambda_i)\,v_i^T u_k+(\lambda_k-\lambda_i)(\delta v_i)^T u_k=0$$

Or :

$$v_k^T\cdot u_i=\delta_{ki}$$

(δ_{ki} symbole de KRONECKER), d'où :

$$(I)\begin{cases} k=i : & v_i^T\delta A\,u_i=\delta\lambda_i \\ k\neq i : & v_k^T\delta u_i=\dfrac{1}{\lambda_i-\lambda_k}\,v_k^T\,\delta A\,u_i \end{cases}$$

$$(II)\begin{cases} k=i : & v_i^T\delta A\,u_i=\delta\lambda_i \\ k\neq i : & (\delta v_k)^T\cdot u_i=\dfrac{1}{\lambda_k-\lambda_i}\,v_k^T\delta A\,u_i \end{cases}$$

Ces relations obtenues, soit à

déterminer $\delta\lambda_i$, δu_i, δv_i , en supposant connus δA, u_i, v_i .

I°) Les $\delta\lambda_i$ sont déterminés par :

$$\delta\lambda_i = v_i^T . \delta A . u_i$$

2°) Les δu_i , pour i fixé , sont tels que :

$$v_k^T . \delta u_i = \frac{1}{\lambda_i - \lambda_k} . v_k^T \delta A \, u_i$$

cela donne pour $k = 1, 2, \ldots i-1, i+1, \ldots n$ seulement $n-1$ équations avec n inconnues : les composantes de δu_i , soient : $\delta\xi_i^1, \delta\xi_i^2 \ldots \delta\xi_i^n$.

mais si l'on pense qu'en permanence : $v_k^T . u_i = \delta_{ki}$ c'est à dire :

$$\left(v_k^T + (\delta v_k)^T\right) . \left(u_i + \delta u_i\right) = \delta_{ki} \quad ;$$

cela fait :

$$(\delta v_k)^T . u_i + v_k^T (\delta u_i) = 0$$

en négligeant toujours les termes du 2° ordre $(\delta v_k)^T . \delta u_i$.

Donc le système fournissant les δu_i et δv_i ($2.n^2$ inconnues) s'écrira :

$$(3) \begin{cases} v_k^T . \delta u_i = b_{ki} & , \ k \neq i \\ (\delta v_k)^T . u_i = -b_{ki} & , \ k \neq i \end{cases}$$

$$(4) \quad (\delta v_k)^T . u_i + v_k^T . \delta u_i = 0$$

où l'on a posé :

$$b_{ki} = \frac{1}{\lambda_i - \lambda_k} . v_k^T \delta A \, u_i \qquad (i \neq k)$$

La remarque importante à faire est celle-ci :

les relations (4) (pour $i \neq k$) sont toujours satisfaites si (3) le sont

, seules les relations (4) pour $i = k$ sont indépendantes de ces relations (3) . Notre système à $2.n^2$ inconnues est donc :

$$(3) \begin{cases} v_k^T . \delta u_i = b_{ki} & , \ k \neq i \\ (\delta v_k)^T . u_i = -b_{ki} & , \ k \neq i \end{cases}$$

$$(4)' \quad (\delta v_i)^T . u_i + v_i^T . \delta u_i = 0$$

Ce qui nous fait $2n^2 - n$ équations seulement . Il nous manque donc, en ce point de notre étude , n , conditions pour résoudre le problème .

Si A est symétrique, le cas est classique et a été traité par JACOBI [42] .
Voyons le cas général :

Ecrivons que les δu_i sont tels que $u_i^T . u_i = 1$ se maintienne.
Cela donne les **n** conditions désirées :

$$(5) \qquad u_i^T . (\delta u_i) = 0$$

Donc, le système devant fournir les δu_i et δv_i sera :

$$(S) \quad (3) \begin{cases} v_k^T . (\delta u_i) = b_{ki} \\ \\ u_k^T . (\delta v_i) = -b_{ik} \end{cases} \quad (i \neq k) \; ; \; (4) \; u_i^T . (\delta v_i) + v_i^T . (\delta u_i) = 0 \; ; \; (5) \; u_i^T . (\delta u_i) = 0 \quad (i = 1, \ldots n)$$

Nous allons indiquer comment on peut résoudre ce système .

Commençons par déterminer les δu_i.

Ils satisfont à :

$$\begin{cases} v_k^T . (\delta u_i) = b_{ki} \\ u_i^T . (\delta u_i) = 0 \end{cases}$$

Or on sait que

$$T^{-1} = \begin{pmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{pmatrix}$$

donc :

$$E_{ii} . T^{-1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ v_i^T \\ \vdots \\ 0 \end{pmatrix}$$

ensuite

$$T = (u_1, \ldots, u_n) \quad , \quad T . E_{ii} = (0, 0, \ldots, u_i, 0 \ldots 0) \quad , \quad E_{ii} T^T = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ u_i^T \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Il en résulte que le système donnant les composantes de δu_i sera :

$$\left(T^{-1} - E_{ii} T^{-1} + E_{ii} T^T\right) . \delta u_i = \beta_i \quad , \quad \left[I - E_{ii} (I - T^T T)\right] . T^{-1} \delta u_i = \beta_i$$

Or il est très facile de voir que la matrice $E_{ii} (I - T^T T)$ est de carré nul :

$$\left(T^T . T\right)_{ii} = 1$$

donc

$$\left[I - E_{ii} (I - T^T T)\right]^{-1} = \left[I + E_{ii} (I - T^T T)\right]$$

et par suite ,

$$\boxed{\delta u_i = T \cdot \left[I + E_{ii} (I - T^T T) \right] \cdot \beta_i}$$

Enfin le système donnant les δu_i peut sécrire :

$$T^T \cdot \delta \sigma_i = \beta'_i$$

si

$$\beta'_i = - \begin{vmatrix} \beta_{i1} \\ \vdots \\ \sigma_i^T \delta u_i \leftarrow i^{ème} \\ \vdots \end{vmatrix} \quad ; \quad \boxed{\delta \sigma_i = (T^{-1})^T \cdot \beta'_i}$$

Ce qui permet de calculer les éléments cherchés .

Enfin signalons une majoration des $|\delta \lambda_i|$ qui résulte de

$$|\delta \lambda_i| = | \sigma_i^T \delta A \, u_i |$$

Si $\mathcal{C}(T^{-1})$ désigne plus grand des éléments de T^{-1} , (on sait que c'est une norme de la matrice T^{-1}), et $M_A = Max \dfrac{\Phi(AX)}{\| X \|}$ on peut écrire

$$|\delta \lambda_i| \leqslant \mathcal{C}(T^{-1}) \cdot M_{\delta A} .$$

ce qui fixe une borne supérieure pour la variation des valeurs propres .

6°) Cas particulier d'une matrice de forme de FROBENIUS .

Soit la matrice

$$A = \begin{vmatrix} 0 & 0 & \cdots & \cdots & p_n \\ 1 & 0 & & & p_{n-1} \\ 0 & 1 & 0 & & p_{n-2} \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & \cdots & 1 & p_1 \end{vmatrix}$$

Je suppose que son équation carractéristique :

$$F(\lambda) = (-1)^n \cdot \left[\lambda^n - p_1 \lambda^{n-1} - p_2 \lambda^{n-2} \cdots \cdots - p_n \right] = 0$$

n'a que des racines simples $\lambda_1, \lambda_2, \cdots \cdots, \lambda_n$.

Il est facile de voir que si l'on pose :

$$f_0(\lambda) = p_n$$

$$f_1(\lambda) = p_n + \lambda \, p_{n-1}$$

$$\vdots$$

$$f_i(\lambda) = p_n + \lambda \, p_{n-1} + \cdots\cdots + p_{n-i} \cdot \lambda^i$$

c'est à dire que les $f_i(\lambda)$ peuvent être calculés (Schéma de HORNER) par : $f_i(\lambda) = f_{i-1}(\lambda) + \lambda^i \cdot \mu_{n-i}$

le vecteur propre u_i de cette matrice correspondant à λ_i est de composantes

$$u_i = \begin{vmatrix} \xi_n \cdot f_0(\lambda_i)/\lambda_i \\ \xi_n \cdot f_1(\lambda_i)/\lambda_i^2 \\ \vdots \\ \xi_n \cdot f_{k-1}(\lambda_i)/\lambda_i^k \\ \vdots \\ \xi_n \cdot 1 \end{vmatrix} \quad , (\xi_n \neq 0, \text{ qq.})$$

La vecteur propre v_i à gauche est alors :

$$v_i = \begin{vmatrix} \eta_i \\ \eta_i \cdot \lambda_i \\ \vdots \\ \eta_i \cdot \lambda_i^{k-1} \\ \vdots \\ \eta_i \cdot \lambda_i^{n-1} \end{vmatrix} \quad , (\eta_i \neq 0, \text{ qq})$$

supposons que A varie de sorte que

$$\delta A = \begin{vmatrix} o & o & \cdots & o & \delta\mu_n \\ o & o & \cdots & o & \delta\mu_{n-1} \\ \vdots & & & & \\ o & & \cdots & o & \delta\mu_1 \end{vmatrix}$$

ce qui revient à dire que les coéfficients de l'équation caractérisque varient des $\delta\mu_i$. Il résulte des formules ci-dessus que :

$$\delta\lambda_i = \xi_n \cdot \eta_i \left(\delta\mu_n + \lambda_i \delta\mu_{n-1} + \cdots + \lambda_i^{n-1} \delta\mu_1 \right)$$

Or si l'on écrit que $v_i^T \cdot u_i = 1$, on trouve que ξ_n , η_i doivent être tels :

$$\frac{\xi_n \cdot \eta_i}{\lambda_i} \left[(n-1)\mu_n + (n-2)\mu_{n-1}\lambda_i + \cdots + \mu_2 \lambda^{n-2} + \lambda^n \right] = 1$$

ce qui donne :

$$\boxed{\delta\lambda_i = \frac{\lambda_i \left[\delta\mu_n + \lambda_i \delta\mu_{n-1} + \cdots + \lambda_i^{n-1} \delta\mu_1 \right]}{n\lambda_i^{n-1} + 2\mu_2 \lambda^{n-2} + \cdots + (n-1)\mu_{n-1}\lambda + n\mu_n}}$$

ou encore : $$\frac{\delta\lambda_i}{\lambda_i} = \frac{\delta\mu_n + \lambda_i \delta\mu_{n-1} + \cdots + \lambda_i^{n-1} \delta\mu_1}{n\mu_n + (n-1)\mu_{n-1}\lambda + \cdots + \lambda_i^{n-1}\mu_1}$$

et si l'on pose :

$$\pi^n f\left(\frac{x}{\pi}\right) = \Phi(x,\pi)$$

on a $n\Phi(x,\pi) = x\Phi'_x + \pi\Phi_\pi$, donc cela redonne la relation classique :

$$\lambda_i \cdot f'(\lambda_i) = n\mu_n + (n-1)\mu_{n-1}\lambda_i + \cdots + \lambda_i^{n-1}\mu_1 ,$$

$$\boxed{\delta\lambda_i = \frac{\delta\mu_n + \lambda_i \delta\mu_{n-1} + \cdots + \lambda_i^{n-1}\delta\mu_1}{f'(\lambda_i)}} .$$

BIBLIOGRAPHIE

On trouve une bibliographie très complète sur l'analyse numérique linéaire dans l'ouvrage :

"Simultaneous linear equations and the détermination of eigenvalues " édité par J.PAIGE et O.TAUSSKY . N.B.S. App.Math.Series .29 (I953)

Cet ouvrage comporte aussi une classification des différents ouvrages cités pour les différentes méthodes utilisées . Il ne cite ,malheureusement que des ouvrages publiés avant I95I .

Pour une bibliographie générale très récente et assez complète on pourra se rapporter à :

" Introduction to matrix analysis " par R.BELLMAN . Mac.GRAW-HILL .(I960).

Les ouvrages cités dans la liste qui suit sont ceux pour les quels un renvoi figure dans le texte .

[I] SPERNER et SCHEIRER "Matrix calculus and Modern Algebra " Chelsea .Pub. Comp . New-York .(I953) Chap .

[2] N.GASTINEL "Utilisation de matrices vérifiant une équation du 2° degré pour la transmutation de matrices " C.R.A.S 7 mars I960 .

[3] E.BODEWIG " Matrix calculus " North-Holland Pub.Comp. Amsterdam (I956) P ages IO et suiv .

[4] N.BOURBAKI " Topologie Générale " Chap IX ,§3 , Déf 5 , p 42 .

[5] N.BOURBAKI " Espaces Vectoriels Topologiques " Chap I,§3, Corol.I, ThI p 36 .

[6] S.BANACH " Théorie des opérations linéaires " Chelsea Pub.Comp? N.Y (I955) ch IV .

[7] N.BOURBAKI "Topologie générale " Chap IX,§3, Prop 7,p 43 .

[8] A.M.OSTROWSKI "Uber Normen von Matrizen " Math .Zeitsch. Bd 69 S p 2 (I955)

T.E.EASTERFIELD "Matix norms and Vector Measures " Duke Math J. Vol 24 pp 663-67I (I955) .

[9] E.BODEWIG Loc .Cit. p II7

[IO] N. GASTINEL "Conditionnement d'un système d'équations linéaires " C.R.A.S. T 248,p2 707-2 709, II mai I959 .

[II] G.H.HARDY,J.E.LITTLEWOOD,G.POLYA "Inequalities" Cambridge Univ.Press (I934

[I2] E.BODEWIG Loc.Cit .p II8

[I3] A.M. TURING "ROUding-off errors in matrix processes " Quart,J.mech.appl. math .I,I948,p287-308 .

[I4] A.T.LONSETH "The propagation of error in linear problems" Trans .Amér. Math .Soc. 62 p I93-2I2.

[I5] J.TODD " The condition of finite segments of the HILBERT matrix" N.B.S. Appl.Math Series.62.39 ,pII2 (I954).

[I6] J.von NEUMANN,H.H.GOLDSTINE "Numerical inverting of matrices of high order Bul.Amer.Math.Soc.Vol.53,No II,pp I02I-I099.(I947)
Formule (7 -5') p I093.

[I7] J.TODD Loc.Cit. p III formule (5-4)

[I8] A.S.HOUSEHOLDER "Introduction to numérical analysis " Mc GRaw-Hill (I956) p 44 , formule (2-08-22) .

[I9] A.R.COLLAR "On the reciprocation of certain matrices" Proc.Roy.Soc . Edingburgh,59,pI95-208 (I939).

[20] A.R. COLLAR "On the reciprocal of a segment of a generalised HILBERT matrix " Poc.Cambridge Phil? Soc. 47,pII-I7 (I95I)

[2I] E.H.LINFOOT , W.M.SHEPHERD "On a set of linear equations"Quart.J.Math. IO,84 (I939)

[22] E.LUKACS ,I.R.SAVAGE "TAbles of inverses of finite segments of the Hilbert matrix" N.B.S. Appl.Math.Series .39 .pI05 (I954)

[23] N.GASTINEL "Sur le choix des pivots dans l'élimination de GAUSS pour la résolution d'un système linéaire " C.R.A.S. T250,p275-277 II janv.I960 .

[24] R.BELLMAN "Matrix analysis " Mc GRAW-HILL .N.Y.(I960) Chap III.

[25] E.BODEWIG Loc.Cit. p38 .

[26] N.GASTINEL "Procédé itératif pour la résolution numérique d'un système d'équations linéaires " C.R.A.S. T 246 ,p 2571 -2574 . 5 mai 1958.

[27] KACZMARZ Bull .Inter.Académie Polonaise des Sciences A 1937,355-7

[28] E.BODEWIG Loc.Cit. p 163.

[29] J.ALLEN "Relaxations methods "

[30] FREEMANN Phil. Mag.34,1943,p 409-416.

[31] SPERNER et SCHEIRER Loc.Cit. Chap IV

[32] J.M.WEDDERBURN "Lectures on matrices " Am.Math.Soc.Colloq.Publ. Vol I7 (1934) .

[33] N.GASTINEL Loc.Cit (2)

[34] R.A.BUCKINGHAM "Numerical Methods" Pitman ed .London .(1957) p 375 .

[35] H.WAYLAND "Expansion of determinantal equations into polynomial form" Quart.Appl.Math.2 (1945) pp 277-306.

[36] N.GASTINEL Loc.Cit.(2).

[38] P.A.WHITE "The computation of eigenvalues and eigenvectors of a matrix" Jour.Soc.Indus.Appl.Math. Vol 6,No 4 ,(Dec 1958) p 418 .

[39] W.GIVENS "Computation of plane unitary rotations transforming a general matrix to triangular form". J.Soc.Indus.Appl.Math. Vol 6,No1,(Mars 1958) P26-50 .

W;GIVENS "Numérical computation of the characteristic values of a réal symétric matrix " Oak Ridge Nat.Lab.Rep. No 1574 (1954) .

[40] P.A.WHITE Loc.Cit. p 425.

[41] C.LANCZOS "An iteration method for the solution of the eigenvalue problem of linear differentialand integral operators " J.Res.Nat.B.S. 45, (1950) pp 255-282 .

[42] C.G.JACOBI ,J;de Crelle Vol 30,(1846) pp51-95 ,Voir aussi : E.BODEWIG Loc.Cit. p 286.

[43] BOURBAKI . Esp . Vect . Topo. Chap II,§ 5 ,Prop 3 ,PP 94 et suiv.

TABLE DES MATIERES

II ° THESE

LE THEOREME DE STONE -WEIERSTRASS

LE THEOREME DE STONE - WEIERSTRASS

-:-:-

I. - INTRODUCTION -

En 1885, WEIERSTRASS[1] a énoncé le théorème célèbre :
"Etant donné une fonction f, de la variable réelle, et à valeurs réelles, dont le domaine de définition est un segment $[a \quad b]$, et qui est continue sur ce segment, il existe toujours un polynôme qui l'approche d'aussi près que l'on veut uniformément, c'est-à-dire ε étant donné, il existe un polynôme P(x) tel que :

$$|f(x) - P(x)| \leqslant \varepsilon \qquad \text{pour tout } x \in [a \quad b] \text{ "} .$$

La démonstration originale de WEIERSTRASS provient du fait, apparemment fort éloigné de la question, que la fonction

$$\psi(x,k) = \frac{1}{k.\sqrt{\pi}} \int_{-\infty}^{+\infty} f(u).e^{-\left(\frac{u-x}{k}\right)^2}.du$$

tend uniformément vers f(x) pour $x \in [a \quad b]$, si $k \to 0$.

Ayant prouvé que $\psi(z,k)$ est une fonction entière de z (dans le plan complexe), on en déduit pour $\psi(x,k)$ (x réel) une série uniformément convergente dans $[a \quad b]$, il s'ensuit qu'il existe un polynôme tel

$$|\psi(x,k) - f(x)| \leqslant \frac{\varepsilon}{2}$$

et puisque k peut être choisi

assez petit afin que $|f(x) - \psi(x,k)| \leqslant \frac{\varepsilon}{2}$ pour tout x, la proposition en résulte.

Les très nombreuses méthodes de démonstration du théorème qui furent présentées par la suite peuvent se diviser en deux catégories (d'après E. BOREL [2]) les unes : de caractère "transcendant" (c'est-à-dire faisant appel à la théorie des fonctions analytiques et du genre de celle de WEIERSTRASS lui-même), les autres : de nature élémentaire.

Les démonstrations de PICARD [3], LERCH [4], VOLTERRA [5], font appel à la théorie des développements en série de FOURIER. Celle de VOLTERRA, par exemple, procède de la façon suivante : il est facile de montrer qu'étant donné f sur $[a \quad b]$, il existe une subdivision finie

$D = (a = x_0, \ldots\ldots\ldots, x_i, \ldots\ldots\ldots\ldots, x_n = b)$ de $[a \quad b]$

telle que si l'on considère la ligne polygonale L_D dont les côtés joignent les points $M_i\ (x_i\ ,\ f(x_i))$ et $M_{i+1}\ (x_{i+1},\ f(x_{i+1}))$, cette ligne brisée L_D puisse être considérée comme le graphe d'une fonction g, continue sur $[a \quad b]$, qui approche uniformément f et de sorte que

$|f(x) - g(x)| \leqslant \varepsilon$ pourtout $x \in [a \quad b]$, quelque soit ε .

Laquestion revient donc à l'approche de la fonction $g(x)$.

VOLTERRA prolonge la définition de $g(x)$ afin que dans $[a \quad c]$, $c > b$, $g(a) = g(c)$, et considère ce prolongement comme la partie du graphe d'une fonction $g_1(x)$ périodique et de période $(c-a)$. Or $g_1(x)$ est continue et n'admet qu'un nombre fini de maxima et minima dans une période ; un théorème classique de la théorie des séries de FOURIER montre que l'on peut écrire :

$$g_1(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{2n\pi}{T} x + b_n \sin \frac{2n\pi}{T} x \right) \quad , T = c - a$$

et la série trigonométrique converge uniformément dans $[a \quad c]$.

Il existe, ε étant donné, un rang N tel que dès que $p > N$ et pour tout $x \in [a \quad c]$

$$\left| g_1(x) - \left\{ a_0 + \sum_{n=1}^{n=p} \left(a_n \cos \frac{2\pi n}{T} x + b_n \sin \frac{2\pi n}{T} x \right) \right\} \right| \leqslant \frac{\varepsilon}{2}$$

Cela prouve déjà l'approximation uniforme de $f(x)$ par des polynômes trigonométriques. Pour passer aux polynômes ordinaires, il suffit de remarquer que $\cos \frac{2\pi}{T} nx$, $\sin \frac{2\pi}{T} nx$ sont égales à des séries entières qui convergent uniformément dans $[a \quad c]$; en ne prenant qu'un nombre fini de termes de ces séries,

On en conclut : il existe un polynôme $P(x)$ tel que

$$\left| \left\{ a_0 + \sum_{k=1}^{k=p} \left(a_n \cos \frac{2\pi}{T} nx + b_n \sin \frac{2\pi}{T} nx \right) \right\} - f(x) \right| \leqslant \frac{\varepsilon}{2} \qquad \text{quel que soit } x \in [a \quad b]$$

De là résulte la proposition.

Dans ces démonstrations de caractère transcendant, il faut signaler la démonstration d'HILBERT [6] qui varie peu de la démonstration de WEIERSTRASS,

mais a le grand mérite de donner explicitement un polynôme d'approximation.
En effet, HILBERT prouve que, si $0 < \alpha < a < b < \beta < 1$
(on peut toujours par un changement de variable supposer que [a b] tombe dans [0 1] strictement), les polynômes en x :

$$P_n(x) = \frac{\int_\alpha^\beta f(u)\left[1-(u-x)^2\right]^n}{\int_{-1}^{+1}(1-u^2)^n}$$

pour $n \to \infty$, convergent uniformément sur [a b] vers f(x).

Les démonstrations classées dans la catégorie "élémentaire" par E. BOREL, sont celles de RUGE [7], LEBESGUE [8], MITTAG-LEFFLER [9], BERNSTEIN [10]. Comme dans la démonstration de VOLTERRA, le problème est ramené à l'approximation uniforme sur [a b], de la fonction g, dont le graphe est une ligne brisée inscrite dans le graphe de f en [a b].

Posons $y_i = f(x_i)$.

Soit $\varphi_i(x) = y_{i-1} + \frac{x-x_{i-1}}{x_i - x_{i-1}}(y_i - y_{i-1})$;

Soit

$$(1) \qquad \varphi(x) = \varphi_1(x) + \sum_{i=1}^{i=n-1}\left[\varphi_{i+1}(x) - \varphi_i(x)\right]\alpha(x-x_i) \ ;$$

$\alpha(x)$ étant la fonction échelon : $\alpha \begin{cases} \alpha(x) = 1 & \text{pour} \quad x \geqslant 0 \\ \alpha(x) = 0 & \text{pour} \quad x < 0 \end{cases}$

D'après (1), il suffirait de trouver une approximation uniforme de l'échelon entre -(b-a) et +(b-a) par des polynômes. Or, cela est impossible, (l'oscillation de α autour de 0 est égale à 1), mais il est possible de trouver une fonction β telleque :

$$|\alpha(x) - \beta(x)| \leqslant \varepsilon'$$

pour tout $x \in [-(b-a) \quad -\eta] \cup [+\eta \quad (b-a)]$

quel que soit ε', η donnés à l'avance.

Si $\psi(x) = \varphi_1(x) + \sum_{i=1}^{i=n-1}\left[\varphi_{i+1}(x) - \varphi_i(x)\right]\beta(x-x_i)$, alors

$$|\varphi(x) - \psi(x)| \leqslant 2n M \varepsilon' \qquad (\text{si } |\varphi(x)| \leqslant M \text{ sur } [a \quad b])$$

pour x en dehors des intervalles $[x_i - \eta \quad x_i + \eta]$

Mais la fonction continue $\varphi_{i+1}(x) - \varphi_i(x)$ est nulle en x_i, on peut choisir η assez petit de sorte que dans $[x_i - \eta \quad x_i + \eta]$, $|\varphi_{i+1}(x) - \varphi_i(x)| \leqslant \varepsilon'$

et dans cet intervalle $|\varphi(x) - \psi(x)| \leq \varepsilon' M_1 + 2h M \varepsilon'$

si $|\alpha(x) - \beta(x)| \leq M_1$ pour tout $x \in [-(b-a) \quad +(b-a)]$, on peut donc dire qu'on peut approcher $\varphi(x)$ par $\psi(x)$ uniformément, donc $f(x)$ par $\psi(x)$.

RUNGE prend pour $\beta(x)$ la fraction rationnelle $1/1+(1-x)^{2k}$

Après avoir ramené (b-a) à être < 1 (par un changement de variables), il prouve que $\beta(x)$ répond à la question pour k assez grand et montre ainsi que l'on peut trouver une approximation de f sur [a b] par une fraction rationnelle. Il indique une méthode [7] pour passer d'une approximation par des fractions rationnelles à une approximation par des polynômes.

MITTAG-LEFFLER propose de prendre $\beta(x) = 1 - 2^{-(1-x)^k}$ (l'intervalle $[-(b-a) \quad +(b-a)]$ toujours ramené à être dans $[-1 \quad +1]$) pour k assez grand, c'est une fonction entière , qui à son tour sera approchée uniformément par un polynôme.

Enfin, LEBESGUE remarque que $\alpha(x)$ est telle que

$$x.[2\alpha(x)-1] = \left\{ \begin{array}{ll} x & \text{pour} \quad x \geq 0 \\ \\ -x & \text{pour} \quad x < \end{array} \right\} = +|x| = +(x^2)^{\frac{1}{2}}$$

si l'on remarque $|x| = (1+z)^{\frac{1}{2}} = 1 + \frac{1}{2}z - \frac{1.3}{2.4}z^2 + \ldots\ldots$

si $z = x^2 - 1$, la série converge pour $|z| < 1$, et même pour $z = \pm 1$, elle converge absolument et uniformément pour $-1 \leq z \leq 1$, donc il existe un polynôme, ε' étantdonné tel $|\,|x| - P(x)| \leq \varepsilon'$ pour tout $x \in [-1 \quad +1]$

Il est clair que pour 0 cela donne $|P(0)| \leq \varepsilon'$ donc : $|\,|x| - (P(x) - P(0))| \leq 2\varepsilon'$

On peut, si $\varepsilon = 2\varepsilon'$, trouver un polynôme sans terme constante tel que

$$|\,|x| - Q(x)| \leq \varepsilon$$

Dès lors, puisque $|x[2\alpha(x)-1] - Q(x)| \leq \varepsilon$

si $xQ'(x) = Q(x)$, $|2\alpha(x) - 1 - Q'(x)| \leq \frac{\varepsilon}{\eta}$ si $\eta \leq |x|$

On est bien arrivé à trouver, si η est donné, un polynôme : $\frac{1}{2}(1 + Q'(x)) = \beta_1(x)$, tel :

$|\alpha(x) - \beta_1(x)| \leq \frac{\varepsilon}{2\eta}$ qui amène l'approximation uniforme de $\alpha(x)$ dans $[-1 \quad -\eta] \cup [\eta \quad +1]$.

LEBESGUE ne présentait pas sa démonstration tout à fait de cette façon, mais elle s'y ramène. Nous verrons le parti qu'en a tiré STONE pour la démonstration de

son théorème.

Enfin, les généralisations diverses du théorème au cas, soit des sommes trigonométriques, soit des fonctions à plusieurs variables, furent obtenues par des procédés analogues par LEBESGUE [8], MONTEL [10], DE LA VALLEE POUSSIN [12], BERNSTEIN [10].

II. - GENERALITES SUR LA DEMONSTRATION DE STONE . - [13], [14]

Il y a trois remarques qui furent le départ de la démonstration de STONE et qui assurent la pleine généralité du théorème par lui prouvé :

- 1) la première est que le problème est de trouver l'approximation uniforme pour les fonctions continues (à valeurs réelles) dont le domaine de définition est un ensemble compact S .
- 2) la deuxième idée est que l'ensemble A des fonctions continues devant assurer cette approximation forment une algèbre de fonctions continues.
- 3) Etant donnés deux points x_1, x_2 distincts de S, quelconques, il est évident qu'il y a : 1°) des fonctions continues f sur S telles que $f(x_1) \neq f(x_2)$.

  2°), si l'approximation de f est réalisée, pour tout ε , une fonction $g \in A$ telle que $|f-g| < \varepsilon$ pour tout $x \in S$,
  donc $|f(x_1) - g(x_1)| < \varepsilon$ et $|f(x_2) - g(x_2)| < \varepsilon$
  et l'hypothèse $f(x_1) \neq f(x_2)$ implique que $g(x_1) \neq g(x_2)$
  pour un choix de ε .

Donc, les fonctions de A devront vérifier cette condition nécessaire : pour tout couple x_1, x_2 de points distincts de S, il existe au moins $g \in A$ telle que $g(x_1) \neq g(x_2)$, on dira que A sépare les points de S

Le théorème de STONE montre que ces conditions sont généralement suffisantes : en voici l'énoncé :

Théorème : Si S est un espace compact, A une algèbre de fonctions continues à valeurs réelles qui sépare les points de S, c'est-à-dire si x_1, x_2 sont distincts $(x_1 \neq x_2)$ et quelconques, il existe toujours une $f \in A$ telle que $f(x_1) \neq f(x_2)$. Alors :

La fermeture uniforme de $\overline{A}$ de A est :

- ou bien l'ensemble de toutes les fonctions continues sur S
- ou bien l'ensemble des fonctions continues qui s'annulent en un point déterminé de S.

Le plan de la démonstration est le suivant :

1°) on prouve que l'ensemble $\overline{A}$ est nécessairement réticulé c'est-à-dire si $f, g \in \overline{A}$, les fonctions $f \cup g = \text{Max}(f,g)$, $f \cap g = \text{Min}(f,g) \in \overline{A}$, par un lemme (I)
(Il est clair que si A est une algèbre de fonctions continues en S, $\overline{A}$ est aussi une algèbre de fonctions continues).

2°) On prouve ensuite que si un ensemble A_1 de fonctions continues en S est réticulé, et qu'une fonction f continue est telle qu'elle puisse être approchée en tout couple x_1, x_2 uniformément par une fonction de A_1 (quels que soient x_1, x_2, ε) il existe $g \in A_1$ telle

$$|f(x_1) - g(x_1)| < \varepsilon \quad , \quad |f(x_2) - g(x_2)| < \varepsilon$$

alors f peut être approchée uniformément dans tout S, c'est-à-dire $f \in \overline{A}_1$.

III. - DEMONSTRATIONS -

a) Un ensemble de fonctions f est une algèbre sur R :

1) si c'est un espace vectoriel :
   $f \pm g \in B$, $\lambda f \in B$ quels que soient $f, g \in B$ et $\lambda \in R$
2) si un produit f.g est défini distributif p.r.p. aux opérations sur l'espace vectoriel

On dira que cette algèbre est uniformément fermée si toute fonction adhérente à B y appartient (au sens de la topologie de la convergence uniforme)

Lemme I : toute algèbre uniformément fermée de fonctions bornées sur un ensemble S (pas forcément compact) est close pour les opérations de treillis Max. et Min., c'est-à-dire est réticulée.

On sait que
$$\begin{cases} \text{Max}(f,g) = \frac{1}{2}(f+g+|f-g|) \\ \text{Min}(f,g) = \frac{1}{2}(f+g-|f-g|) \end{cases}$$

je veux prouver, selon les hypothèses , que si $f,g \in A$: Max(f,g) et Min(f,g) $\in A$

D'après ce qui est écrit plus haut, il suffit de montrer (A est un espace vectoriel sur R) que si une fonction $f \in A$, la fonction $|f| \in A$ aussi.

Par hypothèse, toutes les fonctions de A sont bornées - je prends - $\|f\| = \underset{x \in S}{\text{Max}} |f(x)| \leq 1$

$\underset{x \in S}{\text{Max}} |f(x)| \leq 1$ (sinon je remplace f par λf où λ est une constante $\in R$)

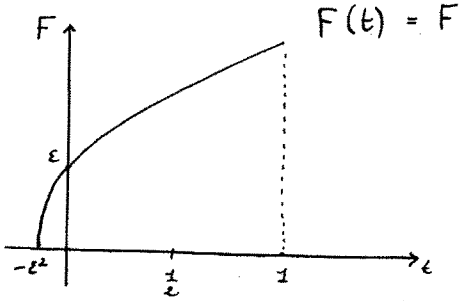
-Il suffit donc de voir si, $f \in A$ et $\|f\| \leq 1$ implique $|f| \in A$.

Puisque A est <u>uniformément fermée</u>, il suffit de montrer que $|f|$ peut être uniformément approchée pardes fonctions de A (hypothèses), c'est-à-dire que, quel que soit ε', il existe $g \in A$ telle $|g - |f|| < \varepsilon'$, pour tout $x \in S$.

-l'idée est d'étudier le même problème dans un cas particulier.

Considérons la fonction $F: t \longrightarrow F(t) = +\sqrt{t+\varepsilon^2} = (t+\varepsilon^2)^{\frac{1}{2}}$, $(t \in [-\varepsilon^2 \quad +\infty])$

et formons le développement de TAYLOR de F(t) autour de $\frac{1}{2}$



$$F(t) = F\left(\tfrac{1}{2}\right) + \left(t-\tfrac{1}{2}\right)F'\left(\tfrac{1}{2}\right) + \frac{1}{2!}\left(t-\tfrac{1}{2}\right)^2 F''\left(\tfrac{1}{2}\right) + \cdots$$

il est clair que ce développement est convergent et <u>uniformément convergent</u> pour t entre $0 \leq t \leq 1$

Si je pose $t = x^2$, je peux donc, étant donné ε trouver un polynôme $P(x^2)$ tel que

$$|f(x^2) - (x^2+\varepsilon^2)^{\frac{1}{2}}| < \varepsilon \quad , \forall x \in [-1 \quad +1]$$

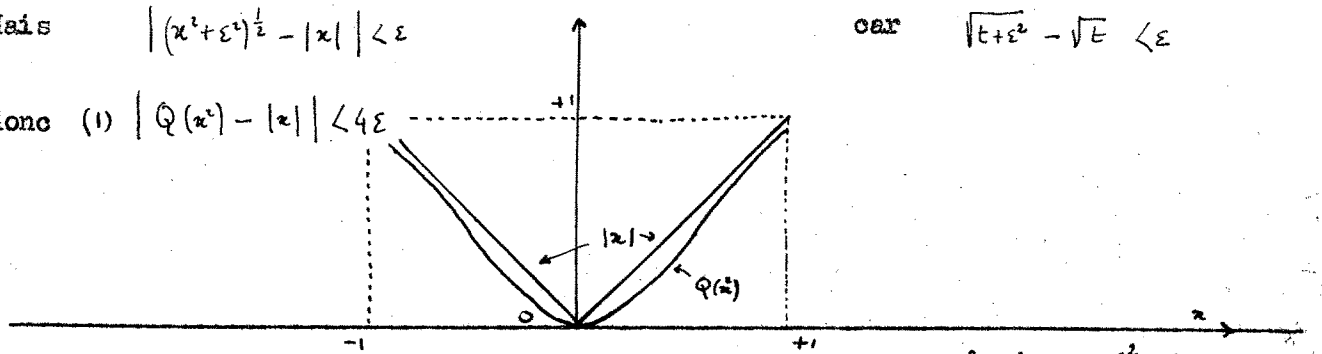
mais alors $|f(0) - \varepsilon| < \varepsilon$, $|f(0)| < 2\varepsilon$ si l'on pose $Q(x) = f(x) - f(0)$

(c'est-à-dire le polynôme P(x) <u>sauf</u> son terme constant)

$$|Q(x^2) - (x^2+\varepsilon^2)^{\frac{1}{2}}| \leq |f(x^2) - (x^2+\varepsilon^2)^{\frac{1}{2}}| + |f(0)| < 3\varepsilon$$

Mais $|(x^2+\varepsilon^2)^{\frac{1}{2}} - |x|| < \varepsilon$ car $\sqrt{t+\varepsilon^2} - \sqrt{t} < \varepsilon$

donc (1) $|Q(x^2) - |x|| < 4\varepsilon$



Si f est donné et puisque Q n'apas de terme constant $\alpha f \in A$, $\alpha_1 f^2 \in A$ ....

.... $Q(f^2) \in A$ et puisque $\|f\| \leq 1$, donc (1) implique :

$$|Q(f^2) - |f|| < \varepsilon' \quad \text{pour } \forall x \in S \quad , \varepsilon' = 4\varepsilon$$

Q.E.D.

b) Prouvons le lemme II . S est ici un compact et A n'a même pas besoin d'être une algèbre.

Mais supposons $A \subset \mathcal{C}(S,R)$ réticulé . ($\mathcal{C}(S,R)$: ensemble des applications continues de S dans R)

Lemme II : Si une fonction $f \in \mathcal{C}(S,R)$ a la propriété de pouvoir être approchée d'aussi près que l'on veut par une fonction de A, en deux points et quels que soient ces deux points p, q de S, alors $f \in \overline{A}$.
C'est à dire il existe une fonction de A qui l'approche uniformément en tout point.

Par hypothèse, quels que soient ε, et p, q, , $\exists\ f_{pq} \in A$ telle

(1) $\quad |f(p) - f_{pq}(p)| < \varepsilon \qquad$ et $\qquad |f(q) - f_{pq}(q)| < \varepsilon \qquad (1')$

Soit

$$U_{pq} = \{ x ;\ f_{pq}(x) < f(x) + \varepsilon \} = \{ x ;\ f_{pq}(x) - f(x) < \varepsilon \}$$

puisque $f_{pq} - f$ est continue et que l'ensemble des réels $< \varepsilon$ est un ouvert $\mathcal{O}_\varepsilon$ de R, U_{pq} est l'image réciproque de $\mathcal{O}_\varepsilon$: il est ouvert dans S.
Il contient p, q de part (1) et (1').
De même :

$$V_{pq} = \{ x ;\ f_{pq}(x) > f(x) - \varepsilon \}$$ est un ouvert de S qui contient p et q.

- Supposons q fixé, et faisons varier dans S, p. Les U_{pq} recouvrent tout S, il existe un recouvrement (propriété de compacité de S) fini de S par des U_{pq} et $p_1, \ldots\ldots\ldots, p_k$ les points qui permettent cela pour q fixé.

Si $x \in S$, il existe donc $U_{p_i q}$ tel $x \in U_{p_i q}$ c'est-à-dire tel $f_{p_i q}(x) < f(x) + \varepsilon$

soit $f_q(x) = \mathrm{Min}\left[f_{p_1 q}, f_{p_2 q}, \ldots, f_{p_k q} \right]$, on a $f_q \in A$ (hypothèse que A est réticulé)

et aussi $f_q(x) < f(x) + \varepsilon$ quel que soit $x \in S$.

De plus, soit $W_q = \bigcap_{p = p_1, \ldots, p_k} V_{pq}$ c'est un ouvert (int. finie d'ouverts) contient q (chaque V_{pq} contient q). Si $x \in W_q$ on a évidemment :

$$f_{pq}(x) > f(x) - \varepsilon$$ quel que soit p_i $(i = 1, \ldots\ldots, k)$, donc :

$$f_q(x) > f(x) - \varepsilon$$

En résumé, on vient de trouver, pour tout q et ε, un W_q et une fonction $f_q \in A$ tels :

1) quel que soit $x \in S$: $f_q(x) < f(x) + \varepsilon$
2) pour $x \in W_q$: $f_q(x) > f(x) - \varepsilon$
3) W_q est ouvert et contient q

Les W_q , pour q variant dans S, recouvrent S, donc il existe (propriété de compacité de S) un recouvrement fini de S par de tels W_q : $W_{q_1}, W_{q_2}, \ldots W_{q_m}$

Prenons ; $f_\varepsilon = \mathrm{Max}\left[f_{q_1}, f_{q_2} \ldots f_{q_m}\right]$, $f_\varepsilon \in A$

(hypothèse que A est réticulé)

Quel que soit x: 1) $f_\varepsilon(x) < f(x) + \varepsilon$.

2) pour x donné, il existe un indice j tel $x \in W_{q_j}$ et tel

$f_{q_j}(x) > f(x) - \varepsilon$ donc $f_\varepsilon(x) > f(x) - \varepsilon$

Par suite, on vient de trouver $f_\varepsilon \in A$ tel que $f(x) - \varepsilon < f_\varepsilon(x) < f(x) + \varepsilon$

ou

$$\left| f(x) - f_\varepsilon(x) \right| < \varepsilon \quad , \quad \forall x \in S$$

Q.E.D.

C) La démonstration des deux lemmes obtenus, passons à la démonstration du théorème de STONE lui-même.

— I°) Supposons que pour tout point $x \in S$ il existe une $f \in A$ telle que $f(x) \neq 0$ (c'est le cas si les fonctions constantes sur S, $\in A$).
Alors si $x_1 \neq x_2$, il existe un f tel $0 \neq f(x_1) \neq f(x_2) \neq 0$

En effet, il existe $f_1 \in A$ telle que : $f_1(x_1) \neq f_1(x_2)$,

(hypothèse). De plus, il existe $f_2 \in A$ telle : $f_2(x_2) \neq 0$

- Supposons, parexemple, que $f_1(x_1) \neq 0$ on aura : $0 \neq f_1(x_1) \neq f_1(x_2)$

si $f_1(x_2) = 0$, considérons la fonction $\alpha \cdot \dfrac{f_1(x)}{f_1(x_1)} + f_2(x) = \varphi(x)$

elle est de A,

(α nombre réel.)

Il est clair que je peux choisir α de sorte que : 1) $\varphi(x_1) \neq \varphi(x_2)$

2) $f_2(x_1) + \alpha \neq 0$

- Conclusion : Dans le cas où, pour tout x, il existe une $f \in A$ telle que $f(x) \neq 0$ et si A sépare S , il existe pour tout couple $x_1 \neq x_2$, une $f \in A$ telle

$$0 \neq f(x_1) \neq f(x_2) \neq 0$$

Posons : $f(x_1) = X_1$, $f(x_2) = X_2$; $0 \neq X_1 \neq X_2 \neq 0$

Cette remarque faite,

— a) soit $\overline{A}$ la fermeture uniforme de l'algèbre A, c'est une algèbre, et puisque ces fonctions sont à domaine de définition sur un compact S, chacune est bornée. Le lemme II prouve qu'elle est réticulée.

— b) D'autre part, soit φ une fonction continue quelconque, et posons :

$\varphi(x_1) = a$ et $\varphi(x_2) = b$ x_1, x_2 étant 2 points quelconques.

Soit f la fonction déterminée plus haut . Il est facile de déterminer un polynôme en X tel $\alpha X^2 + \beta X$ tel $\begin{cases} \alpha X_1^2 + \beta X_1 = a \\ \alpha X_2^2 + \beta X_2 = b \end{cases}$

(On prend : $\alpha = \dfrac{b X_1 - a X_2}{X_1 X_2 (X_2 - X_1)}$, $\beta = \dfrac{b X_1^2 - a X_2^2}{X_1 X_2 (X_2 - X_1)}$)

alors : $g = \alpha f^2 + \beta f \in A$ et pour x_1, x_2 prend les mêmes valeurs que φ .

Il est donc clair que φ fonction continue quelconque est "approchée" en chaque couple aussi près que l'on veut par des fonctions de A : le lemme I $\Rightarrow \varphi \in \bar{A}$ qui est donc identique à $\mathcal{C}(S, R)$

— II°) Supposons que pour un point particulier x_0 , toutes les $f \in A$ s'annulent. Je dois prouver que si g est une fonction qui est continue et s'annule pour $x = x_0$, $g \in \bar{A}$

Supposons augmenter l'ensemble A des fonctions constantes sur S. et considérons l'algèbre A_1 engendrée par A et ces fonctions.

L'ensemble A_1 ainsi obtenu est une algèbre qui sépare les points de S, et pour tout x, il existe bien une $f \in A$ telleque $f(x) \neq 0$

Donc, pour A_1, la première partie s'applique. Toute fonction de A_1 est de la forme $f + c$, ($f \in A$) et g étant donné ainsi que ε , il existe une $f+c$ telle que sur tout S , $| f + c - g | < \frac{\varepsilon}{2}$

donc en x_0 : $|c| < \frac{\varepsilon}{2}$

si bien

$$| f(x) - g(x) | < \varepsilon \quad , \forall x \in S$$

Q.E.D.

IV. - APPLICATIONS. -

Nous citerons pour terminer les applications du théorème de STONE, et , pour commencer, le théorème général de WEIERSTRASS qui en résulte directement.

I. Théorème de WEIERSTRASS - (approximation par des polynômes).

Si X est un ensemble fermé borné de R^n, toute fonction f continue sur X peut être approchée uniformément sur X par des polynômes en $(x_1, \ldots\ldots, x_n)$. Si X contient l'origine $(0, \ldots\ldots, 0)$ la fonction f peut être approchée uniformément par des polynômes s'annulant à l'origine si, et seulement si, f s'annule elle-même à l'origine.

II. Théorème d'approximation par des fonctions trigonométriques -

Soit f une fonction à valeurs réelles de la variable θ réelle et périodique, de période 2π et continue.

f peut être uniformément approchée pardes polynômes trigonométriques de la forme :

$$f(\theta) = \frac{a_0}{2} + \sum_{k=1}^{k=N} (a_k \cos k\theta + b_k \sin k\theta)$$

Le raisonnement est classique (déjà fait par HILBERT) : sur le cercle unité $(x_1^2 + x_2^2 = 1)$, si l'on pose :

$$F(x_1, x_2) = r \cdot f(\theta) \quad , \quad r = +\sqrt{x_1^2 + x_2^2}$$

$F(x_1, x_2)$ est une fonction continue sur ce cercle qui est fermé, borné, donc compact.

Il existe un polynôme $P(x_1, x_2)$ qui approche $F(x_1, x_2)$ uniformément sur le cercle uniformément.

Donc :

$$P(\cos\theta, \sin\theta) = \sum_{i,j \leq N} c_{ij} \cos^i\theta \sin^j\theta \quad , \text{ approche } f(\theta) \text{ uniformément.}$$

Il suffit de transformer les produits $\cos^i\theta$, $\sin^j\theta$ en ne faisant intervenir que des expressions de la forme $\cos p\theta$, $\sin q\theta$ pour obtenir le théorème.

Citons enfin des théorèmes d'approximations très analogues, obtenus en "rendant compact" R ou une partie de R.

III. Théorème d'approximation par des fonctions de LAGUERRE -

Toute fonction continue, à valeurs réelles, définie sur $0 \leq x < +\infty$ et qui est telle que $\lim_{x \to \infty} f(x) = 0$ peut être approchée uniformément par des fonctions de la forme : $e^{-\alpha x} \cdot p(x)$ $\left(p(x)\right.$ (polynôme en x)$\left.\right)$ (fonctions deLAGUERRE).

IV. - Théorème d'approximation par des fonctions d'HERMITE -

Toute fonction continue, à valeurs réelles, définie sur $-\infty < x < +\infty$ et telle que $\lim_{x \to \pm\infty} f(x) = 0$ peut être approchée uniformément par des fonctions de la forme :

$e^{-\alpha^2 x^2} \cdot P(x)$ (P(x) : polynôme en x) : fonctions d'HERMITE.

Enfin, indiquons, pour terminer, que STONE [15] a étendu son théorème dans des conditions un peu plus larges : fonctions à valeurs dans C, et définies sur des espaces localement compacts.

Il donne, commeexemples d'applications, en topologie générale :

- une démonstration du théorème de LEBESGUE-URYSOHN sur l'extension de la définition d'une fonction continue sur une partie fermée d'un espace topologique à tout cet espace.
- une démonstration du théorème de DIEUDONNE indiquant que si X est le produit d'une famille X_ι $(\iota \in I)$, d'espaces compacts, toute fonction continue à valeurs réelles sur X peut être approchée uniformément par des sommes finies de produits finis de fonctions d'une variable sur X.
