LEÇONS D'ANALYSE NUMÉRIQUE

DEUXIÈME PARTIE

J.F. DEBONGNIE

ISBN-13 : 978-2-9600313-6-2
Dépôt légal : D/2009/0480/35

METHODES DIRECTES DE RESOLUTION DES SYSTEMES LINEAIRES

1. Le problème qui nous intéresse dans ce chapitre consiste à résoudre le système matriciel

$$Ax = b$$

où A est une matrice, x et b, des vecteurs. Dans tout ce chapitre, nous nous tiendrons aux conventions de notation suivantes: les matrices sont notées par une lettre latine majuscule, les vecteurs par une minuscule, lorsqu'ils sont disposés verticalement (un vecteur disposé horizontalement sera noté avec le signe de transposition ^{T} , par exemple, x^T) ; pour les scalaires, nous utiliserons des lettres grecques.

Nous nous intéresserons aussi à l'inversion des matrices, ainsi qu'au calcul de leur déterminant. Une place particulière sera donnée aux matrices symétriques définies positives qui jouent un grand rôle dans de nombreuses branches de la physique.

2. Il convient d'abord de remarquer que la règle de CRAMER, si utile pour établir l'existence des solutions, est tout-à-fait impropre aux calculs numériques dès que la dimension de la matrice dépasse 3, en raison du grand nombre d'opérations qu'elle implique. Elle s'écrit en effet

$$x_i = \frac{dtm(c_{(1)}, \ldots, c_{(i-1)}, b, c_{(i+1)}, \ldots, c_{(n)})}{dtm\ A} ,$$

en notant $c_{(i)}$ les vecteurs-colonnes de la matrice A. Or, le calcul des déterminants par la règle des mineurs nécessite n! multiplications. En effet, un déterminant de dimension 2 en nécessite bien 2 = 2! ; si cette vrègle est vraie à l'ordre (n-1), un déterminant d'ordre n nécessite n produits par des déterminants d'ordre (n-1) tous différents, ce qui fait n(n-1)! = n! multiplications.

Voici le décompte complet des opérations significatives (*) du calcul:

(*) Sur ordinateur, les multiplications et divisions sont beaucoup plus lentes que les additions et soustractions. Aussi se limite-t-on souvent à ne dénombrer que les multiplications et divisions, que l'on appelle alors opérations significatives. Ceci vaut pour l'algèbre linéaire. Dans des calculs plus complexes, introduisant des calculs de fonctions transcendantes, ces dernières s'évaluent encore beaucoup plus lentement.

- Calcul de dtm A ..... n!
- Calcul des numérateurs : n déterminants..... n.n!
- Divisions ..... n

soit au total, $(n+1)\,n! + n = (n+1)! + n \approx (n+1)!$.

Or, faut-il le rappeler, la fonction factorielle croît très rapidement: ainsi, par exemple, si le système compte 10 équations, ce qui est loin d'être énorme, on obtient $11! = 3{,}99.10^7$ opérations significatives. A titre de comparaison, d'autres méthodes, que nous allons étudier, impliquent environ n^3, soit ici, 1000 opérations significatives: cela revient à dire 39900 fois moins de temps en machine. S'il faut une seconde pour réaliser les mille opérations de la méthode de Gauss, il faudra la bagatelle de 11 heures pour mener à bien celles de la méthode de Cramer!

3. Les méthodes de résolution des systèmes matriciels peuvent être divisées en deux grandes classes: les méthodes directes, dans lesquelles la solution s'obtient en un nombre fini d'opérations, et les méthodes itératives, qui consistent à construire une suite convergeant vers la solution du système. Signalons encore une troisième classe de méthodes, fondées sur des considérations statistiques, que l'on appelle méthodes de MONTE-CARLO.

Le présent chapitre est consacré aux méthodes directes. Les méthodes itératives seront étudiées à part. En ce qui concerne les méthodes de Monte Carlo, le lecteur pourra se référer aux ouvrages spécialisés [7,8] .

4. SYSTEMES A MATRICE DIAGONALE

Le cas le plus simple que l'on puisse imaginer est le système diagonal, naturellement scindé en n équations à une variable chacune:

$$\begin{cases} a_{11}\, x_1 = b_1 \\ \cdots\cdots\cdots \\ a_{nn}\, x_n = b_n \end{cases},$$

dont la solution s'écrit trivialement

$$x_1 = \frac{b_1}{a_{11}}\ ,\ \ldots\ ,\ x_n = \frac{b_n}{a_{nn}}\ ,$$

pour autant que ces rapports aient un sens. (n opérations significatives)

5. SYSTEMES A MATRICE TRIANGULAIRE SUPERIEURE

Ce sont des systèmes de la forme

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ & a_{22} & a_{23} & \cdots & a_{2n} \\ & & a_{33} & \cdots & a_{3n} \\ & 0 & & \ddots & \vdots \\ & & & & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \\ b_n \end{bmatrix}$$

Leur résolution est presque aussi simple que celle des systèmes diagonaux: il suffit de résoudre successivement les équations de la dernière à la première:

$$x_n = \frac{b_n}{a_{nn}}$$

$$x_{n-1} = \frac{1}{a_{n-1,n-1}} (b_{n-1} - a_{n-1,n}\, x_n) \qquad (x_n \text{ est déjà connu ! })$$

soit en général, pour $k \neq n$,

$$\boxed{x_k = \frac{1}{a_{kk}} \left(b_k - \sum_{i=k+1}^{n} a_{ki}\, x_i \right)}$$

C'est ce que l'on appelle l'algorithme de remontée.

Nombre d'opérations significatives

- Calcul de x_n : 1
- Calcul de x_k , $k \neq n$: (n-k) multiplications et une division, soit (n-k+1) O.S. (opérations significatives)

Total: $1 + \sum_{k=1}^{n-1} (n-k+1) = 1 + \frac{(n+2)(n-1)}{3} = \frac{n(n+1)}{2}$

Le nombre d'opérations croît donc sensiblement comme le carré de la dimension du système.

6. METHODE DE SIMPLE ELIMINATION DE GAUSS

6.1 -Cette méthode consiste à transformer un système plein en un système à matrice triangulaire (triangularisation), puis à résoudre ce dernier par l'algorithme de remontée.

Soit le système

$$\begin{cases} a_{11}\, x_1 + a_{12}\, x_2 + \ldots + a_{1n}\, x_n = b_1 \\ a_{21}\, x_1 + a_{22}\, x_2 + \ldots + a_{2n}\, x_n = b_2 \\ \ldots\ldots\ldots \\ a_{n1}\, x_1 + a_{n2}\, x_2 + \ldots + a_{nn}\, x_n = b_n \end{cases}$$

Le principe de la méthode consiste à éliminer x_1 de toutes les équations à partir de la deuxième, en leur soustrayant un multiple approprié de la première. Supposant $a_{11} \neq 0$, on passe, pour la ligne i, de

$$a_{i1}\, x_1 + a_{i2}\, x_2 + \ldots + a_{in}\, x_n = b_i$$

à

$$\underbrace{\left(a_{i1} - \frac{a_{i1}}{a_{11}}\, a_{11}\right)}_{=\,0}\, x_1 + \left(a_{i2} - \frac{a_{i1}}{a_{11}}\, a_{12}\right) x_2 + \ldots + \left(a_{in} - \frac{a_{i1}}{a_{11}}\, a_{1n}\right) x_n = \left(b_i - \frac{a_{i1}}{a_{11}}\, b_1\right) ,$$

c'est-à-dire que l'on arrive au système équivalent

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots\ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \ldots\ldots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & \ldots\ldots & a_{3n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & & a_{nn}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

avec

$$\begin{cases} a_{1j}^{(1)} = a_{1j} \quad , \quad j = 1, \ldots, n \qquad , \quad b_1^{(1)} = b_1 \\ a_{ij}^{(1)} = \left(a_{ij} - \dfrac{a_{i1}}{a_{11}}\, a_{1j}\right) , \begin{Bmatrix} i = 2, \ldots, n \\ j = 1, \ldots, n \end{Bmatrix} \quad , \quad b_i^{(1)} = b_i - \dfrac{a_{i1}}{a_{11}}\, b_1 \quad , \quad i = 2, \ldots, n \end{cases}$$

Cette opération nous a permis d'annuler tous les termes sous la diagonale de la première colonne. Comme nous n'avons fait que retrancher un multiple de la première ligne à chacune des autres, <u>le déterminant n'a pas varié au cours de la transformation</u>. On a donc

$$\mathrm{dtm}\ A = a_{11} \cdot \mathrm{dtm} \begin{bmatrix} a_{22}^{(1)} & \ldots\ldots & a_{2n}^{(1)} \\ \vdots & & \vdots \\ a_{n2}^{(1)} & & a_{nn}^{(1)} \end{bmatrix} .$$

Pour continuer, il nous suffira d'appliquer la même procédure au sous-système devenu indépendant

$$\begin{bmatrix} a_{22}^{(1)} & \ldots\ldots & a_{2n}^{(1)} \\ \vdots & & \vdots \\ a_{n2}^{(1)} & & a_{nn}^{(1)} \end{bmatrix} \begin{bmatrix} x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_2^{(1)} \\ \vdots \\ b_n^{(1)} \end{bmatrix} ,$$

ce qui donnera

$$\begin{bmatrix} a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_2^{(2)} \\ b_3^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix}$$

et ainsi de suite, ce qui mènera finalement au système triangulaire

$$\begin{bmatrix} a_{11}^{(n-1)} & a_{12}^{(n-1)} & \cdots & a_{1n}^{(n-1)} \\ & a_{22}^{(n-1)} & \cdots & a_{2n}^{(n-1)} \\ & 0 & \ddots & \vdots \\ & & & a_{nn}^{(n-1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(n-1)} \\ b_2^{(n-1)} \\ \vdots \\ b_n^{(n-1)} \end{bmatrix}$$

dont la résolution est aisée. Il est intéressant de noter que <u>le déterminant de la matrice n'a pas changé au cours des transformations</u>, si bien que

$$\text{dtm } A = a_{11}^{(n-1)} \cdots a_{nn}^{(n-1)} = a_{11} \cdot a_{22}^{(1)} \cdot \ldots \cdot a_{nn}^{(n-1)}$$
$$= \Pi_1 \cdot \Pi_2 \cdot \ldots \cdot \Pi_n ,$$

où l'on note

$$\Pi_i = a_{ii}^{(i-1)}$$

les nombres qui ont servi de diviseurs dans les éliminations et que l'on appelle <u>pivots</u>.

En résumé, les opérations à effectuer à l'étape i sont les suivantes:

- $\Pi_i = a_{ii}^{(i-1)}$
- Pour $j = i+1, \ldots, n$ et $k = i+1, \ldots, n$:

$$a_{ji}^{(i)} = 0$$

$$a_{jk}^{(i)} = a_{jk}^{(i-1)} - \frac{a_{ji}^{(i-1)}}{a_{ii}^{(i-1)}} a_{ik}^{(i-1)}$$

- Pour $j = i+1, \ldots, n$:

$$b_j^{(i)} = b_j^{(i-1)} - \frac{a_{ji}^{(i-1)}}{a_{ii}^{(i-1)}} b_i^{(i-1)}$$

6.2 - Pivotage non diagonal partiel

On objectera à juste titre que l'on pourrait, à une certaine étape i, tomber sur un pivot $a_{ii}^{(i-1)}$ nul. A ce stade, on obtiendrait donc la matrice

$$\begin{bmatrix} a_{11} & \cdots & \cdots & a_{1n}^{(i-1)} \\ & a_{22}^{(1)} & \cdots & a_{2n}^{(i-1)} \\ & & \ddots & \vdots \\ & & a_{i,i+1}^{(i-1)} & a_{in}^{(i-1)} \\ & 0 & \vdots & \vdots \\ & & a_{ni}^{(i-1)} & a_{n}^{(i-1)} \end{bmatrix}$$

dont le déterminant, égal à celui de A, vaut également

$\text{dtm } A = \text{dtn } (A^{(i-1)}) = \Pi_1 \Pi_2 \cdots \Pi_{i-1} \text{ dtm } B^{(i-1)}$, avec

$$B^{(i-1)} = \begin{bmatrix} a_{ii}^{(i-1)} & \cdots & a_{in}^{(i-1)} \\ \vdots & & \vdots \\ a_{ni}^{(i-1)} & \cdots & a_{nn}^{(i-1)} \end{bmatrix}$$

Le déterminant de $B^{(i-1)}$ ne peut donc être nul sans entraîner la nullité de dtm A . Supposant A inversible, un terme au moins de la première ligne de $B^{(i-1)}$ dot être différent de zéro. Soit ce terme non nul $a_{ik}^{(i-1)}$. Il suffit alors de considérer le système équivalent obtenu en permutant les colonnes i et k de la matrice et les variables x_i et x_k pour ramener ce terme $a_{ik}^{(i-1)}$ sur la diagonale:

$$\begin{bmatrix} a_{11}^{(i-1)} & \cdots & a_{1i}^{(i-1)} & \cdots & a_{1k}^{(i-1)} & \cdots & a_{1n}^{(i-1)} \\ & & a_{ii}^{(i-1)} & \cdots & a_{ik}^{(i-1)} & \cdots & a_{in}^{(i-1)} \\ & & a_{ni}^{(i-1)} & \cdots & a_{nk}^{(i-1)} & \cdots & a_{nn}^{(i-1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_i \\ x_k \\ x_n \end{bmatrix} =$$

$$\begin{bmatrix} a_{11}^{(i-1)} & \cdots & a_{1k}^{(i-1)} & \cdots & a_{1i}^{(i-1)} & \cdots & a_{1n}^{(i-1)} \\ & & a_{ik}^{(i-1)} & \cdots & a_{ii}^{(i-1)} & \cdots & a_{in}^{(i-1)} \\ & & a_{nk}^{(i-1)} & \cdots & a_{ni}^{(i-1)} & \cdots & a_{nn}^{(i-1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_k \\ x_i \\ x_n \end{bmatrix}$$

Ceci permet de continuer les calculs. Mais après résolution, il faudra

permuter x_i et x_k pour obtenir le vrai vecteur-solution.

Bien entendu, il se peut que l'on soit obligé d'effectuer plusieurs p permutations de colonnes. Alors, on devra faire subir, après résolution, les mêmes permutations au vexteur x, mais dans l'ordre inverse, c'est-à-dire en commençant par la dernière pour terminer par la première, de manière à revenir pas à pas au vecteur originel.

Pour organiser ces permutations, il convient de réserver en mémoire un vecteur, que nous noterons ICOL, à n composantes, de telle façon que ICOL(i) soit le numéro de la colonne où l'on a trouvé le pivot sur la ligne i. Il y a donc eu permutation chaque fois que $ICOL(i) \neq i$. On notera que l'on a toujours $ICOL(i) \geqslant i$, et $ICOL(n) = n$. Pour revenir au vecteur x dans le bon ordre, il suffit de balayer le vecteur ICOL de la composante n à la composante 1 et, chaque fois que $ICOL(i) \neq i$, il faut permuter x_i et $x_{ICOL(i)}$.

Enfin, le déterminant de A vaut

$$\text{dtm } A = \Pi_1 \Pi_2 \ldots\ldots \Pi_n \cdot (-1)^p \quad ,$$

où p est le nombre de permutations effectuées ou, ce qui revient au même, le nombre de composantes ICOL(i) différentes de i.

6.3 - Stratégie de recherche des pivots

Au lieu de se limiter à chercher un pivot non-diagonal si le terme diagonal est nul, il est préférable de choisir systématiquement (à droite de la diagonale, celle-ci comprise) le plus grand terme de la ligne. Le test éventuel de nullité sera fait à meilleur escient et, en outre, on soustraira des lignes suivantes un plus petit multiple de la ligne du pivot, ce qui offre un double avantage:

- Les erreurs d'arrondi seront moindres
- L'indépendance des lignes sera mieux préservée.

6.4 - Nombre d'opérations significatives

Pour diminuer le nombre d'opérations, il est judicieux de transformer l'algorithme comme suit: à l'étape i, après permutation des colonnes, on calcule, dans l'ordre:

$$\begin{cases} \bullet\ \Pi_i = a_{ii}^{(i-1)} \\ \bullet\ a_{ii}^{(i)} = 1 \\ \bullet\ \begin{cases} a_{ij}^{(i)} = \dfrac{a_{ij}^{(i-1)}}{\Pi_i} \quad , \quad j = i+1, \ldots, n & (n-1)\ \text{O.S.} \\ b_i^{(i)} = \dfrac{b_i^{(i-1)}}{\Pi_i} & 1\ \text{O.S.} \end{cases} \\ \bullet\ a_{kl}^{(i)} = a_{kl}^{(i-1)} - a_{ki}^{(i-1)}\ a_{il}^{(i)} \quad , \quad \begin{cases} k = i+1, \ldots, n \\ l = i+1, \ldots, n \end{cases} & (n-i)^2\ \text{O.S.} \\ \bullet\ b_k^{(i)} = b_k^{(i-1)} - a_{ki}^{(i-1)}\ b_i^{(i)} \quad , \quad k = i+1, \ldots, n & (n-i)\ \text{O.S.} \\ \bullet\ a_{ki}^{(i)} = 0 \qquad k = i+1, \ldots, n \end{cases}$$

Cette procédure permet de travailler dans la matrice même. La seule différence par rapport à ce qui a été exposé plus haut consiste en la division de la ligne i par le pivot. On obtient ainsi, à l'étape i,

- pour la matrice: $(n-i)^2 + (n-i)$ O.S.
- pour le second membre: $(n-i+1)$ O.S.

La triangularisation de la matrice demande donc

$$\sum_{i=1}^{n-1} (n-i)^2 + \sum_{i=1}^{n-1} (n-i) = \sum_{j=1}^{n-1} j^2 + \sum_{j=1}^{n-1} j =$$

$$= \frac{(n-1)n(2n-1)}{6} + \frac{n(n-1)}{2} \approx \frac{n^3}{3} \quad \text{O.S.}$$

Celle du second membre,

$$\sum_{i=1}^{n-1} (n-i+1) = \sum_{j=2}^{n} j = \frac{(n-1)(n+2)}{2} \approx \frac{n^2}{2} \quad \text{O.S.} \ ,$$

à quoi il faut ajouter une remontée, soit encore $n^2/2$ O.S.
Pour trouver simultanément les solutions correspondant à p seconds membres, il faut donc environ

$$\frac{n^3}{2} + p\ n^2 \quad \text{O.S.}$$

6.5 - Calcul du déterminant

Il convient de prendre garde au fait que le déterminant, produit des pivots,

peut être très grand ou très petit. Ainsi, si tous les pivots sont de l'ordre de 10, dans une matrice 100 x 100, le déterminant sera de l'ordre de 10^{100}. Il est donc prudent de calculer, plutôt que le déterminant lui-même, les grandeurs

$$\log |\mathrm{dtm}\ A| = \sum_i \log |\Pi_i|$$

$$\mathrm{sign}\ |\mathrm{dtm}\ A| = \prod_i \mathrm{sign}\ \Pi_i \cdot (-1)^p \quad ,$$

où p est le nombre de permutations de colonnes effectuées.

6.6 - Test de singularité

En principe, la singularité se manifeste par la nullité de tous les termes $a_{ij}^{(i-1)}$, $j = 1, \ldots, n$ à une étape donnée. En réalité, les choses sont moins simples, car on a affaire à des zéros calculés et donc entachés d'erreurs d'arrondis. Il faut donc effectuer un test du type

$$|\ \text{pivot}\ | \leq \varepsilon \quad , \qquad \varepsilon \text{ petit.}$$

Mais on court un double risque:

- Si ε est trop grand, on croira avoir affaire à un matrice singulière, alors qu'elle est inversible.

- Si ε est trop petit, on risque de ne pas détecter une véritable singularité. L'effet est en général désastreux, car à un certain moment, tous les termes $a_{ij}^{(i-1)}$ sont des faux zéros, sans signification. On écrira alors

$$a_{kl}^{(i)} = a_{kl}^{(i-1)} - a_{ki}^{(i-1)} \frac{a_{il}^{(i-1)}}{a_{ii}^{(i-1)}} \quad ,$$

et le quotient du second membre prendra des valeurs quasi-aléatoires.

Lorsque les termes de la matrice A sont d'un ordre de grandeur comparable, il est raisonnable de poser, par exemple,

$$\varepsilon = \frac{\frac{1}{n^2} \sum_{ij} |\ a_{ij}|}{2^{2p/3}} \quad ,$$

p étant le nombre de chiffres binaires de la représentation des nombres en machine.

Le cas des matrices aux termes disproportionnés est plus délicat. Bien que ces matrices soient en général mal conditionnées, on peut souvent améliorer la situation en posant le test

$$\Pi_i \leq 2^{-2p/3} \cdot \sup_j |\ a_{ij}| \quad ,$$

mais ceci suppose la mémorisation des termes maximaux de chaque ligne

avant triangularisation.

Il faut d'ailleurs noter que dans de telles matrices, le choix du pivot est perturbé par la disproportion des termes de la matrice. Une manière de porter remède à cette situation consiste à effectuer une mise à échelle des colonnes : étant donné la matrice écrite sous la forme de vecteurs-colonnes,

$$A = \left[c_{(1)}, \ldots\ldots, c_{(n)} \right] ,$$

le système à résoudre est

$$c_{(1)}\, x_1 + \ldots\ldots + c_{(n)}\, x_n = b \quad .$$

En effectuant le changement de variables

$$x_1 = \|c_{(1)}\|\, y_1 \ , \ \ldots\ldots \ , \ x_n = \|c_{(n)}\|\, y_n \ ,$$

on obtient le système

$$\hat{A}\, y = b \ ,$$

avec

$$\hat{A} = \left[\frac{c_{(1)}}{\|c_{(1)}\|} \ , \ \ldots \ , \ \frac{c_{(n)}}{\|c_{(n)}\|} \right] \ ,$$

matrice dont les colonnes ont toutes la norme 1.

7. METHODE D'ELIMINATION COMPLETE DE GAUSS-JORDAN

7.1- L'inversion d'une matrice peut être présentée comme suit: partant du système

$$A \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \ ,$$

on désire trouver la matrice A^{-1} telle que

$$A^{-1} \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad .$$

Proposons-nous de travailler par étapes, en essayant de construire successivement les matrices $A^{(1)}, \ldots, A^{(n)}$ telles que

$$A^{(1)} \begin{bmatrix} y_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \ , \quad A^{(2)} \begin{bmatrix} y_1 \\ y_2 \\ x_3 \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ y_3 \\ \cdot \\ y_n \end{bmatrix} \quad , \ \ldots \ , \ A^{(n)} \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

La matrice $A^{(n)}$ sera alors A^{-1} . Nous écrirons également $A^{(0)}$ pour A.
Le problème fondamental qui se pose est donc, en termes imagés, <u>d'intervertir</u> x_i <u>et</u> y_i <u>dans le système</u>.

Soit donc le système

$$B \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} ,$$

où l'on veut intervertir x_i et y_i . La ligne i de ce système s'écrit

$$b_{ii}\, x_i + \sum_{k \neq i} b_{ik}\, x_k = y_i \quad ,$$

dont on tire, si $b_{ii} \neq 0$,

$$\frac{1}{b_{ii}}\, y_i - \sum_{k \neq i} \frac{b_{ik}}{b_{ii}}\, x_k = x_i \ .$$

Le problème est donc déjà résolu pur la ligne i. Pour les autres, on a

$$b_{ji}\, x_i + \sum_{k \neq i} b_{jk}\, x_k = y_k$$

et, en réintroduisant la valeur trouvée pour x_i ,

$$\frac{b_{ji}}{b_{ii}}\, y_i - \sum_{k \neq i} b_{ji}\, \frac{b_{ik}}{b_{ii}}\, x_k + \sum_{k \neq i} b_{jk}\, x_k = y_k \ ,$$

soit

$$\frac{b_{ji}}{b_{ii}}\, y_i + \sum_{k \neq i} \left(b_{jk} - \frac{b_{ji}\, b_{ik}}{b_{ii}} \right) x_k = y_k \ .$$

Dans le cas qui nous intéresse, le passage de $A^{(i-1)}$ à $A^{(i)}$ se fait donc par les opérations suivantes:

<u>Terme ii</u> : $a_{ii}^{(i)} = \dfrac{1}{a_{ii}^{(i-1)}}$ 1 O.S.

<u>Ligne i, sauf</u> a_{ii} :

$$a_{ij}^{(i)} = - \frac{a_{ij}^{(i-1)}}{a_{ii}^{(i-1)}} = - a_{ij}^{(i-1)}\, a_{ii}^{(i)} \qquad (n-1) \text{ O.S.}$$

<u>Termes non situés sur la ligne ou la colonne i</u> :

$$a_{jk}^{(i)} = a_{jk}^{(i-1)} - \frac{a_{ji}^{(i-1)}\, a_{ik}^{(i-1)}}{a_{ii}^{(i-1)}} = a_{jk}^{(i-1)} + a_{ji}^{(i-1)}\, a_{ik}^{(i)} \qquad (n-1)^2 \text{ O.S.}$$

Colonne i, sauf a_{ii} :

$$a_{ki}^{(i)} = \frac{a_{ki}^{(i-1)}}{a_{ii}^{(i-1)}} = a_{ki}^{(i-1)} \; a_{ii}^{(i)} \qquad (n-1) \text{ O.S.}$$

Dans cet ordre, les opérations peuvent être effectuées dans la matrice elle-même, sans détruire les termes dont on a encore besoin. Une procédure de ce type s'appelle pivotage, et le terme $a_{ii}^{(i-1)}$ est appelé pivot. Un pivotage nécessite n^2 O.S.. L'inversion se réalise donc en n^3 O.S..

7.2 - Il est à remarquer que tous les termes situés en-dessous et à droite du pivot subissent la même transformation que dans la triangularisation de Gauss. Aussi les pivots sont-ils égaux dans les deux méthodes, et on a encore

$$\text{dtm } A = \Pi_1 . \; \ldots \; \Pi_n \; .$$

De la même façon, on peut, et c'est fortement conseillé, utiliser une stratégie de recherche du pivot maximun sur la ligne (à droite de la diagonale). Les permutations faites sur les colonnes de A se retrouveront dans le vecteur x à la fin des calculs, c'est-à-dire que l'on obtiendra

$$x_{perm} = A_{perm}^{-1} \; y.$$

Pour revenir à la vraie matrice A^{-1}, il faut donc faire subir à A_{perm}^{-1} la transformation qui ramène les éléments du vecteur x à leur place, c'est-à-dire permuter ses lignes: il faut donc, en fin de calcul, refaire sur les lignes de A^{-1} toutes les permutations de colonnes, et en sens inverse. Ici aussi, le déterminant de la matrice A est donné par

$$\text{dtm } A = \Pi_1 \; \ldots . \; \Pi_n \; (-1)^p \quad ,$$

p étant le nombre de permutations effectuées. Tout ce qui a été dit sur les tests de singularité pour la triangularisation s'applique encore ici.

8. PIVOTAGE NON DIAGONAL COMPLET

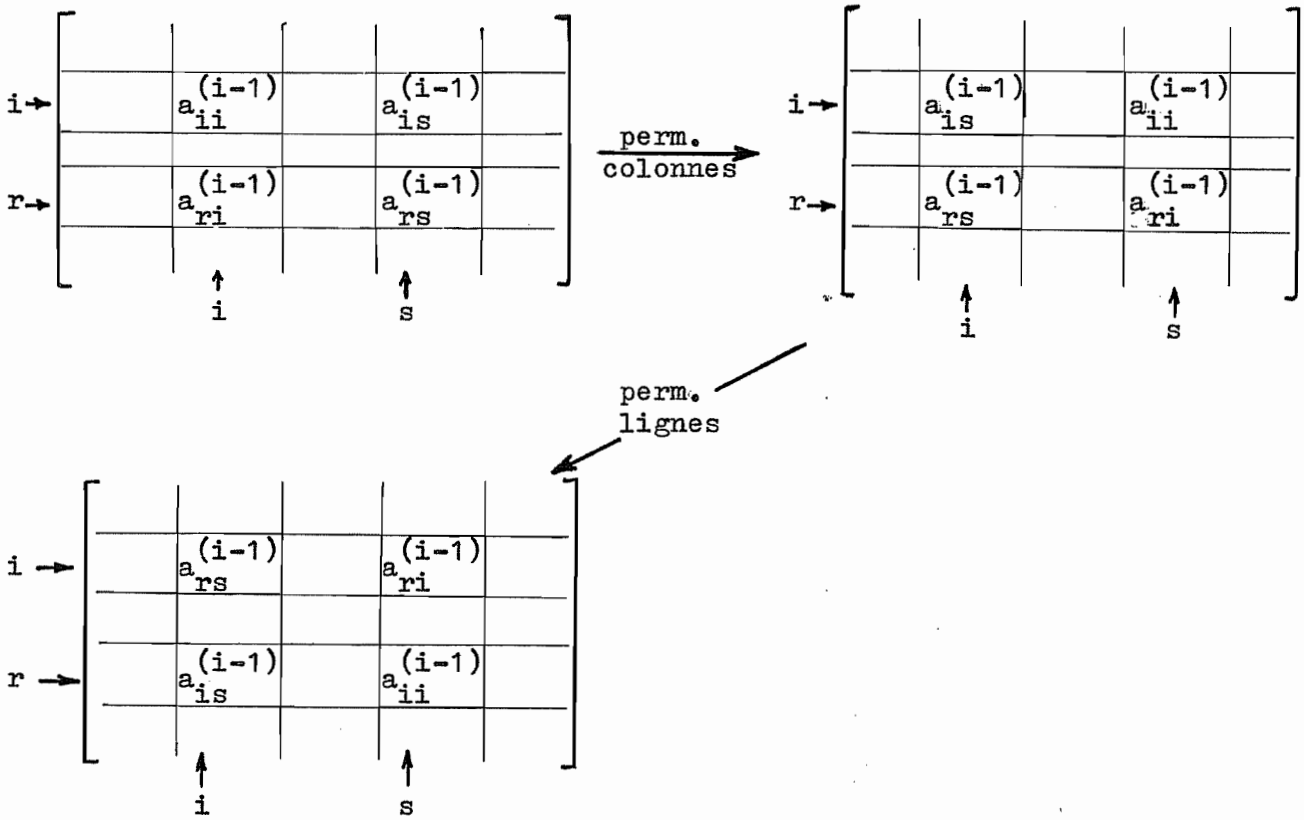
A l'étape i, on peut également, au lieu de chercher le plus grand terme en valeur absolue parmi les $a_{ij}^{(i-1)}$, j = i, ..., n , chercher le plus grand des $|a_{kl}^{(i-1)}|$, k = i, ..., n , l = i, ... , n , c'est-à-dire dans toute la sous-matrice $B^{(i-1)}$ représentée ci-dessous:

$$A^{(i-1)} = \left[\begin{array}{c|c} * & * \\ \hline * & B^{(i-1)} \end{array}\right] \leftarrow i$$

$$\uparrow$$

$$i$$

Soit $a_{rs}^{(i-1)}$ ce terme. On permuta alors la ligne i avec la ligne r et la colonne i avec la colonne s, de manière à amener le terme $a_{rs}^{(i-1)}$ en position (i,i) :



Ce faisant, on transforme le système

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{i-1} \\ x_i \\ \vdots \\ x_r \\ \vdots \\ x_n \end{bmatrix} = A^{(i-1)} \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ y_i \\ \vdots \\ y_s \\ \vdots \\ y_n \end{bmatrix} \quad \text{en} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_{i-1} \\ x_r \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = A_{perm}^{(i-1)} \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ y_s \\ \vdots \\ y_i \\ \vdots \\ x_n \end{bmatrix}$$

Après inversion, si c'est la seule permutation, on obtiendra donc

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_r \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = A_{perm}^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_s \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

Pour revenir à la véritable inverse, il faudra donc permuter les lignes i et s et les colonnes i et r de A^{-1}_{perm} .

Dans le cas général, il faut mémoriser dans deux vecteurs, ICOL et ILIGN, les permutations de lignes et de colonnes: si le pivot n° i est la composante $a^{(i-1)}_{rs}$, on aura ICOL(i) = s et ILIGN(i) = r. Après inversion, il faut balayer ces deux vecteurs dans le sens inverse et effectuer, dans cet ordre, sur les colonnes, les permutations indiquées par ILIGN et sur les lignes, les permutations indiquées par ICOL.

Le déterminant vaut alors

$$\text{dtm } A = \Pi_1 \ldots \Pi_n \, (-1)^p \, (-1)^q \ ,$$

où q est le nombre de permutations de lignes, et p, le nombre de permutations de colonnes.

Dans cette méthode, la singularité de la matrice se détecte par une matrice $B^{(i-1)}$ entièrement peuplée de zéros.

9. METHODE DE GAUSS-JORDAN PAR BLOCS

Au lieu d'intervertir x_1 et y_1, x_2 et y_2, etc... , on peut également en intervertir plusieurs à la fois. Ecrivons le système sous forme partitionnée:

$$\begin{bmatrix} A_{PP} & A_{PR} \\ A_{RP} & A_{RR} \end{bmatrix} \begin{bmatrix} x_P \\ x_R \end{bmatrix} = \begin{bmatrix} y_P \\ y_R \end{bmatrix}$$

avec A_{PP} et A_{RR} carrées, de dimensions n_P et n_R , $n = n_P + n_R$.

On en tire aisément

$$A_{PP} \, x_P + A_{PR} \, x_R = y_P \ ,$$

soit, si A_{PP} est inversible,

$$x_P = A^{-1}_{PP} \, y_P - A^{-1}_{PP} \, A_{PR} \, x_R \ .$$

En réintroduisant cet résultat dans la seconde équation,

$$A_{RP} \, x_P + A_{RR} \, x_R = y_R \ ,$$

on obtient

$$A_{RP} \, A^{-1}_{PP} \, y_P + (A_{RR} - A_{RP} \, A^{-1}_{PP} \, A_{PR}) \, x_R = y_R \ ,$$

ce qui mène à

$$B \begin{bmatrix} y_P \\ x_R \end{bmatrix} = \begin{bmatrix} x_P \\ y_R \end{bmatrix} \ ,$$

avec

$$B = \left[\begin{array}{c|c} A_{PP}^{-1} & -A_{PP}^{-1} A_{PR} \\ \hline A_{RP} A_{PP}^{-1} & A_{RR} - A_{RP} A_{PP}^{-1} A_{PR} \end{array}\right] = \begin{bmatrix} B_{PP} & B_{PR} \\ B_{RP} & B_{RR} \end{bmatrix}$$

Il va sans dire que cette matrice ne diffère en rien de celle que l'on aurait obtenue en intervertissant tour à tour x_1 et y_1, ..., x_{n_P} et y_{n_P}, c'est-à-dire que l'on a

$$B = A^{(n_P)} .$$

La condition $a_{ii}^{(i-1)} \neq 0$ est ici remplacée par dtm $A_{pp} \neq 0$.

A ce stade, notons que de la décomposition évidente

$$\begin{bmatrix} A_{PP} & A_{PR} \\ A_{RP} & A_{RR} \end{bmatrix} = \begin{bmatrix} A_{PP} & 0 \\ A_{RP} & I \end{bmatrix} \begin{bmatrix} I & A_{PP}^{-1} A_{PR} \\ 0 & A_{RR} - A_{RP} A_{PP}^{-1} A_{PR} \end{bmatrix} ,$$

on déduit immédiatement

$$\text{dtm } A = \text{dtm} \begin{bmatrix} A_{PP} & 0 \\ A_{RP} & I \end{bmatrix} . \text{ dtm} \begin{bmatrix} I & A_{PP}^{-1} A_{PR} \\ 0 & A_{RR} - A_{RP} A_{PP}^{-1} A_{PR} \end{bmatrix} ;$$

Or,

$$\text{dtm} \begin{bmatrix} A_{PP} & 0 \\ A_{RP} & I \end{bmatrix} =$$

$$\text{dtm} \left[\begin{array}{c|ccc} A_{PP} & 0 & 0 & 0 \\ \hline * & \begin{matrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ & & & 0 \end{matrix} & \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix} & \begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix} \\ * & 0 \cdots 0 & 1 & 0 \\ * & 0 \cdots 0 & 1 & 0 \\ * & 0 \cdots 0 & 0 & 1 \end{array}\right] = \text{dtm } A_{PP} ,$$

par la règle des mineurs. De même,

$$\text{dtm} \begin{bmatrix} I & A_{PP}^{-1} A_{PR} \\ 0 & A_{RR} - A_{RP} A_{PP}^{-1} A_{PR} \end{bmatrix} = \text{dtm } (A_{RR} - A_{RP} A_{PP}^{-1} A_{PR})$$

on obtient donc la formule

$$\text{dtm } A = \text{dtm } A_{PP} . \text{ dtm } (A_{RR} - A_{RP} A_{PP}^{-1} A_{PR}) .$$

C'est la <u>règle de FROBENIUS-SCHUR</u>, qui vaut chaque fois que dtm $A_{PP} \neq 0$. Dans le cas où la matrice A_{PP} peut être inversée par pivotage diagonal,

elle se ramène à

$$\text{dtm } A = \Pi_1 \ldots \Pi_n \text{ dtm } B_{RR} \ ,$$

formule que nous avons obtenue dans la section 6.2.

* 10. RECHERCHE DES SINGULARITES D'UNE MATRICE RECTANGULAIRE

Rappelons que le rang d'une matrice (rectangulaire) A est la dimension de la plus grande matrice carrée non singulière que l'on peut en extraire par suppression de lignes et de colonnes. La matrice carrée en question (ou l'une de ses réalisations) sera appelée noyau dur de A.

Supposons qu'après certaines permutations de lignes et de colonnes et r pivotages de Gauss-Jordan, on obtienne la matrice (les accents circonflexes rappellent le fait que des permutations ont été effectuées)

$$\hat{B} = \begin{matrix} r \\ m \end{matrix} \overset{\overset{r}{\longleftrightarrow}}{\overset{\updownarrow}{\left[\begin{array}{c|c} \hat{B}_{PP} & \hat{B}_{PR} \\ \hline \hat{B}_{RP} & \hat{B}_{RR} \end{array}\right]}} \overset{\updownarrow}{m} \\ \overset{\longleftrightarrow}{n} = \left[\begin{array}{c|c} \hat{A}_{PP}^{-1} & -\hat{A}_{PP}^{-1}\hat{A}_{PR} \\ \hline \hat{A}_{RP}\hat{A}_{PP}^{-1} & \hat{A}_{RR} - \hat{A}_{RP}\hat{A}_{PP}^{-1}\hat{A}_{PR} \end{array}\right]$$

Les permutations ont, certes, été effectuées en cours de route. Mais il est aisé de se rendre compte qu'il eût été équivalent de les faire avant de commencer.

Montrons que la condition nécessaire et suffisante pour que $\hat{A}_{PP}$ soit le noyau dur de $\hat{A}$ est que $\hat{B}_{RR}$ soit nulle.

Cette condition est nécessaire . En effet, dans la matrice

$$\hat{A} = \left[\begin{array}{cc} \hat{A}_{PP} & \hat{A}_{PR} \\ \hat{A}_{RP} & \hat{A}_{RR} \end{array}\right] \ ,$$

les (m-r) dernières lignes doivent être combinaisons linéaires des r premières. En notant $l_{(i)}^T$ la i^e ligne, ceci s'écrit

$$\begin{bmatrix} \hat{l}_{(r+1)}^T \\ \cdot \\ \cdot \\ \hat{l}_{(m)}^T \end{bmatrix} = \begin{bmatrix} \lambda_{r+1,1}\,\hat{l}_{(1)}^T + \ldots + \lambda_{r+1,r}\,\hat{l}_{(r)}^T \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_{m,1}\,\hat{l}_{(1)}^T + \ldots + \lambda_{m,r}\,\hat{l}_{(r)}^T \end{bmatrix} = \underbrace{\begin{bmatrix} \lambda_{r+1,1} \ldots \lambda_{r+1,r} \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \lambda_{m,1} \ldots \lambda_{m,r} \end{bmatrix}}_{R} \begin{bmatrix} \hat{l}_{(1)}^T \\ \\ \\ \hat{l}_{(r)}^T \end{bmatrix}$$

soit

$$\left[\begin{array}{cc} \hat{A}_{RP} & \hat{A}_{RR} \end{array}\right] = R \left[\begin{array}{cc} \hat{A}_{PP} & \hat{A}_{PR} \end{array}\right] \ ,$$

relation qui se scinde en

$$\hat{A}_{RP} = R\,\hat{A}_{PP} \quad , \quad \hat{A}_{RR} = R\,\hat{A}_{PR} \quad :$$

Il en résulte

$$\hat{B}_{RR} = \hat{A}_{RR} - \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR} = R\,\hat{A}_{PR} - R\,\hat{A}_{PP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR} = R\,(\hat{A}_{PR} - \hat{A}_{PR}) = 0 \ .$$

La <u>suffisance</u> de cette condition est évidente, car si

$$\hat{B}_{RR} = \hat{A}_{RR} - \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR} = 0 \ ,$$

on a évidemment

$$\hat{A}_{RR} = \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR}$$

et, de toute façon,

$$\hat{A}_{RP} = \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PP} \ ,$$

donc

$$\hat{A}_{RP}\,\hat{A}_{RR} = \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PP}\,\hat{A}_{PR} = R\,\hat{A}_{PP}\,\hat{A}_{PR} \ ,$$

ce qui exprime une relarion linéaire entre les (m-r) dernières lignes et les r premières.

Bien plus, il est possible de déterminer à ce stade les <u>singularités de la matrice $\hat{A}$</u> (solutions de $\hat{A}\,x = 0$), <u>qui ne sont autres que les colonnes de la matrice</u>

$$\begin{bmatrix} \hat{B}_{PR} \\ I \end{bmatrix} = \begin{bmatrix} -\hat{A}_{pp}^{-1}\,\hat{A}_{PR} \\ I \end{bmatrix} \ .$$

Pour démontrer cette assertion, notons d'abord que ces colonnes sont linéairement indépendantes, grâce à la présence de la sous-matrice I. Dès lors, si l'on peut montrer que le produit de $\hat{A}$ par cette matrice est bien nul, ce sont <u>les seules</u> singularités indépendantes, car leur nombre (n - rang(A)) est précisément le nombre de singularités indépendantes de $\hat{A}$.

De fait,

$$\hat{A}\begin{bmatrix} -\hat{A}_{PP}^{-1}\,\hat{A}_{PR} \\ I \end{bmatrix} = \begin{bmatrix} \hat{A}_{PP} & \hat{A}_{PR} \\ \hat{A}_{RP} & \hat{A}_{RR} \end{bmatrix}\begin{bmatrix} -\hat{A}_{PP}^{-1}\,\hat{A}_{PR} \\ I \end{bmatrix} =$$

$$\begin{bmatrix} -\hat{A}_{PP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR} + \hat{A}_{PR} \\ -\hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR} + \hat{A}_{RR} \end{bmatrix} = \begin{bmatrix} -\hat{A}_{PR} + \hat{A}_{PR} \\ \hat{B}_{RR} \end{bmatrix} = 0 \ .$$

Pour repasser aux singularités de A, notons que

$$\hat{A}\,\hat{x} = 0 \iff Ax = 0 \quad ,$$

où x est le vecteur $\hat{x}$ auquel on a appliqué , dans l'ordre inverse, les permutations de colonnes subies par A. Il suffit donc d'effectuer, sur les lignes de la matrice

$$\begin{bmatrix} \hat{B}_{PR} \\ I \end{bmatrix}$$

les permutations subies par les colonnes de A, dans l'ordre inverse.

* 11. RESOLUTION D'UN SYSTEME SINGULIER. PSEUDO-INVERSE

Soit à résoudre le problème

$$A\,x = y \quad ,$$

dans le cas où A est une matrice carrée singulière. Il ne peut exister de solution à ce problème que moyennant certaines conditions sur le second membre. Effectuons en effet les permutations de lignes et de colonnes amenant le noyau dur de A en tête. On obtient ainsi le système permuté

$$\hat{A}\,\hat{x} = \hat{y} \quad ,$$

avec

$$\hat{A} = \left[\begin{array}{c|c} \hat{A}_{PP} & \hat{A}_{PR} \\ \hline \hat{A}_{RP} & \hat{A}_{RR} \end{array}\right] \begin{matrix} \updownarrow r \\ \\ \end{matrix} \quad \updownarrow n$$

$$\leftrightarrow n$$

où $\hat{A}_{PP}$ est le noyau dur. Le système se partitionne en

$$\begin{cases} \hat{A}_{PP}\,\hat{x}_P + \hat{A}_{PR}\,\hat{x}_R = \hat{y}_P \\ \hat{A}_{RP}\,\hat{x}_P + \hat{A}_{RR}\,\hat{x}_R = \hat{y}_R \end{cases}$$

On déduit de la première équation

$$\hat{x}_P = -\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR}\,\hat{x}_R + \hat{A}_{PP}^{-1}\,\hat{y}_P \quad ,$$

ce qui donne pour la seconde

$$(\hat{A}_{RR} - \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{A}_{PR})\,\hat{x}_R = \hat{y}_R - \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{y}_P \quad .$$

Comme la matrice du premier membre est nulle, on obtient la condition de compatibilité

$$\hat{y}_R = \hat{A}_{RP}\,\hat{A}_{PP}^{-1}\,\hat{y}_P$$

qui signifie tout simplement que les combinaisons linéaires existant entre les lignes de $\hat{A}$ doivent se retrouver dans le second membre.

Si cette condition est satisfaite, on trouve aisément une <u>solution particulière</u> du système en posant $\hat{x}_R = 0$. Il vient alors

$$\hat{x}_{Po} = \hat{A}_{PP}^{-1} \hat{y}_P$$

et la seconde équation devient

$$\hat{y}_R = \hat{A}_{RP} \hat{x}_{P_o} = \hat{A}_{RP} \hat{A}_{PP}^{-1} \hat{y}_P \quad ,$$

en conformité avec la condition de compatibilité. Cette solution particulière s'écrit encore

$$\begin{bmatrix} \hat{x}_{P_o} \\ \hat{x}_{R_o} \end{bmatrix} = \begin{bmatrix} \hat{A}_{PP}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{y}_P \\ \hat{y}_R \end{bmatrix} \quad .$$

Il existe en fait de nombreuses solutions du système considéré. Soient $\hat{x}_{(1)}$ et $\hat{x}_{(2)}$ deux d'entre elles: on a donc

$$\hat{A}\, \hat{x}_{(1)} = \hat{y} \qquad , \qquad \hat{A}\, \hat{x}_{(2)} = \hat{y}$$

et, par différence,

$$\hat{A}\, (\hat{x}_{(1)} - \hat{x}_{(2)}) = 0 \quad .$$

Deux solutions peuvent donc différer d'une solution de $\hat{A}\, \hat{x} = 0$, c'est-à-dire d'une singularité de A. En d'autres termes, <u>la solution générale de l'équation $\hat{A}\, \hat{x} = \hat{y}$ se présente comme la somme d'une solution particulière et d'une combinaison des singularités de $\hat{A}$</u> . Nous savons que celles-ci sont les colonnes de la matrice

$$\begin{bmatrix} - \hat{A}_{PP}^{-1} \hat{A}_{PR} \\ I \end{bmatrix} \quad ,$$

ce qui permet d'écrire la solution générale sous la forme

$$\begin{bmatrix} \hat{x}_P \\ \hat{x}_R \end{bmatrix} = \begin{bmatrix} \hat{x}_{P_o} \\ \hat{x}_{R_o} \end{bmatrix} + \begin{bmatrix} - \hat{A}_{PP}^{-1} \hat{A}_{PR} \\ I \end{bmatrix} \begin{bmatrix} \lambda_{r+1} \\ \cdot \\ \cdot \\ \lambda_n \end{bmatrix} \quad .$$

où apparaît le vecteur

$$\hat{1}_R = \begin{bmatrix} \lambda_{r+1} \\ \vdots \\ \lambda_n \end{bmatrix}$$

dont les composantes sont <u>arbitraires</u>. Explicitant notre solution particulière, on a encore

$$\begin{bmatrix} \hat{x}_P \\ \hat{x}_R \end{bmatrix} = \begin{bmatrix} \hat{A}_{PP}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{y}_P \\ \hat{y}_R \end{bmatrix} + \begin{bmatrix} -\hat{A}_{PP}^{-1}\hat{A}_{PR} \\ I \end{bmatrix} \hat{1}_R$$

ou encore, en tenant compte de la présence des zéros,

$$\begin{bmatrix} \hat{x}_P \\ \hat{x}_R \end{bmatrix} = \begin{bmatrix} \hat{A}_{PP}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{y}_P \\ \hat{1}_R \end{bmatrix} + \begin{bmatrix} 0 & -\hat{A}_{PP}^{-1}\hat{A}_{PR} \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{y}_P \\ \hat{1}_R \end{bmatrix}$$

soit, finalement,

$$\begin{bmatrix} \hat{x}_P \\ \hat{x}_R \end{bmatrix} = \underbrace{\begin{bmatrix} \hat{A}_{PP}^{-1} & -\hat{A}_{PP}^{-1} \\ 0 & I \end{bmatrix}}_{\hat{C}} \begin{bmatrix} \hat{y}_P \\ \hat{1}_R \end{bmatrix}$$

Pour calculer cette matrice $\hat{C}$, il suffit d'appliquer à $\hat{A}$ l'algorithme de Gauss-Jordan avec pivotage complet. Il tombe en défaut lorsque l'on obtient la matrice

$$\hat{B} = \begin{bmatrix} \hat{A}_{PP}^{-1} & -\hat{A}_{PP}^{-1}\hat{A}_{PR} \\ \hat{A}_{RP}\hat{A}_{PP}^{-1} & 0 \end{bmatrix} ,$$

dont $\hat{C}$ se déduit par annulation du bloc inférieur gauche et remplacement du bloc inférieur droit par la matrice unité. On vérifie l'admissibilité du second membre juste avant cette substitution: il suffit de calculer le vecteur

$$\hat{y}_R - \hat{A}_{RP}\hat{A}_{PP}^{-1}\hat{y}_P ,$$

qui doit être nul.

On peut encore procéder différemment. En effet, on a

$$\hat{A}\hat{C} = \begin{bmatrix} \hat{A}_{PP} & \hat{A}_{PR} \\ \hat{A}_{RP} & \hat{A}_{RR} \end{bmatrix} \begin{bmatrix} \hat{A}_{PP}^{-1} & -\hat{A}_{PP}^{-1}\hat{A}_{PR} \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ \hat{A}_{RP}\hat{A}_{PP}^{-1} & 0 \end{bmatrix} ,$$

si bien que, si le second membre $\hat{y}$ est admissible,

$$\hat{A}\hat{x} = \hat{A}\hat{C} \begin{bmatrix} \hat{y}_P \\ \hat{1}_R \end{bmatrix} = \begin{bmatrix} \hat{y}_P \\ \hat{A}_{RP}\hat{A}_{PP}^{-1}\hat{y}_P \end{bmatrix} = \hat{y} ,$$

ce que l'on peut vérifier a posteriori. L'égalité ci-dessus étant vraie pour tout $\hat{1}_R$, on peut en particulier choisir $\hat{1}_R = \hat{y}_R$, ce qui donne

$$\hat{A}\hat{C}\hat{y} = \hat{y}$$

chaque fois que $\hat{y}$ est admissible. Le fait, pour la matrice $\hat{C}$, de se com-

porter comme une inverse à droite de $\hat{A}$ chaque fois que le second membre est admissible permet de la qualifier de pseudo-inverse de $\hat{A}$.

* 12. SYSTEMES A MATRICE RECTANGULAIRE COUCHEE

La méthode de la pseudo-inverse s'applique également aux systèmes à matrice rectangulaire "couchée", c'est-à-dire comportant plus de colonnes que de lignes. Au système rectangulaire

$$\begin{matrix} m \\ \updownarrow \end{matrix} \overset{\overset{m}{\longleftrightarrow}}{\underset{n}{\longleftrightarrow}} \begin{bmatrix} A_{PP} & A_{PR} \end{bmatrix} \begin{bmatrix} x_P \\ x_R \end{bmatrix} = y_P \quad ,$$

on peut en effet substituer le système carré singulier

$$\begin{bmatrix} A_{PP} & A_{PR} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_P \\ x_R \end{bmatrix} = \begin{bmatrix} y_P \\ 0 \end{bmatrix} \quad ,$$

ce qui ramène au problème de la section 11.

Il arrive assez fréquemment que l'on ne s'intéresse à la solution que si la matrice $\begin{bmatrix} A_{PP} & A_{PR} \end{bmatrix}$ est de rang maximal, le cas contraire étant à considérer comme une erreur de construction de la matrice. On peut dans ce cas se limiter à un pivotage partiel. En effet, si la matrice est effectivement de rang maximal m, on trouvera toujours, sur les premières lignes, un pivot non nul, et ces m pivotages laisseront les (n-m) dernières lignes égales à zéro, car si $j > m$, $i < m$,

$$\begin{cases} a_{ji}^{(i)} = a_{ji}^{(i-1)} / a_{ii}^{(i-1)} = 0 \\ a_{jk}^{(i)} = a_{jk}^{(i-1)} - \dfrac{a_{ji}^{(i-1)} \; a_{ik}^{(i-1)}}{a_{ii}^{(i-1)}} = 0 - 0 = 0 \; . \end{cases}$$

Arrivé à la $(m+1)^e$ ligne, on trouvera de toute manière une matrice $\hat{B}_{RR}$ peuplée de zéros, sans qu'il soit nécessaire de le vérifier. (Si l'on obtient une ligne de zéros avant la $(m+1)^e$ ligne, il faut, en vertu de nos hypothèses, conclure que la matrice est erronée). On remplace alors $\hat{B}_{RR} = 0$ par la matrice unité, ce qui donne

$$\hat{C} = \begin{bmatrix} \hat{A}_{PP}^{-1} & - \hat{A}_{PP}^{-1} \hat{A}_{PR} \\ 0 & I \end{bmatrix} \quad ,$$

et nous savons que la solution générale du système est

$$\begin{bmatrix} \hat{x}_P \\ \hat{x}_R \end{bmatrix} = \hat{C} \begin{bmatrix} \hat{y}_P \\ \hat{1}_R \end{bmatrix} \quad .$$

Comme on n'a effectué que des permutations de <u>colonnes</u> , seul l'ordre des composantes de $\hat{x}$ a changé, mais non celui du vecteur

$$\begin{bmatrix} \hat{y}_P \\ \hat{l}_R \end{bmatrix} ,$$

ce qui signifie que

$$\begin{bmatrix} y_P \\ l_R \end{bmatrix} = \begin{bmatrix} \hat{y}_P \\ \hat{l}_R \end{bmatrix}$$

et

$$x = C \begin{bmatrix} y_P \\ l_R \end{bmatrix} .$$

13. METHODES DE GAUSS-JORDAN POUR LES MATRICES SYMETRIQUES DEFINIES POSITIVES

13.1 - Pour introduire les matrices symétriques définie positives (en abrégé: s.d.p.), considérons le problème courant consistant à chercher un minimum d'une fonction $\phi(x)$. Pour que cette fonction soit <u>stationnaire</u> en $x = y$, il faut que

$$\text{grad } \phi(y) = 0 .$$

Si, de plus, y est un point <u>minimal</u>, il existe une certaine boule B(0) telle que pour tout $z \in B(0)$, $\phi(y+z) \geq \phi(y)$, l'égalité n'ayant lieu que si $z = 0$. Pour autant que $\phi \in C^2(B(y))$, on peut écrire

$$\phi(y+z) = \phi(y) + z^T \text{grad } \phi(y) + \tfrac{1}{2} z^T H z + o(\|z\|^2) ,$$

H étant la <u>matrice hessienne</u> de ϕ , définie par

$$H_{ij} = \left(\frac{\partial^2 \phi}{\partial x_i \partial x_j} \right)_{x=y} .$$

Cette matrice est visiblement symétrique. Comme $\text{grad } \phi(y) = 0$, on a

$$\phi(y+z) = \phi(y) + \tfrac{1}{2} z^T H z + o(\|z\|^2) ,$$

si bien que pour les faibles valeurs de $\|z\|$, le signe de $(\phi(y+z) - \phi(y))$ est déterminé par celui de $z^T Hz$. On aura donc un minimum si pour tout z <u>tel que $\|z\| \leq \eta$</u>, la forme quadratique $z^T Hz$ vérifie

$$z^T Hz > 0 \quad \text{si} \quad z \neq 0 .$$

Il est clair que cette inégalité subsite après multiplication de z par tout scalaire $\lambda > 0$, si bien que la condition ci-dessus équivaut à dire que <u>pour tout z non nul</u>,

$$z^T Hz > 0 .$$

Ceci justifie l'intérêt de la définition suivante: Une matrice symétrique A est dite définie positive si, pour tout vecteur $x \neq 0$, on a $x^T Ax > 0$.

13.2 - Les propriétés des matrices s.d.p. sont nombreuses:

a) Une matrice s.d.p. ne peut être singulière

Si elle l'était, il existerait un vecteur $x \neq 0$ tel que $Ax = 0$, ce qui entraînerait $x^T Ax = 0$.

b) Les valeurs propres d'une matrice s.d.p. sont positives

A, étant symétrique, admet n vecteurs propres $z_{(1)}, \ldots, z_{(n)}$ orthonormés, auxquels correspondent les valeurs propres $\lambda_1, \ldots, \lambda_n$.

On a

$$0 < z_{(k)}^T A\, z_{(k)} = \lambda_k\, z_{(k)}^T\, z_{(k)} = \lambda_k \quad .$$

c) Inversement, une matrice symétrique dont toutes les valeurs propres sont positives est s.d.p.

Il suffit en effet de développer $x \neq 0$ dans la base des vecteurs propres:

$$x = \sum_{k=1}^{n} \alpha_k\, z_{(k)} \quad ,$$

pour obtenir

$$x^T Ax = \sum_k \sum_l \alpha_k \alpha_l\, z_{(k)}^T A\, z_{(l)} = \sum_{k,l} \alpha_k \alpha_l \lambda_k \delta_{kl} = \sum_k \alpha_k^2 \lambda_k > 0 \quad .$$

d) Une matrice s.d.p. a son déterminant positif

En effet,

$$\mathrm{dtm}\, A = \lambda_1 \lambda_2 \ldots \lambda_n$$

e) Les termes diagonaux d'une matrice s.d.p. sont positifs

En effet, en notant

$$e_{(i)}^T = (0, \ldots, 0, \underset{\uparrow}{1}, 0, \ldots, 0) \quad ,$$
$$\qquad\qquad\qquad\qquad\quad i$$

on a visiblement

$$a_{ii} = e_{(i)}^T A\, e_{(i)} > 0 \quad .$$

f) L'inverse d'une matrice s.d.p. est s.d.p.

En effet,

$$x^T A^{-1} x = (x^T A^{-T})\, A\, (A^{-1} x) > 0 \quad .$$

g) Quelle que soit la partition d'une matrice s.d.p. de la

$$A = \begin{bmatrix} A_{RR} & A_{RC} \\ A_{CR} & A_{CC} \end{bmatrix} \quad ,$$

avec A_{RR} et A_{CC} carrées, ces dernières sous-matrices sont s.d.p.

En effet, pour $x_R \neq 0$,

$$x_R^T A_{RR} x_R = \begin{bmatrix} x_R^T & 0 \end{bmatrix} A \begin{bmatrix} x_R \\ 0 \end{bmatrix} > 0$$

et pour $x_C \neq 0$,

$$x_C^T A_{CC} x_C = \begin{bmatrix} 0 & x_C^T \end{bmatrix} A \begin{bmatrix} 0 \\ x_C \end{bmatrix} > 0 \ .$$

h) On appelle sous-déterminants principaux d'une matrice A les déterminants

$$a_{11} \ , \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \ , \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \ , \ \ldots \ , \quad \text{dtm A}$$

Cela étant, une matrice s.d.p. a ses sous-déterminants principaux positifs. Ce sont en effet les déterminants de matrices du type A_{RR} ci-dessus.

i) La réciproque de cette propriété constitue le théorème de SYLVESTER:

Une matrice symétrique dont les sous-déterminants principaux sont positifs est s.d.p..

C'est le critère de positive définition le plus connu, encore qu'il ne soit pas le plus maniable.

Nous démontrerons ce théorème par récurrence sur la dimension de la matrice:

- Pour n = 1 , $A = \alpha$, dtm $A = \alpha > 0$ entraîne $\alpha x^2 \ 0$.

- Si le théorème est vrai pour la dimension (n-1), il l'est également pour la dimension n

En effet, on peut décomposer A en

$$A = \left[\begin{array}{c|c} A_{RR} & b \\ \hline b^T & \alpha \end{array}\right] \begin{array}{l} \updownarrow (n-1) \\ \updownarrow \ 1 \end{array}$$

et, comme le théorème est vrai pour les matrices de dimension (n-1), A_{RR} est définie positive. De plus, la règle de Frobenius-Schur donne

$$0 < \text{dtm } A = \text{dtm } A_{RR} \ . \ (\alpha - b^T A_{RR}^{-1} b)$$

soit, comme dtm $A_{RR} > 0$,

$$\alpha > b^T A_{RR}^{-1} b \ .$$

Soit alors un vecteur $x = \begin{bmatrix} x_R \\ \xi \end{bmatrix}$. On a

$$x^T A x = x_R^T A_{RR} x_R + \xi b^T x_R + x_R^T b \xi + \alpha \xi^2$$

$$> x_R^T A_{RR} x_R + \xi b^T x_R + x_R^T b \xi + \xi b^T A_{RR}^{-1} b \xi =$$

$$= (x_R^T A_{RR}) A_{RR}^{-1} (A_{RR} x_R) + \xi b^T A_{RR}^{-1} (A_{RR} x_R) + (x_R^T A_{RR}) A_{RR}^{-1} b\xi$$

$$+ \xi b^T A_{RR}^{-1} b\xi$$

$$= (x_R^T A_{RR} + \xi b^T) A_{RR}^{-1} (A_{RR} x_R + \xi b) \geq 0 \ ,$$

puisque A_{RR}^{-1} est s.d.p. comme A_{RR} .

13.3 - Relations avec l'élimination de Gauss-Jordan

a) Si A est s.d.p. , le pivotage diagonal ne peut tomber en défaut et donne toujours des pivots positifs.

En effet, on a

$$\Pi_1 = a_{11}$$

$$\Pi_1 \Pi_2 = a_{11} a_{22}^{(1)} = \begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{matrix}$$

$$\Pi_1 \Pi_2 \Pi_3 = a_{11} a_{22}^{(1)} a_{33}^{(2)} = \begin{matrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{matrix}$$

.....

$$\Pi_1 \Pi_2 \ \ldots \Pi_n = \text{dtm A},$$

car dans le pivotage diagonal, les transformations ne dépendent pas de la taille des matrices.

b) Réciproquement, une matrice symétrique est définie positive si, lors d'une élimination de Gauss ou Gauss-Jordan par pivotage DIAGONAL, tous les pivots sont positifs.

La démonstration ci-dessus ne comporte en effet que des égalités. Il s'agit du critère de positive définition dont l'application demande le plus petit nombre d'opérations.

13.4 - Remarque - Pour obtenir la meilleure stabilité numérique, il convient de faire les pivotages dans un certain ordre, défini par ce critère, que le prochain pivot sera le plus grand des termes diagonaux non pivotés.

* 14. MISE EN OEUVRE DE LA METHODE DE GAUSS-JORDAN SUR UNE MATRICE STOCKEE SYMETRIQUE

Soit A une matrice symétrique à laquelle le pivotage diagonal peut s'appliquer. C'est le cas des matrices s.d.p., mais également de bien d'autres matrices (voir, par exemple, exercice 1).

Observons qu'il n'est besoin, pour stocker une matrice symétrique, que de sa triangulaire inférieure $\{ a_{ij}, i=1,\ldots, n ; j = 1, \ldots, i \}$.

On profite de cet état de choses en mettant la matrice en mémoire sous la forme d'un <u>vecteur</u> s:

$$s^T = (a_{11}, a_{21}, a_{22}, a_{31}, a_{32}, a_{33}, a_{41}, \ldots \quad)$$

La correspondance entre la matrice A et le vecteur s est aisée à déterminer: pour $j \leqslant i$,

$$a_{ij} = s_k \ ,$$

où k est le nombre de termes de (i-1) lignes arrêtées à la diagonale, augmenté de j, ce qui donne

$$k = 1 + 2 + \ldots + (i-1) + j = \frac{i(i-1)}{2} + j \ .$$

C'est ce que nous appellerons le <u>stockage symétrique</u>. Il ramène la mémoire nécessaire à $n(n+1)/2$ composantes, soit un peu plus que la moitié d'une matrice stockée de la manière classique.

Ceci posé, dans le processus de Gauss-Jordan, on obtient, après (i-1) pivotages <u>diagonaux</u>, la matrice

$$A^{(i-1)} = \begin{bmatrix} A_{PP}^{-1} & -A_{PP}^{-1} A_{PR} \\ A_{RP} A_{PP}^{-1} & A_{RR} - A_{RP} A A_{PP}^{-1} A_{PR} \end{bmatrix} = \begin{bmatrix} A_{PP}^{(i-1)} & -A_{RP}^{(i-1)T} \\ A_{RP}^{(i-1)} & A_{PP}^{(i-1)} \end{bmatrix} \begin{matrix} \updownarrow i-1 \\ \updownarrow n-i+1 \end{matrix}$$

La matrice $A^{(i-1)}$ obtenue en cours de route <u>n'est donc pas symétrique</u> , mais il existe toujours des relations entre les termes symétriquement disposés:

$$\left\{ \begin{array}{l} \cdot\ A_{PP}^{(i-1)} \text{ est symétrique, donc } a_{kl}^{(i-1)} = a_{lk}^{(i-1)} \quad \text{pour } k \ \& \ l \leqslant i-1 \\ \cdot\ A_{RR}^{(i-1)} \text{ est symétrique, donc } a_{kl}^{(i-1)} = a_{lk}^{(i-1)} \quad \text{pour } k \ \& \ l \geqslant i \\ \cdot\ A_{PR}^{(i-1)} = -A_{RP}^{(i-1)T} \ , \quad \text{donc } a_{kl}^{(i-1)} = -a_{lk}^{(i-1)} \end{array} \right.$$

$$\text{pour} \left\{ \begin{array}{lll} k \geqslant i & \& & l \leqslant i-1 \\ l \geqslant i & \& & k \leqslant i-1 \end{array} \right.$$

Grâce à ces relations, il est possible d'effectuer tout le calcul en ne conservant que la triangulaire inférieure. Reprenons en effet les opérations classiques:

a) <u>Pour k & l $\neq$ i</u>

$$a_{kl}^{(i)} = a_{kl}^{(i-1)} - \frac{a_{ki}^{(i-1)} \ a_{il}^{(i-1)}}{a_{ii}^{(i-1)}}$$

On notera que

$$\text{Pour } \begin{cases} k<i \ , \ a_{ki}^{(i-1)} = - a_{ik}^{(i-1)} \ , \text{ car } k \leqslant i-1 \text{ et } i \geqslant i \ . \\ k>i \ , \ a_{ki}^{(i-1)} \text{ est connu.} \end{cases}$$

$$\text{Pour } \begin{cases} l<i \ , \ a_{il}^{(i-1)} \text{ est connu} \\ l>i \ , \ a_{il}^{(i-1)} = a_{li}^{(i-1)} \ , \text{ car } l \geqslant i \text{ et } i \geqslant i \end{cases}$$

Il en résulte les trois cas suivants:

$$\begin{cases} \text{Pour } k < i \ , \ l \leqslant k, \ a_{kl}^{(i)} = a_{kl}^{(i-1)} + \dfrac{a_{ik}^{(i-1)} \ a_{il}^{(i-1)}}{a_{ii}^{(i-1)}} \\ \text{Pour } k > i \ , \ l < i, \ a_{kl}^{(i)} = a_{kl}^{(i-1)} - \dfrac{a_{ki}^{(i-1)} \ a_{il}^{(i-1)}}{a_{ii}^{(i-1)}} \\ \text{Pour } k > i, \ i < l \leqslant k, \ a_{kl}^{(i)} = a_{kl}^{(i-1)} - \dfrac{a_{ki}^{(i-1)} \ a_{li}^{(i-1)}}{a_{ii}^{(i-1)}} \end{cases}$$

b) <u>Termes $a_{ij}^{(i)}$, $j < i$</u>

$$a_{ij}^{(i)} = - \frac{a_{ij}^{(i-1)}}{a_{ii}^{(i-1)}}$$

c) <u>Termes $a_{ji}^{(i)}$, $j > i$</u>

$$a_{ji}^{(i)} = \frac{a_{ji}^{(i-1)}}{a_{ii}^{(i-1)}}$$

d) <u>Pivot</u>

$$a_{ii}^{(i)} = \frac{1}{a_{ii}^{(i-1)}}$$

On peut gagner quelques opérations en procédant comme suit (dans la matrice elle-même)

(1) <u>Pivot</u> : $a_{ii}^{(i)} = \dfrac{1}{a_{ii}^{(i-1)}}$ 1 O.S.

(2) <u>Pour k i</u> :

$\alpha_k = a_{ik}^{(i-1)} \cdot a_{ii}^{(i)}$ 1 O.S.

Pour $l \leqslant k$, $a_{kl}^{(i)} = a_{kl}^{(i-1)} + \alpha_k \ a_{il}^{(i-1)}$ k O.S.

(3) Ligne i :

Pour $j<i$, $a_{ij}^{(i)} = - a_{ij}^{(i-1)} \; a_{ii}^{(i)}$ (i-1) O.S.

(4) Pour k i:

Pour $l<i$, $a_{kl}^{(i)} = a_{kl}^{(i-1)} + a_{ki}^{(i-1)} \; a_{il}^{(i)}$ (i-1) O.S.

$\beta_k = a_{ki}^{(i-1)} \; a_{ii}^{(i)}$ 1 O.S.

Pour $i<l\leq k$, $a_{kl}^{(i)} = a_{kl}^{(i-1)} - \beta_k \; a_{li}^{(i-1)}$ (k-i) O.S.

(5) Colonne i :

Pour $j>i$, $a_{ji}^{(i)} = a_{ji}^{(i-1)} \circ a_{ii}^{(i)}$ (n-i) O.S.

Le nombre d'opérations significatives vaut alors, pour le pivotage numéro i,

(1) : 1 O.S.

(2) : $2 + 3 + \ldots + i = \dfrac{(i+2)(i-1)}{2} = \dfrac{i^2 + i - 2}{2}$ O.S.

(3) : (i-1) O.S.

(4) : $(i+1) + \ldots + n = \dfrac{(n-i+1)(n-i)}{2} = \dfrac{n^2 - i^2 + n - i}{2}$ O.S.

(5) : (n-i) O.S.

Somme : $n + \dfrac{n^2 - i^2 + n - i}{2} + \dfrac{i^2 + i - 2}{2} = \dfrac{n^2 + 3n - 2}{2}$

Pour le n n pivotages, cela donne

$$\frac{n^3 + 3n^2 - 2n}{2} \approx n^3/2$$

opérations significatives, soit environ la moitié de ce que demande une matrice complète. A ce gain, il faut ajouter que la place occupée en mémoire est également environ deux fois moindre.

*15. MISE EN OEUVRE DE LA TRIANGULARISATION DE GAUSS SUR UNE MATRICE STOCKEE SYMETRIQUE

Remarquons que, dans la matrice $A^{(i-1)}$ obtenue après (i-1) pivotages diagonaux à partir de A symétrique,

$$A^{(i-1)} = \begin{bmatrix} a_{11} & & & * \\ & \ddots & & \\ & & a_{i-1,i-1}^{(i-2)} & \\ 0 & & a_{ii}^{(i-1)} \ldots\ldots & a_{in}^{(i-1)} \\ & & \vdots & \vdots \\ & & a_{ni}^{(i-1)} \ldots\ldots & a_{nn}^{(i-1)} \end{bmatrix}$$

la matrice carrée

$$\begin{bmatrix} a_{ii}^{(i-1)} & \cdots & a_{in}^{(i-1)} \\ \vdots & & \vdots \\ a_{ni}^{(i-1)} & \cdots & a_{nn}^{(i-1)} \end{bmatrix}$$

est symétrique, car obtenue par des transformations du type

$$a_{kl}^{(j)} = a_{kl}^{(j-1)} - \frac{a_{kj}^{(j-1)} \, a_{jl}^{(j-1)}}{a_{jj}^{(j-1)}}$$

qui préservent la symétrie pour $k,l > j$. Il suffit donc de connaître la triangulaire supérieure, que l'on stockera sous la forme d'un vecteur s, tel que

$$a_{ij} = s_k \quad ,$$

avec

$$k = \frac{(j-1)j}{2} + i \quad .$$

On écrit alors

$$a_{kl}^{(i)} = a_{kl}^{(i-1)} - \frac{a_{ik}^{(i-1)} \, a_{il}^{(i-1)}}{a_{ii}^{(i-1)}} \qquad \text{pour} \quad k,l > i \quad .$$

16. RESOLUTION D'UN SYSTEME PAR ORTHOGONALISATION

Si l'on considère la matrice A sous la forme de vecteurs lignes,

$$A = \begin{bmatrix} l_{(1)}^T \\ \vdots \\ l_{(n)}^T \end{bmatrix} \quad ,$$

le système $Ax = b$ peut être écrit sous la forme

$$l_{(1)}^T x = b_1 \; , \; \ldots \; , \; l_{(n)}^T x = b_n \quad .$$

On commence par normer la première équation:

$$\hat{l}_{(1)} = l_{(1)} / \|l_{(1)}\| \quad , \quad \hat{b}_1 = b_1 / \|l_{(1)}\| \quad .$$

On orthogonalise alors les ligne k, $k = 2,\ldots,n$, à toutes les précédentes:

$$l_{(k)}^* = l_{(k)} - \sum_{j=1}^{k-1} (l_{(k)}^T \hat{l}_{(j)}) \hat{l}_{(j)} \quad , \quad b_k^* = b_k - \sum_{j=1}^{k-1} (l_{(k)}^T l_{(j)}) b_j \quad ,$$

puis on la norme:

$$\hat{l}_{(k)} = l_{(k)}^* / \|l_{(k)}^*\| \quad , \quad \hat{b}_k = b_{(k)}^* / \|l_{(k)}^*\| \quad .$$

A la fin de ce processus, on obtient un système de la forme

$$B x = c \quad ,$$

avec une matrice B orthogonale. La solution est donc donnée par

$$x = B^T c \ .$$

En outre, le déterminant est donné, au signe près, par

$$|\mathrm{dtm}\ A| = \|l^*_{(1)}\| \ . \ \ldots \ . \ \|l^*_{(n)}\| \ .$$

En effet, l'orthogonalisation sans normation ne fait que soustraire à chaque ligne une combinaison des autres et ne modifie donc pas le déterminant. La matrice

$$C = \begin{bmatrix} l^{*T}_{(1)} \\ \vdots \\ l^{*T}_{(n)} \end{bmatrix}$$

a donc le même déterminant que A, ce qui entraîne encore

$$(\mathrm{dtm}\ A)^2 = \mathrm{dtm}\ (C\ C^T).$$

Or, $CC^T = \mathrm{diag}(\ \|l^*_{(1)}\|^2, \ldots, \|l^*_{(n)}\|^2)$, a pour déterminant le produit des carrés des normes des $l^*_{(i)}$.

Calculons le nombre d'opérations significatives que nécessite cette méthode. Pour la ligne k, il faut effectuer:

- Le calcul des produits scalaires avec les k lignes précédentes ..... (k-1)n O.S.
- L'orthogonalisation ..... (k-1)n O.S.
- La correction du second membre ..... (k-1) O.S.
- Le calcul du carré de la norme ..... n O.S.
- (La racine carrée)
- La normation ..... (n+1) O.S.

ce qui donne (2n+1)k O.S. + 1 racine carrée. Au total, pour traiter toutes les lignes, il faut

$$\sum_{k=1}^{n} (2n+1)k = (2n+1)\,\frac{(n+1)n}{2} \approx n^3 \quad \text{O.S.} \ ,$$

plus n racines carrées. Cela fait trois fois plus d'opérations que l'élimination de Gauss.

17. METHODE DE CHOLESKI

La méthode de Choleski consiste à chercher une décomposition d'une matrice symétrique A, de la forme

$$A = L\ L^T \ ,$$

L étant une matrice triangulaire inférieure inversible.

17.1 - Réalité de la transformation

Montrons qu'une matrice symétrique A admet une décomposition $A = LL^T$

avec L triangulaire inférieure inversible et REELLE si et seulement si elle est définie positive.

a) La condition est nécessaire, car si $A = LL^T$, on a toujours

$$x^T A x = x^T L L^T x = \|L^T x\|^2 > 0 \qquad \text{si } x \neq 0 \ .$$

b) La condition est suffisante. Raisonnons par récurrence sur la dimension de la matrice. Pour n = 1, on a $A = \alpha$, $\alpha > 0$, $L = \sqrt{\alpha}$, et la propriété est triviale. Si elle est vraie pour toutes les matrices de dimension (n-1), montrons qu'elle est encore vraie pour la dimension n. Soit en effet une matrice A de dimension n, que l'on partitionne comme suit:

$$A = \left[\begin{array}{c|c} B & b \\ \hline b^T & \alpha \end{array}\right] \begin{array}{l} \updownarrow n-1 \\ \updownarrow 1 \end{array} \ .$$

La sous-matrice B étant s.d.p. (voir section 13), il existe une factorisation $B = MM^T$, M triangulaire inférieure inversible réelle. Cherchons L sous la forme

$$L = \begin{bmatrix} M & 0 \\ l^T & \lambda \end{bmatrix} \ .$$

Les conditions pour que $A = LL^T$ sont

$$\begin{bmatrix} B & b \\ b^T & \alpha \end{bmatrix} = \begin{bmatrix} M & 0 \\ l^T & \lambda \end{bmatrix} \begin{bmatrix} M^T & l \\ 0 & \lambda \end{bmatrix} = \left[\begin{array}{c|c} MM^T & Ml \\ \hline l^T M^T & l^T l + \lambda^2 \end{array}\right] \ ,$$

soit

$$\begin{cases} MM^T = B \ , \text{ déjà vérifié} \\ Ml = b \ , \text{ soit } \ l = M^{-1} b \ \text{ (sans problème)} \\ l^T l + \lambda^2 = \alpha \ . \end{cases}$$

C'est la dernière relation qui prête à litige. En effet, il faut que

$$\lambda^2 = \alpha - l^T l > 0 \ .$$

On a

$$\alpha - l^T l = \alpha - b^T M^{-T} M^{-1} b = \alpha - b^T B^{-1} b ,$$

et ce nombre est positif, car la règle de Frobenius-Schur donne

$$\text{dtm } A = \text{dtm } B \cdot (\alpha - b^T B^{-1} b) > 0 \ .$$

17.2 - Calcul de la matrice L

On a évidemment, pour $i \geq j$,

$$a_{ij} = \sum_{k=1}^{n} l_{ik} l_{jk} = \sum_{k=1}^{j} l_{ik} l_{jk} \ .$$

a) Pour j = 1 (première colonne)

On a d'abord

$$a_{11} = l_{11}^2$$

soit

$$l_{11} = \sqrt{a_{11}}$$

et, pour les autres termes,

$$a_{i1} = \sum_{k=1}^{1} l_{ik}\, l_{1k} = l_{i1}\, l_{11} \quad ,$$

ce qui donne

$$l_{i1} = \frac{a_{i1}}{l_{11}} \quad .$$

b) <u>Pour la j^e colonne, en supposant les (j-1) précédentes déjà calculées</u>

On a d'abord

$$a_{jj} = \sum_{k=1}^{j} l_{jk}\, l_{jk} = l_{jj}^2 + \sum_{k<j} l_{jk}^2 \quad ,$$

soit

$$l_{jj} = \sqrt{a_{jj} - \sum_{k<j} l_{jk}^2} \quad ;$$

puis, pour i j,

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\, l_{jk} = l_{ij}\, l_{jj} + \sum_{k<j} l_{ik}\, l_{jk} \quad ,$$

soit

$$l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k<j} l_{ik}\, l_{jk}\right) ,$$

les termes du second membre étant tous connus.

17.3 - <u>Exploitation de la factorisation</u>

Partant de

$$A\, x = b \quad ,$$

on écrit

$$L\, L^T\, x = b$$

et, posant

$$y = L^T\, x,$$

on ramène le système à

$$L\, y = b \quad ,$$

système triangulaire facile à résoudre. Ayant obtenu y, on déduit x du système

$$L^T\, x = y \quad ,$$

lui aussi triangulaire. (Deux remontées après factorisation)

17.4 - Nombre d'opérations significatives

a) Terme diagonal l_{jj} : (j-1) produits, 1 racine carrée

b) Autres termes l_{kj} : (j-1) termes demandant chacun (j-1) produits et une division, soit $j(j-1)$ O.S.

Pour la ligne j, on a donc au total $(j+1)(j-1) = j^2 - 1$ O.S. ,en négligeant la racine carrée. Pour la matrice entière, il faudra donc

$$\sum_{j=1}^{n}(j^2 - 1) = \sum_{j=1}^{n} j^2 - n = \frac{n(n+1)(2n+1)}{6} - n \approx n^3/3 .$$

A cela, il faut ajouter, pour chaque second membre, deux remontées, soit environ n^2 O.S. . Comme on le voit, c'est très comparable à une triangularisation de Gauss.

* 18. STABILITE D'UN SYSTEME MATRICIEL PAR RAPPORT AUX DONNEES

18.1-Dans ce paragraphe, nous utiliserons anticipativement les normes matricielles du type

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

étudiées en détail dans le chapitre relatif aux méthodes itératives de résolution des systèmes matriciels.

18.1 - Effet d'une perturbation du second membre

Supposons que, dans le système $Ax = b$, A soit une matrice connue exactement, mais que b doit entaché d'une erreur δb. La solution obtenue sera donc perturbée d'un vecteur δx, défini par

$$A(x + \delta x) = b + \delta b ,$$

soit

$$A\,\delta x = \delta b.$$

Dans quelle mesure l'erreur relative, mesurée par $\|\delta x\|/\|x\|$, sera-t-elle petite si $\|\delta b\|/\|b\|$ est petit? On peut mesurer l'amplification de l'erreur relative par le rapport

$$\gamma(A, b) = \sup_{\delta b \neq 0} \left(\frac{\|\delta x\|}{\|x\|} \cdot \frac{\|b\|}{\|\delta b\|} \right)$$

appelé conditionnement du système $Ax=b$. Ce nombre dépend à la fois de la matrice A et du second membre b. Plus il est grand, plus la solution sera sensible à une perturbation du second membre. Les systèmes à grand conditionnement sont dits mal conditionnés.

On peut donner des expressions plus simples du conditionnement, car, comme

$$\delta x = A^{-1}\,\delta b,$$

on a

$$\gamma(A,\ b) = \sup_{\delta b \neq 0} \left(\frac{\|b\|}{\|x\|} \quad \frac{\|A^{-1}\,\delta b\|}{\|\delta b\|} \right) = \frac{\|A^{-1}\|\ \|b\|}{\|x\|}$$

ou encore

$$\gamma(A,b) = \frac{\|A^{-1}\|\ \|b\|}{\|A^{-1}\ b\|} \ .$$

En termes de la solution x, on a encore

$$\gamma(A,\ b) = \frac{\|A^{-1}\|\ \|Ax\|}{\|x\|} \ .$$

Evidemment,

$$\frac{\|\delta x\|}{\|x\|} \leq \gamma(A,\ b)\ \frac{\|\delta b\|}{\|b\|} \ ;$$

bien plus, c'est la meilleure borne possible, car il existe toujours un δb tel que $\|A^{-1}\delta\ b\| = \|A^{-1}\|\ \|\delta b\|$.

Lorsque l'on désire résoudre plusieurs systèmes, on préfère une borne enveloppant toutes les précédentes. On définit alors le <u>conditionnement de la matrice A par</u>

$$\gamma(A) = \sup_{b \neq 0} \gamma(A,\ b).$$

Il est facile de voir que

$$\gamma(A) = \sup_{b \neq 0} \frac{\|A^{-1}\|\ \|b\|}{\|A^{-1}b\|} = \|A^{-1}\| \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \|A^{-1}\|\ \|A\| \ .$$

On a donc, quel que soit b,

$$\frac{\|\delta x\|}{\|x\|} \leq \gamma(A)\ \frac{\|\delta b\|}{\|b\|} \ .$$

On notera que, comme $A\ A^{-1} = I$,

$$1 = \|I\| \leq \|A\|\ \|A^{-1}\| = \gamma(A) \ .$$

18.2 - <u>Stabilité de la solution par rapport à une perturbation simultanée de la matrice et du second membre</u>

Supposons à présent A <u>et</u> b perturbés. On résout donc

$$(A + \delta A)(x + \delta x) = b + \delta b \ ,$$

ce qui donne

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

et

$$\delta x = A^{-1}\,\delta b - A^{-1}\,\delta A\ x - A^{-1}\,\delta A\,\delta x \ .$$

On en déduit immédiatement

$$\|\delta x\| \leq \|A^{-1}\|\ \|\delta b\| + \|A^{-1}\,\delta A\|\ \|x\| + \|A^{-1}\delta\ A\|\ \|\delta x\| \ ,$$

soit

$$\|\delta x\| \leqslant \frac{\|A^{-1}\| \, \|\delta b\| + \|A^{-1} \delta A\| \, \|x\|}{1 - \|A^{-1} \delta A\|} \quad . \qquad (1)$$

La perturbation de la solution reste donc toujours finie si $\|A^{-1} \delta A\| < 1$. Cette condition garantit en fait que <u>la matrice $(A + \delta A)$ reste inversible.</u> En effet, une condition <u>nécessaire</u> de singularité de $A + \delta A$ est $\|A^{-1} \delta A\| \geqslant 1$, car pour toute solution y de

$$(A + \delta A)\, y = 0 \ ,$$

on a également

$$y = - A^{-1} \delta A \, y \ ,$$

ce qui entraîne

$$\|y\| \leqslant \|A^{-1} \delta A\| \, \|y\|$$

soit

$$\|A^{-1} \delta A\| \geqslant 1 \ .$$

Pour passer aux erreurs relatives, divisons par $\|x\|$ les deux membres de (1): on obtient, en supposant $\|A^{-1}\| \, \|\delta A\| < 1$,

$$\frac{\|\delta x\|}{\|x\|} \quad \frac{\|A^{-1}\| \, \frac{\|\delta b\|}{\|x\|} + \|A^{-1}\| \, \|\delta A\|}{1 - \|A^{-1}\| \, \|\delta A\|} =$$

$$\frac{\|A^{-1}\| \, \frac{\|\delta b\|}{\|Ax\|} \, \frac{\|Ax\|}{\|x\|} + \gamma(A) \, \frac{\|\delta A\|}{\|A\|}}{1 - \gamma(A) \, \frac{\|\delta A\|}{\|A\|}}$$

$$\frac{\gamma(A)}{1 - \gamma(A) \, \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

18.3 - <u>Inégalité inverse</u>

Dans le cas d'une perturbation du second membre seul, on a encore

$$\begin{cases} \delta b = A \delta x \ , \text{ donc } & \|\delta b\| \leqslant \|A\| \, \|\delta x\| \\ x = A^{-1} b \ , \text{ donc } & \|A^{-1}\| \, \|b\| \geqslant \|x\| \end{cases}$$

Divisant des deux résultats, on obtient

$$\frac{\|\delta x\|}{\|x\|} \geqslant \frac{1}{\gamma(A)} \, \frac{\|\delta b\|}{\|b\|}$$

On le constate, dans le cas d'une matrice mal conditionnée ($\gamma(A)$ grand), selon le vecteur b et selon sa perturbation b, l'erreur relative sur la solution peut aussi bien être très petite que très grande par rapport à celle sur le second membre: il n'y a presque plus de liaison entre elles. Au contraire, dans un système idéalement conditionné, $\gamma(A)=1$ et les

deux erreurs relatives sont égales.

18.4 - Critère des colonnes

Nous essaierons, dans les paragraphes qui suivent, de dégager des règles pratiques simples pour détecter un mauvais conditionnement et, dans une certaine mesure, pour y porter remède. Tout d'abord, observons que la structure même du conditionnement,

$$\gamma(A) = \|A\| \, \|A^{-1}\|$$

entraîne

$$\gamma(A) = \gamma(A^{-1}) \quad ,$$

ce dernier mesurant l'effet sur b d'une perturbation de x.

Cela étant, appelons $c_{(1)}, \ldots, c_{(n)}$ les colonnes de la matrice A. On a

$$b = A\,x = x_1\, c_{(1)} + x_2\, c_{(2)} + \ldots + x_n\, c_{(n)} \quad .$$

Quitte à réordonner les x_i, on peut supposer

$$\|c_{(1)}\| \geq \|c_{(2)}\| \geq \ldots \geq \|c_{(n)}\| \, .$$

Considérons alors le cas $x^T = (0, \ldots, 1)$, $(x)^T = (1, 0, \ldots, 0)$. On a donc

$$\|x\| = 1 \, , \quad \|\delta x\| = 1$$

et, comme

$$b = c_{(n)} \, , \qquad b' = c_{(1)} \, ,$$

il vient

$$\gamma(A) \geq \frac{\|\delta b\|}{\|b\|} \, \frac{\|x\|}{\|\delta x\|} = \frac{\|c_{(1)}\|}{\|c_{(n)}\|} \, .$$

Ainsi, le conditionnement est AU MOINS EGAL au rapport de la plus grande norme de colonne à la plus petite. Il est d'ailleurs exactement égal à ce rapport dans le cas d'une matrice à colonnes orthogonales. En effet, dans ce cas particulier (en adoptant les normes euclidiennes), on a

$$\|Ax\|^2 = \sum_i x_i^2 \, \|c_{(i)}\|^2 \, ,$$

forme quadratique dont le maximum et le minimum sur la sphère unité correspondent à $x^T = (1, 0, \ldots, 0)$ et $x^T = (0, \ldots, 0, 1)$.

De ceci, il découle qu'il est raisonnable d'effectuer une mise à échelle des colonnes.

18.5 - Critère des valeurs propres

Ce critère ne s'emploie pas comme tel; mais sert aux raisonnements ultérieurs. Le voici:

Soit A une matrice diagonalisable, de valeurs propres ordonnées par ordre décroissant des valeurs absolues:

$$|\lambda_1| \geq \ldots \geq |\lambda_n|$$

Son conditionnement est supérieur au rapport $|\lambda_1| / |\lambda_n|$

En effet, en notant $z_{(i)}$ le vecteur propre normé correspondant à la valeur propre λ_i , on a

$$A\, z_{(i)} = \lambda_i\, z_{(i)} \cdot$$

Si donc on pose $x = z_{(n)}$ et $\delta x = z_{(1)}$, on aura $b = \lambda_n\, z_{(n)}$ et $\delta b = \lambda_1\, z_{(1)}$, ce qui entraîne

$$\gamma(A) \geqslant \frac{\|\delta b\|}{\|b\|} \frac{\|x\|}{\|\delta x\|} = \frac{|\lambda_1|}{|\lambda_n|} \cdot$$

18.6 - Conditionnement d'une matrice triangulaire

Soit A une matrice triangulaire dont tous les termes diagonaux sont différents. Son conditionnement est supérieur au rapport du plus grand en valeur absolue au plus petit en valeur absolue de ces termes diagonaux.

En effet, pour une telle matrice, les termes diagonaux sont les valeurs propres.

Dès lors, une matrice triangulaire dont les termes diagonaux ont des valeurs absolues très différentes est certainement mal conditionnée.

18.7 - Application à la triangularisation de Gauss

On vérifie aisément que la triangularisation de Gauss, avec division de la ligne du pivot par celui-ci, équivaut à décomposer la matrice A en deux matrices triangulaires:

$A = B\, C$,

avec C triangulaire supérieure de diagonale unitaire et B triangulaire inférieure, produit des matrices élémentaires $B_{(n)} \cdots B_{(1)}$,

$$B_{(1)} = \begin{bmatrix} a_{11} & & & 0 \\ \vdots & 1 & & \\ & & \ddots & \\ a_{n1} & 0 & & 1 \end{bmatrix}, \quad B_{(2)} = \begin{bmatrix} 1 & & & \\ 0 & a_{22}^{(1)} & & 0 \\ & & 1 & \\ 0 & a_{n2}^{(1)} & 0 & 1 \end{bmatrix}, \ \cdots$$

soit

$$B = \begin{bmatrix} a_{11} & & & 0 \\ & a_{22}^{(1)} & & \\ & & \ddots & \\ * & & & a_{nn}^{(n-1)} \end{bmatrix}$$

La triangularisation consiste à passer du système

$A\, x = B\, C\, x = b$

au système équivalent

$$C\ x = B^{-1}\ b\ .$$

Or, pour autant que tous les pivots soient différents, le conditionnement de B est supérieur au rapport du plus grand au plus petit, en valeur absolue. Donc, si les pivots sont très différents en valeur absolue, l'erreur sur le second membre modifié peut être très grande.

Exercice 1 - On considère une matrice symétrique A de la forme

$$A = \begin{bmatrix} A_{PP} & A_{PR} \\ A_{RP} & 0 \end{bmatrix} \begin{matrix} \updownarrow r \\ \updownarrow m \end{matrix}$$

$$\overset{\longleftrightarrow r}{\quad} \qquad \overset{\longleftrightarrow m}{\quad}$$

où A_{PP} est s.d.p. et $A_{PR} = A_{RP}^T$. On suppose que les colonnes de A_{PR} sont linéairement indépendantes.

Montrer que a) $r \geqslant m-r$

b) On peut inverser cette relation par la méthode de Gauss-Jordan à pivotage diagonal, et r pivots seront positifs, les (m-r) autres négatifs.

Solution

a) On ne peut avoir plus de r vecteurs indépendants à r dimensions, donc on doit avoir

$$m-r \leqslant r.$$

b) La matrice A_{PP} étant définie positive, on aura r pivots positifs. Après ces r pivotages, on obtiendra la matrice

$$\begin{bmatrix} A_{PP}^{-1} & - A_{PP}^{-1} A_{PR} \\ A_{RP} A_{PP}^{-1} & - A_{RP} A_{PP}^{-1} A_{PR} \end{bmatrix}$$

Examinons la matrice

$$(- B_{RR}) = A_{RP} A_{PP}^{-1} A_{PR} \quad .$$

Pour $x_R \neq 0$, on doit avoir $A_{PR} x_R \neq 0$, car les colonnes de A_{PR} sont indépendantes. Dès lors,

$$x_R^T (-B_{RR}) x_R = x_R^T A_{RP} A_{PP}^{-1} A_{PR} x_R = (A_{PR} x_R)^T A_{PP}^{-1} (A_{PR} x_R) > 0 \quad ,$$

car A_{PP}^{-1} est définie positive. B_{RR} est donc une matrice définie positive changée de signe, ce qui fournira au pivotage diagonal (m-r) pivots négatifs.

Exercice 2 - On veut résoudre l'équation matricielle

$$Kx = My \ ,$$

avec K et M s.d.p., x = inconnue, y = donnée variable. Montrer que ce problème peut être résolu par un algorithme tenant compte de la symétrie des matrices.

Solution: on décompose K en $K = LL^T$ par Choleski. Alors,

$$LL^T x = My \quad , \quad L^T x = L^{-1} M L^{-T} (L^T y) \quad , \quad L^{-1} M L^{-T} \text{ symétrique.}$$

Il faut donc: a) calculer L ; b) calculer $L^T y$; c) calculer $L^{-1} M L^{-T}$;

d) calculer $L^T x$ e) en déduire x (système triangulaire).

METHODES ITERATIVES DE RESOLUTION DES SYSTEMES LINEAIRES

1. Les méthodes itératives de résolution des systèmes linéaires consistent à construire une suite $x^{(k)}$ de vecteurs tels que

$$x^{(k)} \longrightarrow A^{-1} b.$$

On peut distinguer deux classes importantes de méthodes itératives, à savoir:

- Les <u>méthodes itératives simples</u>, consistant à transformer le système $Ax = b$ en un système équivalent $x = B x + c$, et à chercher la solution comme limite de la suite

$$x^{(k+1)} = Bx^{(k)} + c ,$$

avec un point de départ $x^{(o)}$ arbitraire.

- Les méthodes de <u>minimisation d'une forme quadratique</u>, réservées aux matrices symétriques définies positives.

2. NORME D'UNE MATRICE

2.1 - Nous aurons besoin de la notion de norme d'une matrice en tant qu'opérateur.

Rappelons d'abord que dans R^n, de nombreuses normes sont possibles. Les trois plus courantes sont

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$$

$$\|x\|_\infty = \sup_{i=1,\ldots,n} |x_i| .$$

Ces normes sont <u>équivalentes</u>, c'est-à-dire bornées l'une par rapport à l'autre: on vérifie aisément que

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq n \|x\|_\infty .$$

Le choix d'une norme pour les vecteurs étant fait, on définit la norme d'une matrice A comme le plus grand facteur d'amplification qu'elle introduit:

$$\boxed{\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}} .$$

Cette définition implique les relations suivantes:

a) $\|\lambda A\| = |\lambda| \, \|A\|$ pour tout $\lambda \in R$

$$b)\quad \|A+B\| = \sup_{x\neq 0} \frac{\|(A+B)x\|}{\|x\|} \leqslant \sup_{x\neq 0} \frac{\|\,\|Ax\| + \|Bx\|\,\|}{\|x\|}$$

$$\leqslant \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} + \sup_{x\neq 0} \frac{\|Bx\|}{\|x\|} = \|A\| + \|B\| .$$

$$c)\quad \|AB\| = \sup_{x\neq 0} \frac{\|ABx\|}{\|x\|} \leqslant \sup_{x\neq 0} \frac{\|A\|\,\|Bx\|}{\|x\|} = \|A\|\,\|B\|$$

2.2 - Calculons à présent les normes matricielles <u>subordonnées</u> aux trois normes vectorielles définies ci-dessus:

a) <u>Norme 1</u>

Posant $y = Ax$, on a

$$y_i = \sum_j a_{ij}\, x_j$$

donc

$$|y_i| \leqslant \sum_j |a_{ij}|\,|x_j|$$

Il en découle

$$\sum_i |y_i| \leqslant \sum_{ij} |a_{ij}|\,|x_j| = \sum_j \Big(\sum_i |a_{ij}|\Big)\,|x_j|$$

$$\sup_{j=1,\ldots,n} \Big(\sum_i |a_{ij}|\Big) \cdot \sum_j |x_j|$$

soit

$$\frac{\|y\|_1}{\|x\|_1} \leqslant \sup_{j=1,\ldots,n} \sum_i |a_{ij}| = \sup_{j=1,\ldots,n} \|c_{(j)}\|_1 \quad ,$$

où $c_{(i)}$ est le vecteur-colonne n° i de la matrice A .

Pour prouver qu'il s'agit bien de la <u>meilleure</u> borne supérieure, montrons qu'elle peut être atteinte. Il suffit en effet de repérer la colonne j_o dont la norme 1 est la plus grande et de considérer le vecteur x tel que $x_{j_o} = 1$, $x_i = 0$ si $i \neq j_o$. Pour ce vecteur,

$$y_i = a_{ij_o} \quad , \quad \|y\|_1 = \sum_j |a_{ij_o}| .$$

Ainsi, nous pouvons affirmer que

$$\boxed{\|A\|_1 = \sup_{j=1,\ldots,n} \sum_i |a_{ij}| = \sup_{j=1,\ldots,n} \|c_{(j)}\|_1}$$

b) <u>Norme 2</u>

On a visiblement

$$\|y\|_2^2 = y^T y = x^T A^T A x .$$

Or, la matrice A^TA, symétrique, admet une base de vecteurs propres orthonormés $z_{(1)}, \ldots, z_{(n)}$ correspondant aux valeurs propres $\lambda_1, \ldots, \lambda_n$. Développons x dans cette base:

$$x = \sum_{i=1}^{n} \alpha_i z_{(i)} .$$

Il vient

$$A^TA\, x = \sum_{i=1}^{n} \alpha_i A^TA\, z_{(i)} = \sum_{i=1}^{n} \alpha_i \lambda_i z_{(i)}$$

et

$$x^T A^TA\, x = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i z_{(i)}^T \alpha_j \lambda_j z_{(j)} = \sum_{i=1}^{n} \alpha_i^2 \lambda_i ,$$

en tenant compte des relations d'orthogonalité des vecteurs propres. On a donc

$$\|Ax\|_2^2 = \sum_{i=1}^{n} \alpha_i^2 \lambda_i \leq \lambda_{max} \sum_{i=1}^{n} \alpha_i^2 = \lambda_{max} \|x\|_2^2 ,$$

ce qui entraîne directement

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} \leq \lambda_{max} .$$

Cette valeur peut être atteinte en choisissant pour x un vecteur propre associé à λ_{max}. Par conséquent,

$$\boxed{\|A\|_2 = \sqrt{\lambda_{max}(A^TA)}}$$

Le principal inconvénient de cette norme est la difficulté de calculer les valeurs propres.

c) Norme ∞

Comme

$$y_i = \sum_j a_{ij} x_j ,$$

on a pour tout i

$$|y_i| \leq \sum_j |a_{ij}|\,|x_j| \leq \sum_j |a_{ij}| \sup_j |x_j|$$

$$\leq \sup_i \Big(\sum_j |a_{ij}|\Big) \cdot \sup_j |x_j| .$$

En particulier,

$$\sup_i |y_i| \leq \sup_i \left(\sum_j |a_{ij}| \right) . \sup_j |x_j| \quad ,$$

soit

$$\frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \sup_i \sum_j |a_{ij}| = \sup_i \|l_i\|_1 \quad ,$$

en notant $l_{(1)}^T, \ldots, l_{(n)}^T$ les vecteurs-lignes de A .

Cette borne peut être atteinte. En effet, repérons la ligne i_o dont la norme 1 est maximale. Le vecteur x ~~défini par~~ ayant pour composantes

$$x^T = (\text{sign}(a_{i_o,1}) , \text{sign}(a_{i_o,2}) , \ldots) \qquad x_j = \frac{a_{i_o,j}}{|a_{i_o,j}|} \text{ si } a_{i_o,j} \neq 0, \quad x_j = 0 \text{ sinon}$$

est de norme ∞ égale à un et on a visiblement

$$\|Ax\|_\infty = \sup_i \|l_i\|_1 \quad .$$

Il s'agit donc bien de la meilleure borne supérieure et

$$\boxed{\|A\|_\infty = \sup_i \sum_j |a_{ij}| = \sup_i \|l_i\|_1} \quad .$$

2.3 - Montrons que <u>deux normes matricielles subordonnées à des normes vectorielles équivalentes sont elles-mêmes équivalentes</u>.

En effet, si $\|.\|$ et $|||.|||$ sont deux normes équivalentes, il existe une relation du type

$$\alpha \|x\| \leq |||x||| \leq \mathcal{A} \|x\| \quad ,$$

avec α et $\mathcal{A}$ strictement positifs, valable pour <u>tous</u> les vecteurs x. Dès lors, pour tout $x \neq 0$,

$$\frac{|||Ax|||}{|||x|||} \leq \frac{\mathcal{A}}{\alpha} \frac{\|Ax\|}{\|x\|} \quad \text{et} \quad \frac{\|Ax\|}{\|x\|} \leq \frac{1/\alpha}{1/\mathcal{A}} \frac{|||Ax|||}{|||x|||} \quad ,$$

ce qui entraîne successivement

$$\frac{|||Ax|||}{|||x|||} \leq \frac{\mathcal{A}}{\alpha} \sup_{y \neq 0} \frac{\|Ay\|}{\|y\|} \quad \text{et} \quad \frac{\|Ax\|}{\|x\|} \leq \frac{\mathcal{A}}{\alpha} \sup_{y \neq 0} \frac{|||Ay|||}{|||y|||}$$

et

$$\sup_{x \neq 0} \frac{|||Ax|||}{|||x|||} \leq \frac{\mathcal{A}}{\alpha} \sup_{y \neq 0} \frac{\|Ay\|}{\|y\|} \quad \text{et} \quad \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \leq \frac{\mathcal{A}}{\alpha} \sup_{y \neq 0} \frac{|||Ay|||}{|||y|||} \quad .$$

2.4 - Notons finalement que <u>dans R^n, toutes les normes sont équivalentes à la norme euclidienne [20]</u>. En effet, si $\|.\|$ est une norme quelconque,

a) On a

$$\|x\| = \| \sum_i \xi_i \, e_{(i)} \| ,$$

où $e_{(i)}$ est le i^e vecteur de base. Dès lors,

$$\|x\| \leq \sum_i |\xi_i| \, \|e_{(i)}\| \leq \sqrt{\sum_i |\xi_i|^2} \sqrt{\sum_i \|e_{(i)}\|^2} = \mathcal{A} \sqrt{\sum_i |\xi_i|^2}$$

$$= \mathcal{A} \, \|x_2\|_2 ,$$

où

$$\mathcal{A} = \sqrt{\sum_i \|e_{(i)}\|^2}$$

ne dépend que de la norme choisie.

b) Pour montrer que $\|x\|_2 \leq \alpha \|x\|$, supposons le contraire, c'est-à-dire que, quel que soit α , on puisse trouver un vecteur x tel que

$$\|x\| < \frac{1}{\alpha} \|x\|_2 .$$

Cela implique la possibilité de trouver, pour tout m, un vecteur $x_{(m)}$ tel que

$$\|x_{(m)}\| < \frac{1}{m} \|x_{(m)}\|_2 .$$

On peut évidemment supposer $\|x_{(m)}\|_2 = 1$. Mais alors, la suite $\{x_{(m)}\}$ étant bornée, on peut en extraire une sous-suite $\{x_{(m_k)}\}$ qui converge vers un certain vecteur y. Alors,

$$\|y\| = \lim \|x_{(m_k)}\| = 0 \quad \text{et} \quad \|y\|_2 = \lim \|x_{(m_k)}\|_2 = 1 ,$$

si bien que $y = 0$ et $\|y\|_2 = 1$, ce qui est contradictoire.

Cette propriété entraîne en particulier que <u>toutes les normes de matrices en tant qu'opérateurs de R^n sont équivalentes</u>.

3. CONVERGENCE DES METHODES ITERATIVES SIMPLES

3.2 - Nous commencerons par établir une <u>condition suffisante de convergence</u>: <u>Si, pour une définition quelconque de la norme, $\|B\| < 1$, la suite</u>

$$x^{(k+1)} = B \, x^{(k)} + c$$

<u>partant de $x^{(0)}$ quelconque admet une limite qui est la solution unique de l'équation</u>

$$x = B \, x + c .$$

Tout d'abord, la limite existe bien (c.-à.-d. dtm $(I - B) \neq 0$), car dans le cas contraire, il existerait un vecteur y non nul tel que

$$(I - B) \, y = 0 ,$$

soit

$$y = B\,y\ ,$$

ce qui entraînerait

$$\|y\| \leqslant \|B\|\ \|y\|$$

et

$$\|B\| \geqslant 1\ ,$$

en contradiction avec l'hypothèse.

Cela étant, les itérés successifs vérifient

$$x^{(k+1)} = B\,x^{(k)} + c\ ;$$

soustrayant à cette relation la définition de la limite,

$$x = B\,x + c$$

et notant

$$r^{(k)} = x^{(k)} - x\ ,$$

on obtient

$$r^{(k+1)} = B\,r^{(k)}\ ,$$

ce qui entraîne

$$\|r^{(k+1)}\| \leqslant \|B\|\ \|r^{(k)}\| \leqslant \ldots\ \leqslant \|B\|^{k+1}\|r^{(0)}\| \longrightarrow 0\ .$$

3.3 - Notion de rayon spectral d'une matrice

On appelle rayon spectral d'une matrice B sa plus grand valeur propre en module:

$$\rho(B) = |\lambda(B)|_{max}\ .$$

Le rayon spectral jouit de deux propriétés importantes vis-à-vis des normes matricielles:

a) Quelle que soit la norme matricielle utilisée, on a $\rho(B) \leqslant \|B\|$

En effet, si z est un vecteur propre associé à la plus grande valeur propre en module λ, on a

$$B\,z = \lambda z\ ,$$

ce qui entraîne

$$|\lambda|\ \|z\| \leqslant \|B\|\ \|z\|$$

soit

$$\rho(B) \leqslant \|B\|\ .$$

b) Si $\rho(B) \leqslant q$ et si les valeurs propres éventuelles de module q ont chacune un nombre de vecteurs propres égal à leur multiplicité (ce qui a lieu, notamment, si elles sont simples), il existe une matrice inversible S telle que $\|S^{-1}\,B\,S\|_\infty \leqslant q$ [8]

Démontrons d'abord cette propriété dans le cas où B est diagonalisable. Alors, en choisissant

$$S = [\, z_{(1)}, \ldots, z_{(n)} \,] \quad ,$$

où $z_{(i)}$ sont les vecteurs propres, on a

$$S^{-1} B\, S = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n) \quad ,$$

d'où, à l'évidence,

$$\| S^{-1} B\, S \| = |\lambda(B)|_{\max} = \rho(B) \leq q \quad .$$

* Passons à présent au cas où <u>B n'est pas diagonalisable</u>. Posons

$$\varepsilon = \tfrac{1}{2}\Big(q - \sup_{|\lambda_i| < q} |\lambda_i|\Big).$$

Il existe une matrice S transformant $\frac{1}{\varepsilon} B$ en forme canonique de JORDAN:

$$S^{-1} \frac{1}{\varepsilon} B\, S = \begin{bmatrix} \frac{\lambda_1}{\varepsilon} & \beta_1 & & \\ & \frac{\lambda_2}{\varepsilon} & \beta_2 & \\ & & \ddots & \ddots \\ & & & \frac{\lambda_{n-1}}{\varepsilon} \quad \beta_{n-1} \\ & & & \frac{\lambda_n}{\varepsilon} \end{bmatrix} \quad ,$$

avec $\beta_i = 1$ ou 0. Aux valeurs propres λ_i de module q correspondent $\beta_i = 0$, car leur sous-matrice carrée de Jordan est diagonale du fait que ces valeurs propres possèdent autant de vecteurs propres que leur multiplicité. Dès lors,

$$S^{-1} B\, S = \begin{bmatrix} \lambda_1 & \varepsilon\beta_1 & & \\ & \lambda_2 & \varepsilon\beta_2 & \\ & & \ddots & \ddots \\ & & & \lambda_{n-1} \quad \varepsilon\beta_{n-1} \\ & & & \lambda_n \end{bmatrix}$$

et on a, pour $|\lambda_i| = q$,

$$\sum_j |(S^{-1} B S)_{ij}| = q \quad ,$$

tandis que pour $|\lambda_i| \neq q$,

$$\sum_j |(S^{-1} B S)_{ij}| = |\lambda_i| + \varepsilon\beta_i < q \quad ,$$

donc

$$\| S^{-1} B S \|_\infty \leq q \quad ,$$

ce qui achève la démonstration. *

3.3 - Condition nécessaire et suffisante de convergence

Si l'équation $x = Bx + c$ admet une solution unique, le processus itératif

$$x^{(k+1)} = B\,x^{(k)} + c$$

converge si et seulement si $\rho(B) < 1$.

La condition est nécessaire. Supposons en effet qu'il existe une valeur propre λ telle que $|\lambda| \geqslant 1$, et soit z un vecteur propre associé à cette valeur propre. Alors, si l'on choisit comme point de départ

$$x^{(0)} = x + z ,$$

où x est la solution, on a

$$\begin{cases} x^{(k+1)} = B\,x^{(k)} + c \\ x = B\,x + c . \end{cases}$$

Faisant la différence et notant $r^{(k)} = x^{(k)} - x$, on obtient

$$r^{(k+1)} = B\,r^{(k)} = \ldots\ldots = B^{k+1}\,r^{(0)} = \lambda^{k+1}\,r^{(0)}$$

ce qui entraîne

$$\|r^{(k+1)}\| = |\lambda|^{k+1}\,\|z\| \not\to 0 .$$

La condition est suffisante. Soit q tel que

$$|\lambda|_{max} < q < 1 .$$

Il existe une matrice S telle que

$$\|S^{-1}BS\|_\infty \leqslant q .$$

Posons

$$C = S^{-1}\,B\,S .$$

On a évidemment

$$B = S\,C\,S^{-1}$$

et

$$\|C\|_\infty \leqslant q .$$

Par ailleurs, on montre comme ci-dessus que la différence $r^{(k)}$ entre $x^{(k)}$ et la solution vérifie

$$r^{(k+1)} = B\,r^{(k)} = \ldots\ldots = B^{k+1}\,r^{(0)} .$$

Or,

$$B^{(k)} = \underbrace{SCS^{-1}\,.\,SCS^{-1}\,.\,\ldots\ldots\,.\,SCS^{-1}}_{(k+1)\text{ facteurs}} = S\,C^{k+1}\,S^{-1},$$

ce qui entraîne

$$\|r^{(k+1)}\|_\infty \leqslant \|S\|_\infty\,\|S^{-1}\|_\infty\;q^{k+1}\,\|r^{(0)}\| \longrightarrow 0 .$$

La convergence est donc plus rapide que celle de toute progression géométrique de raison q supérieure à $\rho(B)$.

4. NOTION DE TAUX DE CONVERGENCE

4.1 - Taux de convergence moyen pour k itérations

Pour caractériser la vitesse de convergence, il est naturel de comparer le résidu $r^{(k)}$ au résidu de départ $r^{(0)}$. Comme

$$r^{(k)} = B^k r^{(0)} ,$$

on a

$$\|r^{(k)}\| \leq \|B^k\| \, \|r^{(0)}\| .$$

On peut donc dire qu'en moyenne, chaque itération a divisé la norme du résidu dans le rapport

$$\tau_k = \|B^k\|^{1/k}$$

que l'on appelle taux de convergence moyen pour k itérations.

Cette notion est assez peu satisfaisante pour les raisons suivantes:

a) Le taux moyen dépend de la norme choisie. Il faut donc distinguer $\tau_{k,1}$, $\tau_{k,2}$, $\tau_{k,\infty}$.

b) En général, le taux moyen dépend de k. Cependant, si la matrice B est normale (i.e. $B^T B = B B^T$), elle admet une base de vecteurs propres orthonormés $z_{(1)}, \ldots, z_{(n)}$ auxquels correspondent les valeurs propres $\lambda_1, \ldots, \lambda_n$, que nous supposerons rangées par ordre des modules décroissants. Alors,

$$B^p z_{(i)} = \lambda_i^p z_{(i)}$$

et pour un vecteur quelconque

$$y = \sum_i \alpha_i z_{(i)} ,$$

on obtient aisément

$$\|B^p y\|_2^2 = \sum_i |\lambda_i|^{2p} |\alpha_i|^2 \leq |\lambda_1|^{2p} \sum_i |\alpha_i|^2 = |\lambda_1|^{2p} \|y\|_2^2$$

cette borne pouvant être atteinte dans le cas $y = z_{(1)}$, ce qui implique

$$\|B^p\|_2 = |\lambda_1|^p = \rho^p(B) .$$

On a donc, pour une matrice normale,

$$\tau_{k,2} = \rho(B) .$$

Par contre, $\tau_{k,1}$ et $\tau_{k,\infty}$ peuvent varier fortement avec k.

Pour une matrice quelconque, on n'a aucun résultat de ce type.

4.2 - Taux de convergence asymptotique

Il se trouve qu'à la limite, toutes les définitions du taux de

convergence se réconcilient: quelle que soit la définition de la norme des matrices en tant qu'opérateurs, on a la propriété

$$\lim_{k \to \infty} \|B^k\|^{1/k} = \rho(B) .$$

Démonstration : nous montrerons que $\rho \leq \lim_{k \to \infty} \|B^k\|^{1/k} \leq \rho$.

a) On a évidemment $\|B^k\| \geq \rho^k$. En effet, si z est un vecteur propre relatif à une valeur propre λ de module ρ , on a

$$B^k z = \lambda^k z ,$$

ce qui entraîne

$$|\lambda^k| \|z\| \leq \|B^k\| \|z\|$$

soit

$$\|B^k\| \geq \rho^k .$$

b) Pour tout $\varepsilon > 0$, on a $\rho < \rho + \varepsilon$, donc il existe une matrice inversible S telle que

$$\|S^{-1}BS\|_\infty < \rho + \varepsilon .$$

Alors, par l'équivalence des normes de R^n ,

$$\|B^k\| \leq \mathcal{A} \|B^k\|_\infty = \mathcal{A} \|SS^{-1}B^kSS^{-1}\|_\infty \leq \mathcal{A} \|S\|_\infty \|S^{-1}\|_\infty \|S^{-1}B^kS\|_\infty$$

et

$$\|S^{-1} B^k S\|_\infty = \|(S^{-1}BS)^k\|_\infty \leq \|S^{-1}BS\|_\infty^k < (\rho + \varepsilon)^k ,$$

si bien que

$$\|B^k\|^{1/k} \leq (\mathcal{A} \|S\|_\infty \|S^{-1}\|_\infty)^{1/k} (\rho + \varepsilon) = (\rho + \varepsilon) f(k) ,$$

avec

$$\lim_{k \to \infty} f(k) = 1 ,$$

puisque pour tout nombre positif θ , on a $\theta^{1/k} \to 1$.

c) Rassemblant les deux résultats précédents, on obtient que, pour tout $\varepsilon > 0$,

$$\rho \leq \|B^k\|^{1/k} \leq (\rho + \varepsilon) f(k) ,$$

avec $f(k) \to 1$, ce qui implique

$$\lim_{k \to \infty} \|B^k\|^{1/k} = \rho . \quad \text{C.Q.F.D.}$$

Cette propriété permet de définir le taux de convergence asymptotique comme

$$\tau = \lim_{k \to \infty} \|B^k\|^{1/k} = \rho(B) .$$

Grâce à cette notion , il est possible de comparer deux méthodes itératives simples quant à leur vitesse de convergence. Plus petit est τ, meilleur est l'algorithme.

5. METHODE DE JACOBI

Soit à résoudre le système

$$\begin{cases} a_{11}\, x_1 + \dots + a_{1n}\, x_n = b_1 \\ \quad \dots\dots\dots \\ a_{n1}\, x_1 + \dots + a_{nn}\, x_n = b_n \end{cases}.$$

On écrit chaque équation sous la forme

$$a_{ii}\, x_i^{(k+1)} + \sum_{j \neq i} a_{ij}\, x_j^{(k)} = b_i \quad ,$$

ce qui entraîne

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}\, x_j^{(k)} \right) .$$

Cette méthode simple ne s'applique évidemment que si aucun terme diagonal n'est nul. Pour en étudier la convergence, notons que cette méthode revient à partitionner A en

$$A = \left[\begin{array}{c} \diagdown\!\!\diagdown \\ \diagdown\!\!\diagdown \end{array} \right. \begin{array}{l} U \\ D \\ L \end{array} \left. \right]$$

et à écrire en conséquence le système sous la forme

$$L\, x + D\, x + U\, x = b \quad ,$$

puis à itérér sous la forme

$$D\, x^{(k+1)} = b - L\, x^{(k)} - U\, x^{(k)} \quad ,$$

soit

$$x^{(k+1)} = D^{-1} \left(b - (L + U)\, x^{(k)} \right) .$$

La condition de convergence est donc

$$\rho \left(-D^{-1} (L + U) \right) < 1 .$$

Or, l'équation caractéristique s'écrit

$$\mathrm{dtm} \left(-D^{-1} (L + U) - \lambda I \right) = 0 \quad ,$$

soit, en supposant, comme il se doit, $\mathrm{dtm}\, D = a_{11} \dots a_{nn} \neq 0$,

$$\mathrm{dtm}(D^{-1}) \,.\, \mathrm{dtm} \left(-(L + U) - \lambda D \right) = 0 .$$

Il faut donc que toutes les solutions de l'équation

$$(-1)^n \begin{vmatrix} \lambda a_{11} & a_{12} & \dots\dots & a_{1n} \\ a_{21} & \lambda a_{22} & \dots\dots & a_{2n} \\ \dots\dots & & & \\ a_{n1} & a_{n2} & \dots\dots & \lambda a_{nn} \end{vmatrix} = 0$$

soient inférieures en module à l'unité. Cette condition étant malaisée

d'emploi, on la remplace souvent par la condition <u>suffisante</u>

$$\|B\| < 1 \quad ,$$

avec

$$B = \begin{bmatrix} 0 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & \cdots & \frac{a_{2n}}{a_{22}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{bmatrix} \quad ,$$

ce qui donne

$$\sup_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1 \ .$$

Une matrice vérifiant cette condition, que l'on peut encore écrire

$$|a_{ii}| < \sum_{j \neq i} |a_{ij}| \quad , \quad i = 1, \ldots, n$$

est dite <u>à diagonale dominante</u>. Cette condition est vérifiée notamment par les matrices de capacité en électrostatique.

6. METHODE DE GAUSS-SEIDEL

C'est une transformation de la méthode de JACOBI: lors du calcul de $x_i^{(k+1)}$, on utilise déjà les nouvelles valeurs $x_j^{(k+1)}$, $j < i$, déjà calculées. En d'autres termes, on effectue

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right] .$$

Ceci permet de ne stocker en mémoire qu'un vecteur-solution au lieu de deux dans la méthode de Jacobi. Matriciellement, cela revient à écrire, en utilisant la partition (L,D,U) définie plus haut,

$$D\, x^{(k+1)} = b - L\, x^{(k+1)} - U\, x^{(k)}$$

c'est-à-dire

$$(D + L)x^{(k+1)} = b - U\, x^{(k)}$$

ou encore

$$x^{(k+1)} = (D + L)^{-1} (b - U\, x^{(k)}) \ .$$

La condition de convergence s'écrit donc

$$\rho(- (D + L)^{-1} U) < 1 \ .$$

L'équation caractéristique correspondante,

$$\mathrm{dtm} (- (D + L)^{-1} U - \lambda I) = 0$$

équivaut, si dtm $D = a_{11} \cdots a_{1n} \neq 0$ (condition nécessaire d'applicabilité de la méthode), à

$$(-1)^n \text{ dtm } (U + \lambda (D + L)) = 0 ,$$

soit explicitement

$$\begin{vmatrix} \lambda a_{11} & a_{12} & a_{13} \cdots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & a_{23} \cdots & a_{2n} \\ \cdots\cdots\cdots & & & \\ \lambda a_{n1} & \lambda a_{n2} & \lambda a_{n3} & \lambda a_{nn} \end{vmatrix} = 0$$

Cette condition est très lourde à mettre en oeuvre. Mais ici encore, la condition de dominance diagonale

$$\sup_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} = q < 1$$

est <u>suffisante</u>. En effet, on a

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right)$$

et

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j - \sum_{j>i} a_{ij} x_j \right) ,$$

ce qui entraîne, en notant $r_i^{(k)} = x_i^{(k)} - x_i$,

$$r_i^{(k+1)} = - \sum_{j<i} \frac{a_{ij}}{a_{ii}} r_j^{(k+1)} - \sum_{j>i} \frac{a_{ij}}{a_{ii}} r_j^{(k)} .$$

Pour la <u>première ligne</u>, il vient

$$|r_1^{(k+1)}| \leq \sum_{j>1} \left| \frac{a_{1j}}{a_{11}} \right| |r_j^{(k)}| \leq \sum_{j>1} \left| \frac{a_{1j}}{a_{11}} \right| \|r^{(k)}\|_\infty \leq q \|r^{(k)}\|_\infty ;$$

par <u>récurrence</u>, si

$$|r_j^{(k+1)}| \leq q \|r^{(k)}\|_\infty$$

pour $j<i$, il en est de même pour $|r_i^{(k+1)}|$, car

$$|r_i^{(k+1)}| \leq \sum_{j<i} \left| \frac{a_{ij}}{a_{ii}} \right| \left| r_j^{(k+1)} \right| + \sum_{j>i} \left| \frac{a_{ij}}{a_{ii}} \right| |r_j^{(k)}|$$

$$\sum_{j<i} \left| \frac{a_{ij}}{a_{ii}} \right| q \|r^{(k)}\|_\infty + \sum_{j>i} \left| \frac{a_{ij}}{a_{ii}} \right| \|r^{(k)}\|_\infty$$

$$\sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| \|r^{(k)}\|_\infty \leq q \|r^{(k)}\|_\infty .$$

Au total,

$$\|r^{(k+1)}\|_\infty \leq q \|r^{(k)}\|_\infty .$$

7. METHODE GENERALE $x^{(k+1)} = x^{(k)} + \alpha(A\, x^{(k)} - b)$

Examinons dans quelles conditions l'algorithme

$$x^{(k+1)} = x^{(k)} + \alpha\,(A\, x^{(k)} - b)$$

converge. Il est clair que la limite, si elle existe, vérifie

$$x = x - \alpha(A\, x - b)\ ,$$

soit

$$A\, x = b\ .$$

L'erreur vectorielle $r^{(k)} = x^{(k)} - x$ vérifie visiblement la récurrence

$$r^{(k+1)} = r^{(k)} - \alpha\, A\, r^{(k)} = (I - \alpha A)\, r^{(k)}\ ,$$

et la condition de convergence est donc

$$\rho(I - \alpha A) < 1\ .$$

L'équation caractéristique de cette matrice est

$$\mathrm{dtm}\ (\ (1 - \Lambda)\, I - \alpha\, A) = 0\ ,$$

soit

$$(-1)^n\, \alpha^n\ \mathrm{dtm}\ (A - \frac{1 - \Lambda}{\alpha}\, I) = 0\ .$$

Les valeurs propres Λ_i de $(I - \alpha A)$ sont donc liées aux valeurs propres λ_i de A par l'équation

$$\frac{1 - \Lambda_i}{\alpha} = \lambda_i\ ,$$

ce qui revient à dire

$$\Lambda_i = 1 - \alpha\lambda_i\ .$$

On ne peut donc avoir $|\Lambda_i| < 1$ pour tout i que si toutes les valeurs propres de A (supposées réelles) ont le même signe. Supposons donc toutes les valeurs propres <u>positives</u>, et admettons que l'on ait en outre une estimation de la forme

$$0 < \mu \leq \lambda_i \leq M\ .$$

Le taux de convergence asymptotique vérifie alors

$$\rho(\alpha) = \sup_{\lambda_i} |1 - \alpha\lambda_i| \qquad \sup_{\mu \leq \lambda \leq M} |1 - \alpha\lambda| = \tilde{\rho}(\alpha)\ .$$

Nous désirons choisir α de manière à rendre ce nombre aussi petit que possible. Pour $\alpha > 0$, on a

$$1 - \alpha\mu \geq 1 - \alpha\lambda \geq 1 - \alpha M\ .$$

Pour $\alpha < 1/M$, les deux bornes décroissent pour a croissant et on a $\tilde{\rho}(\alpha)\downarrow$.

Mais pour $\alpha > 1/M$, $(1 - \alpha M)$ devient négatif et son module croît à nouveau, tandis que celui de $(1 - \alpha\mu)$ décroît tant que $\alpha < 1/\mu$. La figure 1 montre clairement que le plus grand des deux modules est minimum pour

$$1 - \alpha\mu = -(1 - \alpha M)$$

soit

$$\alpha (M + \mu) = 2$$

A partir de cette valeur, si α croît, $|1 - \alpha M|$ croît aussi et le maximum s'accroît ; si α diminue, $(1 - \alpha\mu)$ croît, et le maximum s'accroît également. Donc, la valeur optimale est bien

$$\alpha_{opt} = \frac{2}{M + \mu} \quad .$$

En ce point,

$$\tilde{\rho}(\alpha_{opt}) = 1 - \frac{2}{M + \mu} = \frac{M - \mu}{M + \mu} = \frac{1 - \frac{\mu}{M}}{1 + \frac{\mu}{M}} \quad .$$

La vitesse de convergence dépend donc du rapport entre la plus petite et la plus grande valeur propre de la matrice.

8. METHODES DE MINIMISATION

Lorsque la matrice A est symétrique définie positive, la fonction

$$\phi(x) = \tfrac{1}{2}\, x^T A\, x - x^T b$$

admet un point stationnaire lorsque

$$\text{grad}\ \phi = A\, x - b = 0 \ .$$

Il s'agit d'un <u>minimum</u>, car

$$\sum_{ij} \frac{\partial^2 \phi}{\partial x_i\, \partial x_j}\, h_i\, h_j = h^T A\, h > 0 \quad .$$

On peut donc, au lieu de résoudre directement le système matriciel, procéder par minimisation progressive de la fonction ϕ. Le principe des méthodes de minimisation est de se donner une direction de progression $h^{(k)}$ à partir du vecteur $x^{(k)}$, c'est-à-dire de chercher une meilleure approximation de la forme

$$x^{(k+1)} = x^{(k)} + \beta_k\, h^{(k)}$$

et de choisir β_k pour que $\phi(x^{(k+1)})$ soit aussi petit que possible. Notant

$$g^{(k)} = \text{grad}\ \phi(x^{(k)}) = A\, x^{(k)} - b \ ,$$

on a

$$\phi(x^{(k)} + \beta_k h^{(k)}) = \phi(x^{(k)}) + \beta_k h^{(k)T} g^{(k)} + \tfrac{1}{2} \beta_k^2 h^{(k)T} A h^{(k)},$$

et cette expression atteint son minimum lorsque

$$h^{(k)T} \left[g^{(k)} + \sum_k A h^{(k)} \right] = 0 ,$$

condition qui signifie que l'on s'arrête lorsque la direction de progression devient perpendiculaire au gradient .

$$g^{(k+1)} = g^{(k)} + \beta_k h^{(k)}$$

et qui entraîne

$$\beta_k = - \frac{h^{(k)T} g^{(k)}}{h^{(k)T} A h^{(k)}} .$$

Cette valeur permet de calculer

$$x^{(k+1)} = x^{(k)} + \beta_k h^{(k)}$$

et de recommencer.

Chaque itération diminue ϕ, pour autant que $\beta_k \neq 0$. En effet,

$$\phi(x^{(k+1)}) = \phi(x^{(k)}) - \frac{(h^{(k)T} g^{(k)})^2}{h^{(k)} A h^{(k)}} + \tfrac{1}{2} \frac{(h^{(k)T} g^{(k)})^2}{h^{(k)T} A h^{(k)}}$$

$$= \phi(x^{(k)}) - \tfrac{1}{2} \frac{(h^{(k)T} g^{(k)})^2}{h^{(k)T} A h^{(k)}} < \phi(x^{(k)})$$

puisque la matrice A est s.d.p.. Or, cela signifie que l'on se rapproche du minimum, car

$$\|x^{(k)} - x\|_2^2 \leq \frac{1}{\lambda_{min}(A)} (x^{(k)T} - x^T) A (x^{(k)} - x)$$

$$\leq \frac{1}{\lambda_{min}(A)} (x^{(k)T} A x^{(k)} - 2 x^T A x^{(k)} + x^T A x) .$$

$$\leq \frac{1}{\lambda_{min}(A)} (x^T A x - 2 x^{(k)T} b + x^{(k)T} A x^{(k)})$$

$$\leq \frac{1}{\lambda_{min}(A)} (2 \phi(x^{(k)}) + x^T A x)$$

Notant que pour la solution x,

$$\phi(x) = \tfrac{1}{2} x^T A x - x^T b = - \tfrac{1}{2} x^T A x + x^T (A x - b) = - \tfrac{1}{2} x^T A x ,$$

il vient finalement

$$\|x^{(k)} - x\|_2^2 \quad \frac{2}{\lambda_{min}(A)} (\phi(x^{(k)}) - \phi(x)) .$$

Dès lors, si le choix des $h^{(k)}$ est conçu de telle sorte que l'on ne puisse jamais avoir $h^{(k)T}$ grad $\phi = 0$ (avec grad $\phi \neq 0$) pour plus d'un nombre fini s d'itérations, s fixé d'avance, la convergence est assurée, puisque tant que l'on n'est pas arrivé à la solution, on devra se déplacer après s itérations au plus. La suite des $\phi(x^{(k)})$ est donc décroissante et bornée, et converge vers sa borne inférieure, ce qui implique $x^{(k)} \longrightarrow x$.

9. METHODE DE GAUSS-SEIDEL POUR LES MATRICES DEFINIES POSITIVES

Choisissant pour direction de progression

$$h^{(k)} = e_{(k)} \quad (k^e \text{ axe, modulo n}) ,$$

on obtient

$$\beta_k = - \frac{e_{(k)}^T g^{(k)}}{e_{(k)}^T A e_{(k)}} = - \frac{1}{a_{kk}} \left(b_k - \sum_l a_{kl} x_l^{(k)} \right) ,$$

ce qui donne

$$x^{(k+1)} = x^{(k)} - \frac{1}{a_{kk}} (b_k - \sum_l a_{kl} x_l^{(k)}) e_{(k)}$$

soit

$$\begin{cases} x_l^{(k+1)} = x_l^{(k)} , & l \neq k \text{ mod. n} \\ x_k^{(k+1)} = x_k^{(k)} - \frac{1}{a_{kk}} (b_k - \sum_l a_{kl} x_l^{(k)}) \end{cases}$$

Ce n'est rien d'autre que la méthode de Gauss-Seidel. La convergence est assurée, car au bout d'un cycle de n itérations, on aura certainement obtenu $\beta_k \neq 0$, tant que grad $\phi(x^{(k)}) \neq 0$.

* Montrons que la convergence se fait en progression géométrique. A cette fin, notons B la triangulaire inférieure de A (diagonale comprise) et C le reste de la matrice:

$$A = \left[\begin{array}{c} B \quad C \end{array} \right]$$

et considérons la forme complète en n itérations d'un coup:

$$x^{(k+1)} = B^{-1} (b - C x^{(k)}) .$$

Comme on ne peut avoir $g^{(k)T} e_{(i)}$ pour tous les axes, on a nécessairement, en notant $r^{(k)} = x^{(k)} - x$, l'inégalité

$$\frac{r^{(k+1)T} A r^{(k+1)}}{r^{(k)T} A r^{(k)}} < 1 ,$$

soit

$$\psi(r^{(k)}) = \frac{r^{(k)T} C^T B^{-T} A B^{-1} C r^{(k)}}{r^{(k)T} A r^{(k)}} < 1$$

Cette relation est vraie quel que soit le résidu $r^{(k)}$. Sur la sphère $\|r\|_2 = 1$, qui est compacte, la fonction continue $\psi(r)$ atteint sa borne supérieure $\xi^2 < 1$. Comme ψ est homogène de degré zéro, on a toujours

$$(r^{(k)}) = \frac{r^{(k+1)T} A r^{(k+1)}}{r^{(k)T} A r^{(k)}} \leq \xi^2$$

et, de proche en proche,

$$r^{(k)T} A r^{(k)} \leq \xi^{2k} r^{(0)T} A r^{(0)} .$$

Il en découle

$$\|r^{(k)}\|_2^2 \leq \frac{1}{\lambda_{min}(A)} r^{(k)T} A r^{(k)} \leq \frac{\xi^{2k}}{\lambda_{min}(A)} r^{(0)T} A r^{(0)}$$

$$\xi^{2k} \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \|r^{(0)}\|_2^2 ,$$

soit

$$\|r^{(k)}\|_2 \leq \xi^k \left(\frac{\lambda_{max}(A)}{\lambda_{min}(A)}\right)^{\frac{1}{2}} \|r^{(0)}\|_2 \qquad *$$

10. METHODE DE LA PLUS GRANDE PENTE

Puisqu'il convient d'assurer aussi souvent que possible la relation

$$h^{(k)T} g^{(k)} \neq 0,$$

une solution évidente est de progresser dans la direction du gradient, c'est-à-dire de la plus grande pente. On écrit donc

$$x^{(k+1)} = x^{(k)} + \beta_k g^{(k)} = x^{(k)} + \beta_k (A x^{(k)} - b) ,$$

avec

$$\beta_k = - \frac{g^{(k)T} g^{(k)}}{g^{(k)T} A g^{(k)}}$$

La convergence est assurée. Pour évaluer la vitesse de co vergence, notons que, comme

$$x^{(k+1)} = x^{(k)} + \beta_k (A x^{(k)} - b)$$

et

$$x = x + \beta_k (A x - b),$$

on obtient

$$r^{(k+1)} = r^{(k)} + \beta_k A r^{(k)} .$$

Développons alors $r^{(k)}$ dans la base des vecteurs propres orthonormés $z_{(1)}, \ldots, z_{(n)}$, ordonnés de façon que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Pour

$$r^{(k)} = \sum_i \alpha_i z_{(i)} ,$$

on obtient

$$r^{(k+1)} = \sum_i \alpha_i \, (1 + \beta_k \lambda_i) \, z_{(i)} \, .$$

Considérons d'abord une valeur quelconque β de β_k. Il y correspond le vecteur $\tilde{r}^{(k+1)}$. Il vient

$$\tilde{r}^{(k+1)T} \, A \, \tilde{r}^{(k+1)} = \sum_i \alpha_i^2 \lambda_i \, (1 + \beta \lambda_i)^2$$

$$\leqslant \sup_i (1 + \beta \lambda_i)^2 \sum_i \alpha_i^2 \lambda_i = \sup_i (1 + \beta \lambda_i)^2 \, r^{(k)T} A r^{(k)} \, .$$

En choisissant la valeur optimale de la section 7, à savoir,

$$\beta = - \frac{2}{\lambda_1 + \lambda_n} \, ,$$

on obtient

$$\sup_i (1 + \beta \lambda_i) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}$$

et, par conséquent,

$$\tilde{r}^{(k+1)T} \, A \, \tilde{r}^{(k+1)} \leqslant \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^2 \, r^{(k)T} \, A \, r^{(k)} \, .$$

Or, la valeur de β_k est précisément choisie pour minimiser

$$\tfrac{1}{2} \, x^{(k+1)T} \, A \, x^{(k+1)} - b^T \, x^{(k+1)} = \tfrac{1}{2} \, x^{(k+1)T} \, A \, x^{(k+1)} - x^{(k+1)T} \, A \, x$$

$$= \tfrac{1}{2} \, r^{(k+1)T} \, A \, r^{(k+1)} - \tfrac{1}{2} \, x^T \, A \, x \, ,$$

donc

$$r^{(k+1)T} \, A \, r^{(k+1)} \leqslant \tilde{r}^{(k+1)T} \, A \, \tilde{r}^{(k+1)} \leqslant \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^2 \, r^{(k)T} \, A \, r^{(k)}$$

et, de proche en proche,

$$r^{(k)T} \, A \, r^{(k)} \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^{2k} \, r^{(0)T} \, A \, r^{(0)} \, .$$

Finalement, tenant compte de l'encadrement général

$$\lambda_n \, \|y\|_2^2 \leqslant y^T \, A \, y \leqslant \lambda_1 \, \|y\|_2^2 \, ,$$

on obtient

$$\|r^{(k)}\|_2 \leqslant \sqrt{\frac{\lambda_1}{\lambda_n}} \, \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^k \, \|r^{(0)}\|_2 \, .$$

La convergence de la méthode de la plus grande pente est donc du même ordre que celle de l'algorithme optimisé de la section 7, mais avec l'énorme avantage qu'il n'est pas nécessaire d'évaluer les valeurs propres de la matrice.

11. METHODE DU GRADIENT CONJUGUE

11.1 - La méthode du gradient conjugué est une amélioration de la précédente. Au lieu de choisir comme direction de progression

$$h^{(k)} = g^{(k)} ,$$

on s'impose un $h^{(k)}$ de la forme

$$h^{(k)} = g^{(k)} + \sigma_k h^{(k-1)} ,$$

le coefficient σ_k étant <u>choisi lui-même pour minimiser</u> $\phi(x^{(k+1)})$. Pour un σ_k donné, le choix optimal de β_k conduit à

$$\phi(x^{(k+1)}) = \phi(x^{(k)}) - \tfrac{1}{2} \frac{(h^{(k)T} g^{(k)})^2}{h^{(k)T} A h^{(k)}}$$

$$= \phi(x^{(k)}) - \tfrac{1}{2} \frac{\left[(g^{(k)T} + \sigma_k h^{(k-1)T}) g^{(k)} \right]^2}{g^{(k)T} A g^{(k)} + 2 \sigma_k g^{(k)T} A h^{(k-1)} + \sigma_k^2 h^{(k-1)T} A h^{(k-1)}}$$

et comme $h^{(k-1)T} g^{(k)} = 0$, on obtient

$$\phi(x^{(k+1)}) = \phi(x^{(k)}) - \tfrac{1}{2} \frac{(g^{(k)T} g^{(k)})^2}{g^{(k)T} A g^{(k)} + 2 \sigma_k g^{(k)T} A h^{(k-1)} + \sigma_k^2 h^{(k-1)T} A h^{(k-1)}}$$

Le minimum de cette expression s'obtient en minimisant le dénominateur du second terme, ce qui mène à la condition

$$\sigma_k h^{(k-1)T} A h^{(k-1)} + g^{(k)T} A h^{(k-1)}$$

ou encore,

$$\sigma_k = - \frac{h^{(k-1)T} A g^{(k)}}{h^{(k-1)T} A h^{(k-1)}} .$$

Cette condition signifie en fait que la nouvelle direction de progression $h^{(k)}$ vérifie, par rapport à l'ancienne $h^{(k-1)}$, la <u>condition de conjugaison</u>

$$h^{(k-1)T} A h^{(k)} = 0.$$

Que faut-il entendre par là? Par le point $x^{(k)}$ passe le n-ellipsoïde d'équation

$$x^T A x - 2 b^T x = x^{(k)T} A x^{(k)} - 2 b^T x^{(k)} ;$$

les directions $h^{(k-1)}$ et $g^{(k)}$ définissent une variété linéaire à deux dimensions passant par $x^{(k)}$. La section du n-ellipsoïde par la variété en question est une ellipse (fig. 2). La direction $h^{(k-1)}$ est tangante à cette ellipse et $h^{(k)}$, qui lui est conjuguée, est donc dirigée vers le centre de cette ellipse.

Calculons explicitement β_k: on a

$$\beta_k = - \frac{h^{(k)T} g^{(k)}}{h^{(k)T} A h^{(k)}} = - \frac{g^{(k)T} [g^{(k)} + \sigma_k h^{(k-1)}]}{h^{(k)T} A h^{(k)}} = - \frac{\| g^{(k)} \|_2^2}{h^{(k)T} A h^{(k)}}$$

Le calcul de σ_k par la formule donnée ci-dessus serait assez long. On peut l'écourter à partir de la propriété

$$g^{(k+1)T} g^{(k)} = 0$$

qui se vérifie à chaque itération. En effet, on a

$$g^{(k+1)T} g^{(k)} = g^{(k+1)T} (h^{(k)} - \sigma_k h^{(k-1)}) = - \sigma_k g^{(k+1)T} h^{(k-1)}$$

du fait de l'orthogonalité de $g^{(k+1)}$ et $h^{(k)}$; de plus,

$$h^{(k-1)T} g^{(k+1)} = h^{(k-1)T} (g^{(k)} + A h^{(k)}) = 0,$$

en vertu de la condition d'orthogonalité du gradient et de la condition de conjugaison.

Cela étant, le numérateur de $(- \sigma_k)$ s'écrit encore

$$h^{(k-1)T} A g^{(k)} = g^{(k)T} A \left[\frac{x^{(k)} - x^{(k-1)}}{\beta_{k-1}} \right]$$

$$= \frac{1}{\beta_{k-1}} g^{(k)T} (g^{(k)} - g^{(k-1)}) = \frac{\| g^{(k)} \|_2^2}{\beta_{k-1}} ,$$

et son dénominateur se transforme, en tenant compte de la formule de β_k, en

$$h^{(k-1)T} A h^{(k-1)} = - \frac{\| g^{(k-1)} \|_2^2}{\beta_{k-1}} .$$

Finalement,

$$\sigma_k = \frac{\| g^{(k)} \|_2^2}{\| g^{(k-1)} \|_2^2}$$

Pour la première itération, on utilise comme direction de progression celle du gradient.

Dénombrons les opérations nécessaires pour une itération du gradient conjugué qui ne soit pas la première:

- Calcul de $A x^{(k)}$: n^2 multiplications..... n^2 O.S.
- Calcul de $g^{(k)} = A x^{(k)} - b$: n soustractions
- Calcul de σ_k: on suppose $\| g^{(k-1)} \|_2^2$ conservé et il suffit de calculer $\| g^{(k)} \|_2^2$ (n multiplications), puis de diviser par $\| g^{(k-1)} \|_2^2$ .....(n - 1) O.S.
- Calcul de $h^{(k)}$: n multiplications..... n O.S.
- Calcul de $A h^{(k)}$ (n^2 mult.), puis calcul de $h^{(k)T} A h^{(k)}$ (n multiplications), puis division

de $\|g^{(k)}\|_2^2$ par le produit (1 division) .....$n^2 + n + 1$ O. S.

. Calcul de $x^{(k)}$: n multiplications pour $\beta_k\ h^{(k)}$.....n O.S.

Total : $2\,n^2 + 4\,n + 2 \approx 2\,n^2$

C'est le double d'un pivotage de Gauss-Jordan et le sextuple du travail moyen par inconnue d'une résolution par la méthode de Gauss. La méthode du gradient conjugué ne sera donc performante que si elle converge en un très petit nombre d'itérations, c'est-à-dire de l'ordre de n/6, pour concurrencer la méthode de Gauss.

* 11.2 - Propriétés particulières de l'algorithme du gradient conjugué.

Voici une autre présentation de l'algorithme du gradient conjugué, en tant que méthode directe. Nous nous appuierons sur deux lemmes.

Lemme 1 - Si la matrice symétrique définie positive A possède p valeurs propres différentes, et si g est un vecteur quelconque, $A^p\ g$ est une combinaison linéaire de $g, A\,g, \ldots, A^{p-1}g$.

Appelons en effet $z_{(1)}$ la projection de g dans le sous-espace propre relatif à la valeur propre λ_1, $z_{(2)}$ sa projection dans le sous-espace propre relatif à la valeur propre λ_2, ..., et $z_{(p)}$ sa projection dans le sous-espace relatif à λ_p. On a donc

$$g = z_{(1)} + z_{(2)} + \ldots + z_{(p)}$$

et, par des multiplications successives par A,

$$Ag = \lambda_1 z_{(1)} + \ldots + \lambda_p z_{(p)}$$

$$\ldots\ldots\ldots$$

$$A^{p-1}g = \lambda_1^{p-1} z_{(1)} + \ldots + \lambda_p^{p-1} z_{(p)}$$

$$A^p g = \lambda_1^p z_{(1)} + \ldots + \lambda_p^p z_{(p)} .$$

Ainsi, g, Ag, ..., $A^p g$ sont (p+1) combinaisons de p vecteurs indépendants et doivent donc être linéairement dépendantes.

Lemme 2 - Soit A une matrice symétrique définie positive possédant p valeurs propres différentes. Soit encore $x^{(0)}$ un vecteur quelconque, et posons $g^{(0)} = A\,x^{(0)} - b$. Alors, la solution x du système $A\,x = b$ est de la forme

$$x = x^{(0)} + \sum_{j=0}^{p-1} \alpha_j A^j g^{(0)} \qquad (1)$$

Démonstration : Puisque

$$x - x^{(0)} = A^{-1} A\,(x - x^{(0)}) = A^{-1}(b - A\,x^{(0)}) = - A^{-1}\,g^{(0)} ,$$

le problème revient à calculer ce dernier vecteur. De la décomposition en vecteurs propres

$$g^{(0)} = \sum_{i=1}^{p} z_{(i)} \quad ,$$

on déduit

$$A^{-1}\, g^{(0)} = \sum_{i=1}^{p} \frac{1}{\lambda_i}\, z_{(i)} \quad ,$$

tandis que la somme du second membre de (1) vaut

$$\sum_{j=0}^{p-1} \alpha_j\, A^j\, g^{(0)} = \sum_{j=0}^{p-1} \alpha_j \left(\sum_{i=1}^{p} \lambda_i^j\, z_{(i)} \right) = \sum_{i=1}^{p} \left(\sum_{j=0}^{p-1} \alpha_j\, \lambda_i^j \right) z_{(i)} .$$

La comparaison de ces deux expressions mène aux conditions

$$\sum_{j=0}^{p-1} \alpha_j\, \lambda_i^j = -\frac{1}{\lambda_i} \quad ,$$

soit, explicitement,

$$\begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{p-1} \\ \vdots & & & & \vdots \\ 1 & \lambda_p & \lambda_p^2 & \cdots & \lambda_p^{p-1} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_{p-1} \end{bmatrix} = - \begin{bmatrix} 1/\lambda_1 \\ \vdots \\ 1/\lambda_p \end{bmatrix} \qquad (2)$$

Il s'agit d'un système de VANDERMONDE régulier, puisque $\lambda_1 \neq \ldots \neq \lambda_p$, donc la solution existe, ce qui démontre le lemme.

Remarque - Il peut se faire que le vecteur $g^{(0)}$ ait ses composantes dans certains sous-espaces propres nulles. Alors, le développement de $g^{(0)}$ ne contient que $q < p$ vecteurs propres, et la somme (1) se limite à $j = q-1$. (Même démonstration).

A priori, ce résultat semble tout théorique: il est évidemment hors de question de calculer directement les coefficients α_j de la formule (1) par le système (2), ne serait-ce que parce que les valeurs propres ne sont pas connues en général. Du reste, il s'agit d'un système matriciel d'une dimension comparable à celle du système d'origine.

Mais on peut procéder par approximations successives, en calculant

$$x^{(k)} = x^{(0)} + \sum_{j=0}^{k-1} \alpha_j^{(k)}\, A^j\, g^{(0)} \quad ,$$

les $\alpha_j^{(k)}$ étant choisis de manière à minimiser $\phi(x^{(k)})$. On écrira donc

$$\frac{\partial \phi(x^{(k)})}{\partial \alpha_j^{(k)}} = \mathrm{grad}^T\, \phi(x^{(k)})\, \frac{\partial x^{(k)}}{\partial \alpha_j^{(k)}} = g^{(k)T}\, A^j\, g^{(0)} = 0 \quad , \quad j=0, \ldots, k-1 \quad ,$$

c'est-à-dire que $g^{(k)}$ doit être orthogonal au sous-espace enveloppe linéaire

de $g^{(0)}$, $Ag^{(0)}$, ..., $A^{k-1}g^{(0)}$. Ici encore, le nombre de conditions devient rapidement important. Mais on peut tourner la difficulté. En effet, comme

$$r^{(k)} = x^{(k)} - x = A^{-1}(Ax^{(k)} - Ax) = A^{-1}(Ax^{(k)} - b) = A^{-1}g^{(k)},$$

les conditions précédentes équivalent à

$$r^{(k)T}\, Ag^{(0)} = 0\ ,\quad r^{(k)T}\, A^2g^{(0)} = 0,\ \ldots,\ r^{(k)T}\, A^kg^{(0)} = 0\ .$$

De la même façon, $r^{(k+1)}$ est orthogonal à $Ag^{(0)}$, ..., $A^{k+1}g^{(0)}$. Dès lors,

$$x^{(k+1)} - x^{(k)} = r^{(k+1)} - r^{(k)}$$

est orthogonal à $Ag^{(0)}$, ..., $A^kg^{(0)}$; du reste, c'est une combinaison linéaire de $g^{(0)}$, ..., $A^kg^{(0)}$:

$$x^{(k+1)} - x^{(k)} = \sum_{j=0}^{k-1} \left[\alpha_j^{(k+1)} - \alpha_j^{(k)}\right] A^jg^{(0)} + \alpha_k^{(k+1)} A^kg^{(0)}.$$

C'est donc une combinaison linéaire de (k+1) vecteurs soumise à k conditions d'orthogonalité. Ces conditions définissent entièrement sa direction, et on peut écrire

$$x^{(k+1)} - x^{(k)} = \beta_k\, h^{(k)}\ ,$$

$h^{(k)}$ étant une combinaison linéaire quelconque de $g^{(0)}$, ..., $A^kg^{(0)}$ orthogonale à $Ag^{(0)}$, ..., $A^kg^{(0)}$, et β_k étant un scalaire. Or, il est aisé de trouver un tel vecteur $h^{(k)}$, sous la forme

$$h^{(k)} = g^{(k)} + \sigma_k\, h^{(k-1)}\ ,$$

$h^{(k-1)}$ étant un vecteur de même direction que $x^{(k)} - x^{(k-1)}$. En effet,

$$g^{(k)} = A\,x^{(k)} - b = A\,x^{(0)} - b + \sum_{j=0}^{k-1} \alpha_j^{(k)}\, A^{j+1}g^{(0)}$$

$$= g^{(0)} + \sum_{j=0}^{k-1} \alpha_j^{(k)}\, A^{j+1}g^{(0)}$$

est une combinaison linéaire de $g^{(0)}$, ..., $A^kg^{(0)}$ orthogonale à $g^{(0)}$, ..., $A^{k-1}g^{(0)}$ et $h^{(k-1)}$ est une combinaison linéaire de $g^{(0)}$, ..., $A^{k-1}g^{(0)}$ orthogonale à $Ag^{(0)}$, ..., $A^{k-1}g^{(0)}$, donc $g^{(k)} + \sigma_k\, h^{(k-1)}$ est encore une combinaison linéaire de $g^{(0)}$, ..., $A^{k-1}g^{(0)}$, orthogonale à $Ag^{(0)}$, ..., $A^{k-1}g^{(0)}$. Il suffit donc de choisir σ_k pour assurer la condition

$$h^{(k)T}\, A^k\, g^{(0)} = 0\ .$$

Enfin, on peut remplacer dans cette condition $A^k g^{(0)}$ par toute combinaison linéaire de $Ag^{(0)}, \ldots, A^k g^{(0)}$ dont le coefficient de $A^k g^{(0)}$ ne soit pas nul. Une telle combinaison est donnée par le vecteur $Ah^{(k-1)}$, ce qui mène à la condition

$$h^{(k)T} A h^{(k-1)} = 0 ,$$

qui n'est autre que la condition de conjugaison. Connaissant alors la direction de progression, il suffit de minimiser $\phi(x^{(k)} + \beta_k h^{(k)})$ par rapport à β_k. On retrouve ainsi l'algorithme du gradient conjugué. Par conséquent, l'algorithme du gradient conjugué équivaut à construire les meilleures approximations successives de la solution x de la forme

$$x^{(k)} = x^{(0)} + \sum_{j=0}^{k-1} \alpha_j^{(k)} A^j g^{(0)} .$$

Si la matrice A possède p valeurs propres distinctes, on obtient la solution au plus tard après p itérations $(x^{(p)} = x)$.

* 11.3 - Convergence spectrale de la méthode du gradient conjugué

De toutes les combinaisons linéaires de la forme

$$\tilde{x}^{(k)} = \tilde{x}^{(0)} + \sum_{j=0}^{k-1} \tilde{\alpha}_j^{(k)} A^j g^{(0)} ,$$

l'algorithme du gradient conjugué choisit celle qui minimise la fonction $\phi(\tilde{x}^{(k)})$ ou, ce qui est équivalent, la fonction

$$\phi_o(\tilde{x}^{(k)}) = (\tilde{x}^{(k)} - x)^T A (\tilde{x}^{(k)} - x) = \tilde{x}^{(k)} A \tilde{x}^{(k)} - 2 \tilde{x}^{(k)T} b + x^T A x .$$

On obtiendra donc une valeur plus grande pour un autre choix des $\tilde{\alpha}_j^{(k)}$. Nous allons effectuer un choix particulier de la manière suivante: supposant les valeurs propres $\lambda_1, \ldots, \lambda_p$ rangées dans l'ordre décroissant,

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p ,$$

gardons les k plus petites et remplaçons $\lambda_{p-k}, \ldots, \lambda_1$ par l'infini dans l'expression de $A^{-1} g^{(0)}$ obtenue au lemme 2 ci-dessus. Cela donne

$$\tilde{x}^{(k)} = x^{(0)} - \sum_{i=p-k+1}^{p} \frac{z^{(i)}}{\lambda_i}$$

et, par conséquent,

$$\tilde{r}^{(k)} = \tilde{x}^{(k)} - x = -\sum_{i=1}^{p-k} \frac{1}{\lambda_i} z^{(i)} .$$

On en déduit

$$\phi_o(\tilde{x}^{(k)}) = \sum_{i=1}^{p-k} \frac{1}{\lambda_i} z^{(i)T} A \sum_{j=1}^{p-k} \frac{1}{\lambda_j} z^{(j)} = \sum_{i=1}^{p-k} \frac{\|z_{(i)}\|_2^2}{\lambda_i}$$

et

$$\phi_o(x^{(k)}) \leq \phi_o(\tilde{x}^{(k)}) \leq \sum_{i=1}^{p-k} \frac{\|z_{(i)}\|_2^2}{\lambda_i} \leq \frac{1}{\lambda_{p-k}} \sum_{i=1}^{p-k} \|z_{(i)}\|_2^2 ,$$

d'où

$$\|r^{(k)}\|_2^2 \leq \frac{1}{\lambda_p} \phi_o(x^{(k)}) \leq \frac{1}{\lambda_p} \cdot \frac{1}{\lambda_{p-k}} \sum_{i=1}^{p-k} \|z_{(i)}\|_2^2$$

et

$$\|r^{(k)}\|_2 \leq \frac{1}{\sqrt{\lambda_p}} \cdot \frac{1}{\sqrt{\lambda_{p-k}}} \sqrt{\sum_{i=1}^{p-k} \|z_{(i)}\|_2^2}$$

De cette expression, il ressort qu'à chaque itération, le résidu décroît pour deux raisons:

a) Il est multiplié par $\sqrt{\frac{\lambda_{p-k+1}}{\lambda_{p-k}}} < 1$.

b) De moins en moins de termes du développement spectral de $g^{(0)}$ interviennent.

Nous venons donc de montrer que l'algorithme du gradient conjugué converge mieux en termes de ϕ_o que le développement spectral de la solution. Dans de nombreux cas concrets, les praticiens savent que le développement spectral converge assez vite et que, par exemple, il suffit de dix termes pour obtenir une bonne approximation de la solution pour un système de dimension 100. Alors, dix itérations du gradient conjugué mèneront à une approximation comparable. De plus, lorsque les valeurs propres sont multiples, il n'est pas nécessaire d'augmenter le nombre d'itérations, alors que dans les méthodes spectrales (*), il convient de représenter chaque vecteur propre indépendant.

Enfin, il convient de noter le fait remarquable suivant: plus les valeurs propres diffèrent en ordre de grandeur, c'est-à-dire <u>plus la matrice A est mal conditionnée, plus la convergence est rapide</u>.

(*) La <u>méthode spectrale</u> consiste à développer la solution en modes propres:

$$x = A^{-1} b \quad , \qquad b = \sum_i \beta_i \, z_{(i)} \quad ,$$

$$x = \sum_i (\beta_i / \lambda_i) \, z_{(i)} \, .$$

Lorsque les valeurs propres sont suffisamment espacées, on peut ne retenir que les termes correspondants aux plus petites valeurs propres, le reste étant négligeable. Il faut évidemment connaître les vecteurs propres, du moins les premiers. En pratique, cette méthode ne se justifie que pour l'intégration d'équations différentielles du type $\dot{x} = - Ax$.

Exercice 1 - Montrer directement, dans le cas d'une matrice diagonalisable, la relation

$$\lim_{k \to \infty} \|A^k\|^{1/k} = \rho(A) .$$

Solution : Soit S une matrice telle que $S^{-1}AS = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. On peut, sans perte de généralité, supposer que λ_1 est de module maximal:

$$|\lambda_1| = \rho(A).$$

Alors,

$$S^{-1} A^k S = (S^{-1} A S)^k = \mathrm{diag}(\lambda_1^k, \ldots, \lambda_n^k)$$

et

$$\|S^{-1} A S\|_\infty = |\lambda_1|^k ,$$

ce qui entraîne

$$\|A^k\|_\infty = \|S (S^{-1} A^k S) S^{-1}\|_\infty \leq \|S\|_\infty \|S^{-1}\|_\infty |\lambda_1|^k ,$$

et

$$\|A^k\|_\infty^{1/k} = (\|S\|_\infty \|S^{-1}\|_\infty)^{1/k} |\lambda_1| \longrightarrow |\lambda_1| .$$

On a par ailleurs

$$\|S^{-1} A^k S\|_\infty \leq \|S^{-1}\|_\infty \|A^k\|_\infty \|S\|_\infty ,$$

d'où

$$\|A^k\|_\infty^{1/k} \leq \frac{1}{(\|S^{-1}\|_\infty \|S\|_\infty)^{1/k}} \longrightarrow |\lambda_1| .$$

Exercice 2 - Soit l'équation matricielle

$$x = B x + c .$$

On suppose que B est diagonalisable et que toutes ses valeurs propres ont leur module inférieur à l'unité. Montrer que l'équation matricielle ci-dessus admet une solution unique.

Solution: Il suffit de montrer que la matrice (I - B) n'est pas singulière. Si la matrice S diagonalise B, on a

$$\mathrm{dtm}(I - B) = \mathrm{dtm}(S^{-1}(I-B)S) = (1 - \lambda_1)\ldots(1 - \lambda_n) \neq 0$$

* Exercice 3 - Dans l'énoncé de l'exeercice précédent, se débarrasser de l'hypothèse que B est diagonalisable.

Solution : Soient deux vecteurs x et y, et posons

$$\hat{x} = B x + c \quad , \quad \hat{y} = B y + c .$$

Il est clair que

$$\hat{x} - \hat{y} = B(x - y) .$$

Soit alors q un nombre tel que $\rho(B) < q < 1$. Il existe une matrice inversible S telle que

$$\|S^{-1} B S\|_\infty \leq q .$$

On a évidemment

$$S^{-1}(\hat{x} - \hat{y}) = S^{-1} B S S^{-1}(x - y) \quad ,$$

ce qui entraîne

$$\|S^{-1}(\hat{x} - \hat{y})\|_\infty \leq q \, \|S^{-1}(x - y)\|_\infty \quad .$$

L'application $S^{-1}BS$ est donc contractante, et toute suite de la forme

$$z^{(k+1)} = S^{-1} B S z^{(k)} + S^{-1} c$$

converge vers une limite d telle que

$$d = S^{-1}BS\, d + S^{-1}c \quad .$$

Dès lors, à la suite $\{x^{(k)}\}$ définie par la relation de récurrence

$$x^{(k+1)} = B x^{(k)} + c,$$

on peut associer la suite $\{z^{(k)}\}$ qui est de la forme ci-dessus et

$$\|x^{(k)} - S d\|_\infty \leq \|S\|_\infty \, \|z^{(k)} - d\|_\infty \longrightarrow 0,$$

donc $\{x^{(k)}\}$ converge vers la solution Sd .

Exercice 4 - Soit l'équation matricielle

$$x = B x + c \quad .$$

Montrer qu'elle admet une solution unique si et seulement si B n'a aucune valeur propre égale à 1.

Solution : Cette équation équivaut à

$$(I - B) x = c \quad .$$

Elle admet une solution unique si et seulement si

$$\det(I-B) \neq 0 \quad .$$

Or, ce déterminant est nul si et seulement si il existe un vecteur z tel que

$$(I - B)z = 0 \quad .$$

Mais alors, z est vecteur propre de B, de valeur propre égale à 1.

Exercice 5 - Montrer directement, à partir de l'équation du rayon spectral, que la méthode de Gauss-Seidel converge pour une matrice A symétrique définie positive [7].

Solution : En notant L l'extradiagonale inférieure et $U = L^T$ l'extradiagonale supérieure, les valeurs propres de la matrice d'itération vérifient la relation

$$\det(L^T + \lambda (D + L)) = 0.$$

Un vecteur propre (complexe!) vérifie donc

$$L^T z = - \lambda (D + L) z$$

et, en notant $z^* = \bar{z}^T$, on a

$$\lambda = \frac{z^* L^T z}{z^*(D + L)z} \quad .$$

Or, D est définie positive, ce qui implique

$$z^* D z = \sigma > 0 \quad .$$

Posant d'autre part

$$z^* L^T z = \alpha + i\beta \quad ,$$

on a

$$z^* L z = (z^* L^T z)^* = \alpha - i\beta$$

et

$$\lambda = \frac{\alpha - i\beta}{\sigma + \alpha - i\beta} \quad , \quad |\lambda| = \left(\frac{\alpha^2 + \beta^2}{(\sigma + \alpha)^2 + \beta^2} \right)^{\frac{1}{2}} .$$

De plus, A étant définie positive,

$$z^* A z = z^* (D + L + L^T) z = \sigma + 2\alpha > 0 ,$$

donc

$$\sigma + \alpha > -\alpha ,$$

ce qui entraîne

$$|\lambda| < 1.$$

Exercice 6 - Soit A une matrice symétrique définie positive de dimension n, ayant une seule valeur propre. On considère le système

$$A x = b.$$

a) Combien faut-il d'itérations pour le résoudre par la méthode de la plus grande pente?

b) Même question, par la méthode du gradient donjugué

c) Comparer le nombre d'opérations significatives nécessaires par ces deux méthodes à celui qu'exige la méthode de Gauss.

d) Interpréter géométriquement les réponses données aux points a) et b).

Solution :

a) 1, car le résidu est multiplié, après une itération, par un nombre dont le numérateur contient le facteur $(\lambda_1 - \lambda_n) = 0$.

b) 1, car p = 1 .

c) Le gradient conjugué et la plus grande pente coïncident à la première itération. Le nombre d'opérations est $O(n^2)$, alors que la méthode de Gauss exige $O(n^3)$ O.S. .

d) On a dans le cas présent

$$\phi(x) = \tfrac{1}{2} \sum_k (x_k - b_k)^2 + \text{cte.}$$

C'est l'équation d'une sphère, et le gradient pointe directement sur son centre.

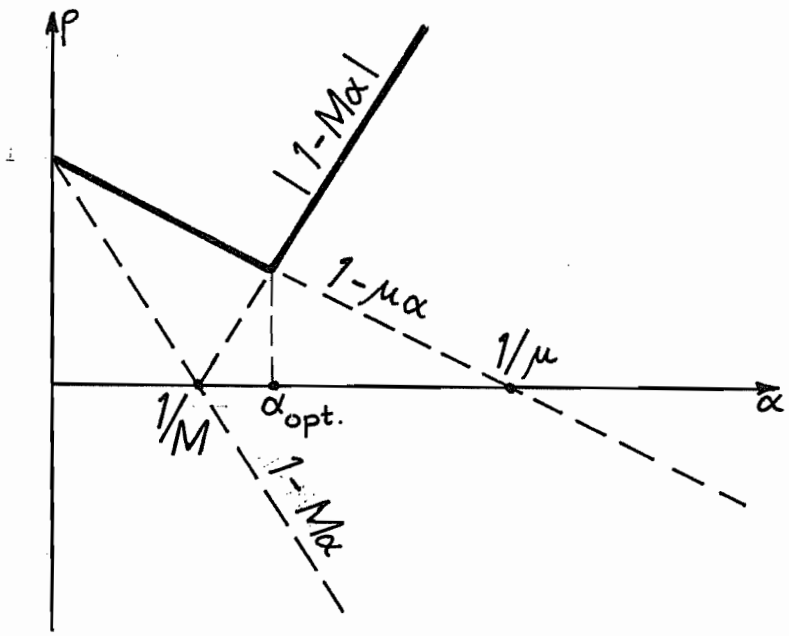


Fig.1

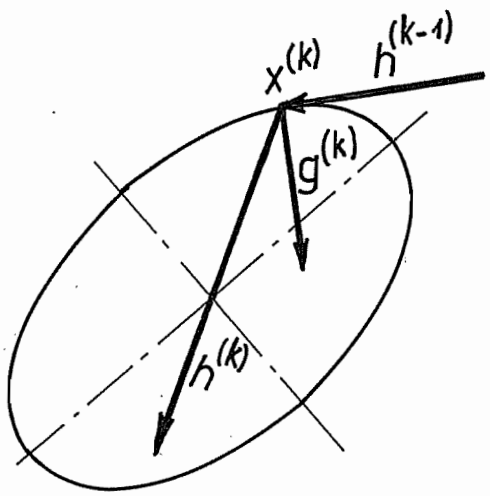


Fig. 2

RECHERCHE DES VALEURS PROPRES DES MATRICES.

Les problèmes de valeurs propres se rencontrent assez fréquemment dans les applications physiques. Citons en particulier les problèmes de vibrations, et de nombreux phénomènes d'instabilité.

Certaines applications nécessitent la connaissance de tout le spectre; certaines autres, par contre, se satisfont largement des quelques plus grandes (ou plus petites) valeurs propres, et des vecteurs propres associés. Ces deux types de problèmes se traitent différemment.

1. DISQUES DE GERSHGORIN

1.1 - Un procédé très simple permettant de situer les valeurs propres d'une matrice quelconque est fondé sur le lemme suivant:

Une matrice A à diagonale strictement dominante, c'est-à-dire vérifiant

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad , \quad i = 1, \ldots, n$$

ne peut être singulière.

En effet, supposons A singulière: il existe donc un vecteur x tel que A x = 0. Soit alors x_k la plus grande en module des composantes de x. On a

$$a_{kk}\, x_k + \sum_{i \neq k} a_{ki}\, x_i = 0 \ ,$$

soit

$$a_{kk}\, x_k = - \sum_{i \neq k} a_{ki}\, x_i \ ,$$

ce qui entraîne

$$|a_{kk}| \leq \sum_{i \neq k} |a_{ki}| \frac{|x_i|}{|x_k|} \leq \sum_{i \neq k} |a_{ki}| \ .$$

1.2 - Soit à présent λ une valeur propre de A. La matrice $(A - \lambda I)$ est donc singulière, si bien qu'en vertu du lemme précédent, il existe une valeur de i pour laquelle

$$\boxed{|a_{ii} - \lambda| \leq \sum_{j \neq i} |a_{ij}|}$$

En d'autres termes, toute valeur propre de la matrice A repose au moins dans l'un des disques de centre a_{ii} et de rayon $\sum_{j \neq i} |a_{ij}|$. Ces disques portent le nom de disques de GERSHGORIN.

2. THEOREME DE STABILITE DES VALEURS PROPRES D'UNE MATRICE SYMETRIQUE

2.1 - Soit A une matrice symétrique, et soit E une perturbation de A

(matrice quelconque n x n). Les valeurs propres μ de (A + E) vérifient

$$\inf_{i=1,\ldots,n} |\mu - \lambda_i| \leq \|E\|_2 .$$

On remarque d'abord que si $\mu = \lambda_1, \lambda_2, \ldots,$ ou λ_n, la proposition est évidente. Soit donc

$$\inf_{i=1,\ldots,n} |\mu - \lambda_i| \neq 0 .$$

La matrice $(A - \mu I + E)$ est singulière, donc il existe un vecteur x tel que

$$(A - \mu I)\, x = - E\, x .$$

Comme μ n'est égal à aucune valeur propre de A, $A - \mu I$ est inversible et

$$x = - (A - \mu I)^{-1} E\, x,$$

ce qui implique

$$\|x\|_2 \leq \|(A - \mu I)^{-1}\|_2 \, \|E\|_2 \, \|x\|_2$$

et

$$\frac{1}{\|(A - \mu I)^{-1}\|_2} = \inf_{i=1,\ldots,n} |\mu - \lambda_i| \leq \|E\|_2 .$$

Ce théorème garantit la <u>stabilité du problème aux valeurs propres par rapport aux perturbations de la matrice</u>. Son application est cependant assez lourde, car le calcul de $\|E\|_2$ nécessite lui-même la recherche de la plus grande valeur propre de la matrice $E^T E$. Mais on peut utiliser une norme plus simple, car

$$\|E\|_2^2 = \lambda_{max}(E^T E) \leq \sum_{i=1}^{n} \lambda_i(E^T E) = tr(E^T E) = \sum_{i,j} e_{ij}^2 .$$

On définit donc la <u>norme euclidienne</u> de la matrice E par

$$\|E\|_E^2 = \sum_{i,j} e_{ij}^2 .$$

Ce n'est pas une norme de E en tant qu'opérateur, mais en tant que tableau. Elle rend néanmoins de grands services. Le théorème précédent permet alors d'affirmer que <u>toute valeur propre μ de la matrice perturbée (A + E) se trouve dans un des disques de centre λ_i et de rayon $\|E\|_E$, que nous appellerons disques euclidiens</u> de (A + E).

En d'autres termes, la dépendance des valeurs propres de la matrice par rapport à ses éléments est <u>continue</u>.

3. THEOREME DE REPARTITION DES VALEURS PROPRES DE LA MATRICE PERTURBEE (A + E) DANS LES DISQUES EUCLIDIENS. LORSQUE LA PERTURBATION E EST SYMETRIQUE.

Soit A une matrice symétrique, et soit E une perturbation symétrique de A. Si la réunion de m disques euclidiens $B_{\|E\|_E}(\lambda_i)$ est disjointe des (n-m) disques restants, la réunion de ces m disques contient exactement m valeurs propres de (A+E) (fig. 1)

La démonstration se fonde sur la continuité de la dépendance des valeurs propres de la matrice par rapport à ses éléments. Considérons la matrice

$$A_\varepsilon = A + \varepsilon E ,$$

ε étant un paramètre variant entre 0 et 1. Il est clair que le rayon des disques euclidiens de la perturbation εE de A est proportionnel à ε. Soit alors G_ε la réunion de m disques disjoints des (n-m) autres. Pour $\varepsilon = 0$, G_0 se réduit à m points, qui sont des valeurs propres. Lorsque ε croît, les valeurs propres μ_ε de A_ε suivent des courbes continues dans $\mathbb{C}$. Elles ne peuvent donc passer d'un disque à l'autre que si leur intersection est non vide, ce qui implique que G_1 contient exactement m valeurs propres.

En particulier, un disque euclidien isolé ne peut contenir qu'une seule valeur propre.

4. METHODE DE JACOBI

La méthode de JACOBI permet d'obtenir simultanément toutes les valeurs propres d'une matrice symétrique A. L'idée est la suivante: il existe une matrice orthonormale S telle que

$$S^T A S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

En appliquant cette matrice, on annule tous les termes non diagonaux. Bien entendu, on ne connaît pas la matrice S, mais on peut aisément trouver une matrice qui annule le terme diagonal de module maximal (ou l'un d'eux s'il y en a plusieurs de même module). Après cela, on annulera dans une nouvelle opération le nouveau terme maximal et ainsi de suite, dans l'espoir d'annuler finalement tous les termes non diagonaux. Pour décrire ce procédé, appelons $A^{(0)}$ la matrice A. On passe de $A^{(k)}$ à $A^{(k+1)}$ en calculant

$$A^{(k+1)} = S^{(k)T} A^{(k)} S^{(k)} ,$$

$S^{(k)}$ étant constfuit comme suit: soit $a_{pq}^{(k)}$ un des termes maximaux en module de $A^{(k)}$. On pose

$$S^{(k)} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \cos\theta_k & & \sin\theta_k & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & -\sin\theta_k & & \cos\theta_k & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \leftarrow p \\ \\ \\ \leftarrow q \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & \uparrow & & & \uparrow & \\ & & p & & & q & \end{matrix}$$

C'est ce que l'on appelle une <u>rotation élémentaire</u>. L'angle θ_k est choisi pour annuler $a_{pq}^{(k+1)}$. Examinons comment se fait le produit. Seules changent les lignes et les colonnes p et q. On a, pour $i,j \neq p,q$,

$$a_{ij}^{(k+1)} = a_{ij}^{(k)}$$

$$a_{ip}^{(k+1)} = \sum_{rs} S_{ri}^{(k)} a_{rs}^{(k)} S_{sp}^{(k)} = \sum_{s} a_{is}^{(k)} S_{sp}^{(k)} = a_{iq}^{(k)} S_{qp}^{(k)} + a_{ip}^{(k)} S_{pp}^{(k)}$$

$$= - a_{iq}^{(k)} \sin\theta_k + a_{ip}^{(k)} \cos\theta_k \ .$$

$$a_{iq}^{(k+1)} = \sum_{rs} S_{ri}^{(k)} a_{rs}^{(k)} S_{sq}^{(k)} = \sum_{s} a_{is}^{(k)} S_{sq}^{(k)} = a_{ip}^{(k)} S_{pq}^{(k)} + a_{iq}^{(k)} S_{qq}^{(k)}$$

$$= a_{ip}^{(k)} \sin\theta_k + a_{iq}^{(k)} \cos\theta_k \ .$$

$$a_{pp}^{(k+1)} = S_{pp}^{(k)} a_{pp}^{(k)} S_{pp}^{(k)} + S_{qp}^{(k)} a_{qp}^{(k)} S_{pp}^{(k)} + S_{pp}^{(k)} a_{pq}^{(k)} S_{qp}^{(k)} + S_{qp}^{(k)} a_{qq}^{(k)} S_{qp}^{(k)}$$

$$= a_{pp}^{(k)} \cos^2\theta_k - a_{qp}^{(k)} \sin\theta_k \cos\theta_k - a_{pq}^{(k)} \sin\theta_k \cos\theta_k + a_{qq}^{(k)} \sin^2\theta_k$$

$$= a_{pp}^{(k)} \cos^2\theta_k + a_{qq}^{(k)} \sin^2\theta_k - 2 a_{pq}^{(k)} \sin\theta_k \cos\theta_k \ .$$

$$a_{qq}^{(k+1)} = S_{pq}^{(k)} a_{pp}^{(k)} S_{pq}^{(k)} + S_{qq}^{(k)} a_{qp}^{(k)} S_{pq}^{(k)} + S_{pq}^{(k)} a_{pq}^{(k)} S_{qq}^{(k)} + S_{qq}^{(k)} a_{qq}^{(k)} S_{qq}^{(k)}$$

$$= a_{pp}^{(k)} \sin^2\theta_k + a_{pq}^{(k)} \sin\theta_k \cos\theta_k + a_{pq}^{(k)} \sin\theta_k \cos\theta_k + a_{qq}^{(k)} \cos^2\theta_k$$

$$= a_{pp}^{(k)} \sin^2\theta_k + a_{qq}^{(k)} \cos^2\theta_k + 2 a_{pq}^{(k)} \sin\theta_k \cos\theta_k \ .$$

$$a_{pq}^{(k+1)} = S_{pp}^{(k)} a_{pp}^{(k)} S_{pq}^{(k)} + S_{pp}^{(k)} a_{pq}^{(k)} S_{qq}^{(k)} + S_{qp}^{(k)} a_{qp}^{(k)} S_{pq}^{(k)} + S_{qp}^{(k)} a_{qq}^{(k)} S_{qq}^{(k)}$$

$$= a_{pp}^{(k)} \sin\theta_k \cos\theta_k + a_{pq}^{(k)} \cos^2\theta_k - a_{qp}^{(k)} \sin^2\theta_k$$

$$- a_{qq}^{(k)} \sin\theta_k \cos\theta_k$$

$$= (a_{pp}^{(k)} - a_{qq}^{(k)}) \sin\theta_k \cos\theta_k + a_{pq}^{(k)} (\cos^2\theta_k - \sin^2\theta_k) .$$

L'annulation de $a_{pq}^{(k+1)}$ sera obtenue si

$$\tfrac{1}{2}(a_{pp}^{(k)} - a_{qq}^{(k)}) \sin 2\theta_k + a_{pq}^{(k)} \cos 2\theta_k = 0 ,$$

ce qui donne la condition

$$\boxed{\operatorname{tg} 2\theta_k = 2 \frac{a_{pq}^{(k)}}{a_{qq}^{(k)} - a_{pp}^{(k)}}}$$

Comme il existe plusieurs angles θ_k possibles, on choisit la détermination $|\theta_k| < \pi/4$. Cependant, cette formule ne vaut pas si $a_{pp}^{(k)} = a_{qq}^{(k)}$. Dans ce cas, l'équation d'annulation de $a_{pq}^{(k+1)}$ se réduit à

$$a_{pq}^{(k)} \cos 2\theta_k = 0,$$

soit

$$\theta_k = \pi/4 .$$

On remarquera d'ailleurs qu'il n'est pas nécessaire de calculer l'angle. Posant en effet

$$X = 2 a_{pq}^{(k)} \quad , \quad Y = a_{qq}^{(k)} - a_{pp}^{(k)} \quad , \quad Z = (X^2 + Y^2)^{\frac{1}{2}} \quad ,$$

on a

$$\operatorname{tg} 2\theta_k = X/Y$$

et

$$2\cos^2\theta_k = 1 + \cos 2\theta_k = 1 + \frac{1}{(1 + \operatorname{tg}^2 2\theta_k)^{\frac{1}{2}}} = 1 + \frac{1}{(1 + X^2/Y^2)^{\frac{1}{2}}}$$

$$= 1 + \frac{|Y|}{(X^2 + Y^2)^{\frac{1}{2}}} = 1 + |Y|/Z ,$$

soit

$$\boxed{\cos\theta_k = \left[\tfrac{1}{2}\left(1 + \frac{|Y|}{Z}\right)\right]^{\frac{1}{2}}}$$

On a de même

$$\sin^2\theta_k = 1 - \cos^2\theta_k = \tfrac{1}{2}(1 - |Y|/Z) ,$$

mais cette formule est très imprécise si $|X| \ll |Y|$ ($\approx 1 - 1$) . On la trasforme en notant que

$$\sin 2\theta_k = 2 \sin\theta_k \cos\theta_k = \operatorname{tg} 2\theta_k \cdot \cos 2\theta_k = \frac{X}{Y}\,(1 + X^2/Z^2)^{-\frac{1}{2}}$$

$$= \frac{X \operatorname{sign} Y}{Z} ,$$

d'où

$$\boxed{\sin\theta_k = \frac{X \operatorname{sign} Y}{2\, Z \cos\theta_k}}$$

5. CONVERGENCE DE LA METHODE DE JACOBI

5.1 - La convergence de la méthode de JACOBI n'est nullement évidente, car en annulant un terme non diagonal, on modifie les autres, y compris ceux que l'on a déjà annulés. Notons cependant que, tout au cours du processus, les valeurs propres des matrices $A^{(k)}$ sont constamment égales à celles de A, puisque, pour toute matrice de rotation, on a

$$\operatorname{dtm}(S^T A S - \lambda I) = \operatorname{dtm}(S^T (A - \lambda I) S) = |\operatorname{dtm} S|^2 \operatorname{dtm}(A - \lambda I)$$

$$= \operatorname{dtm}(A - \lambda I) ,$$

c'est-à-dire que l'équation caractéristique reste invariante. On a aussi la propriété suivante:

$$\|A^{(k)}\|_E^2 = \sum_{ij} a_{ij}^{(k)2} = \sum_{ij} a_{ij}^2 = \|A\|_E^2 .$$

En effet, si T est la matrice de transformation jusqu'à l'étape K,

$$\sum_{ij} a_{ij}^{(k)2} = \operatorname{tr}(A^{(k)T} A^{(k)}) = \operatorname{tr}\,(T^T A\, T\, T^T A\, T) = \operatorname{tr}\,(T^T A^T A T)$$

$$= \sum_{ijkl} T_{ij}^T A_{jk} A_{kl} T_{li} = \sum_{ijkl} \delta_{ji} A_{jk} A_{kl}$$

$$= \sum_{ik} A_{ik} A_{ki} = \operatorname{tr}\,(A^T A) .$$

Or, si l'on examine les relations de transformation, on constate que

a) $\sum_{\substack{i\neq p\\ \neq q}} a_{ii}^{(k+1)2} = \sum_{\substack{i\neq p\\ \neq q}} a_{ii}^{(k)\,2}$ (termes inchangés)

b) L'ensemble des termes

$$\begin{bmatrix} a_{pp}^{(k)} & a_{pq}^{(k)} \\ a_{pq}^{(k)} & a_{qq}^{(k)} \end{bmatrix}$$

se transforme exactement comme une matrice 2x2, ce qui entraîne

$$a_{pp}^{(k+1)\,2} + a_{qq}^{(k+1)\,2} + 2.0 = a_{pp}^{(k)\,2} + a_{qq}^{(k)\,2} + 2\,a_{pq}^{(k)\,2} \ .$$

En combinant ces deux résultats, on obtient

$$\sum_i a_{ii}^{(k+1)\,2} = \sum_i a_{ii}^{(k)\,2} + 2\,a_{pq}^{(k)\,2} \ .$$

Par conséquent, on a nécessairement

$$\sum_{i\neq j} a_{ij}^{(k+1)\,2} = \sum_{ij} a_{ij}^{(k+1)\,2} - \sum_i a_{ii}^{(k+1)\,2} =$$

$$= \sum_{ij} a_{ij}^{(k)\,2} - \sum_i a_{ii}^{(k)\,2} - 2\,a_{pq}^{(k)\,2}$$

$$= \sum_{i\neq j} a_{ij}^{(k)\,2} - 2\,a_{pq}^{(k)\,2} \ .$$

Comme $a_{pq}^{(k)}$ est le plus grand en module des termes extradiagonaux, on a nécessairement (le plus grand étant supérieur à la moyenne)

$$\sum_{j\neq i} a_{ij}^{(k)\,2} \leqslant n(n-1)\,a_{pq}^{(k)\,2} \ ,$$

ce qui entraîne

$$a_{pq}^{(k)\,2} \geqslant \frac{\sum_{i\neq j} a_{ij}^{(k)\,2}}{n(n-1)} \qquad (1)$$

et

$$\sum_{i\neq j} a_{ij}^{(k+1)\,2} \leqslant \left(1 - \frac{2}{n(n-1)}\right) \sum_{i\neq j} a_{ij}^{(k)\,2}$$

Ainsi, si l'on note $A^{(k)} = D^{(k)} + E^{(k)}$, où $D^{(k)}$ est la diagonale de $A^{(k)}$ et $E^{(k)}$ son extradiagonale, on a

$$\|E^{(k+1)}\|_E^2 \leqslant \left(1 - \frac{2}{n(n-1)}\right) \|E^{(k)}\|_E^2$$

et, par conséquent,

$$\|E^{(k)}\|_E^2 \leqslant \left(1 - \frac{2}{n(n-1)}\right)^k \|E^{(0)}\|_E^2 \ ,$$

c'est-à-dire que <u>l'extradiagonale tend vers zéro</u>, au moins en progression géométrique

5.2 - Montrons à présent que les éléments diagonaux convergent vers les valeurs propres. Supposons provisoirement que toutes les valeurs propres sont distinctes, et posons

$$\beta = \inf_{i \neq j} |\lambda_i - \lambda_j| .$$

Alors, pour k suffisamment grand, on peut obtenir

$$\|E^{(k)}\|_E \leqslant \varepsilon < \frac{\beta}{6} \quad , \quad k \geqslant k_o$$

Soit $k \geqslant k_o$. La matrice $A^{(k)} - E^{(k)}$ est une perturbation de $A^{(k)}$ dont les valeurs propres sont les termes diagonaux de cette dernière; les valeurs propres de $A^{(k)}$ étant égales à celles de A, on déduit du théorème de stabilité des valeurs propres l'existence, pour deux valeurs propres λ_i et λ_j, d'éléments $a_{ii}^{(k)}$ et $a_{jj}^{(k)}$ sur la diagonale de $A^{(k)}$ tels que

$$|a_{ii}^{(k)} - \lambda_i| \leqslant \varepsilon \quad , \quad |a_{jj}^{(k)} - \lambda_j| \leqslant \varepsilon .$$

Chacun des disques de centre λ_i et de rayon ε (fig. 2) contient donc <u>un seul</u> élément diagonal, car les relations ci-dessus entraînent

$$|a_{ii}^{(k)} - \lambda_j| \geqslant |\lambda_i - \lambda_j| - |\lambda_i - a_{ii}^{(k)}| > \beta - \varepsilon > \frac{5}{6}\beta > 5\varepsilon .$$

Cependant, ceci ne prouve pas que $a_{ii}^{(k)} \longrightarrow \lambda_i$, car on pourrait très bien imaginer qu'à l'itération (k+1), $a_{ii}^{(k+1)}$ passe dans le disque de centre λ_j, et $a_{jj}^{(k+1)}$ dans celui de centre λ_i. Il y aurait dans ce cas accumulation dans chaque disque, mais non convergence.

Tout revient à montrer que, lors de l'étape $k \mapsto k+1$, $a_{pp}^{(k)}$ ne peut échanger son disque avec $a_{qq}^{(k)}$, c'est-à-dire que $a_{pp}^{(k+1)}$ ne peut se trouver dans le disque de centre λ_q (puisque les autres termes diagonaux ne changent pas).

On a en effet, à partir des formules de transformation,

$$a_{pp}^{(k+1)} - \lambda_q = a_{pp}^{(k)} \cos^2\theta_k + a_{qq}^{(k)} \sin^2\theta_k - 2\, a_{pq}^{(k)} \sin\theta_k \cos\theta_k - \lambda_q$$

$$= (a_{pp}^{(k)} - \lambda_p) \cos^2\theta_k + (a_{qq}^{(k)} - \lambda_q) \sin^2\theta_k$$

$$- 2\, a_{pq}^{(k)} \sin\theta_k \cos\theta_k + (\lambda_p - \lambda_q) \cos^2\theta_k ,$$

d'où

$$|a_{pp}^{(k+1)} - \lambda_q| \geqslant \beta\cos^2\theta - \varepsilon\cos^2\theta_k - \varepsilon\sin^2\theta_k - \varepsilon \sin 2\theta_k$$

$$\geqslant \beta\cos^2\theta - \varepsilon - \varepsilon\sin 2\theta_k > 6\varepsilon \cos^2\theta_k - 2\varepsilon$$

et, puisque $|\theta_k| \leqslant \frac{\pi}{4}$, on a $\cos^2\theta_k \geqslant \frac{1}{2}$, ce qui entraîne

$$|a_{pp}^{(k+1)} - \lambda_q| > \frac{6}{2}\varepsilon - 2\varepsilon = \varepsilon .$$

Donc, $a_{pp}^{(k+1)}$ ne peut se trouver dans le disque de centre λ_q et de rayon ε,

ce qui démontre la convergence dans le cas où toutes les valeurs propres sont différentes.

Examinons à présent le cas où certaines valeurs propres sont confondues. Dans ce cas, on sait que toutes les valeurs propres de la matrice $A^{(k)} - E^{(k)} = \mathrm{diag}(a_{11}^{(k)}, \ldots, a_{nn}^{(k)})$, c'est-à-dire tous les éléments diagonaux de $A^{(k)}$, se trouvent dans un disque de centre λ_i et de rayon $\|E^{(k)}\|_E$. Bien plus, par le théorème de répartition, si tous les disques correspondant à des valeurs propres différentes sont disjoints, chacun contient un nombre d'éléments diagonaux égal à la multiplicité m de la valeur propre, puisqu'il s'agit en fait de m disques confondus. Posons donc

$$\beta = \inf_{\lambda_i \neq \lambda_j} |\lambda_i - \lambda_j|$$

et supposons $\|E^{(k)}\|_E \leq \varepsilon < \beta/6$. Alors, par le même raisonnement que ci-dessus, on trouvera que deux éléments diagonaux situés dans des disques différents ne peuvent changer de disque, ce qui achève la démonstration.

6. CALCUL DES VECTEURS PROPRES PAR LA METHODE DE JACOBI

Lorsque la matrice extradiagonale est suffisamment petite, la matrice

$$T^{(p)} = S^{(0)} \ldots S^{(p)}$$

où p est le nombre d'itérations, vérifie

$$T^{(p)T} A T^{(p)} = A^{(p)} \approx \mathrm{diag}(\lambda_1, \ldots, \lambda_n) ,$$

ce qui laisse penser que ses colonnes constituent une approximation des vecteurs propres. Appelons en effet $c_{(1)}, \ldots, c_{(n)}$ les colonnes de $T^{(p)}$. Alors, si $e_{(i)}$ est le i^e vecteur de la base canonique de R^n,

$$A^{(p)} e_{(i)} = i^e \text{ colonne de } A^{(p)} = i^e \text{ col. de } T^{(p)T} A T^{(p)} = T^{(p)T} A c_{(i)} ,$$

soit

$$T^{(p)T} A c_{(i)} = A^{(p)} e_{(i)} = a_{ii} e_{(i)} + E^{(p)} e_{(i)}$$

Prémultipliant par $T^{(p)}$, on obtient

$$A c_{(i)} = T^{(p)} a_{ii} e_{(i)} + T^{(p)} E^{(p)} e_{(i)} ,$$

ce qui donne le résidu de $c_{(i)}$:

$$r_{(i)} = A c_{(i)} - a_{ii} c_{(i)} = T^{(p)} E^{(p)} e_{(i)}$$

Visiblement, ce résidu est de l'ordre de grandeur de $\|E^{(p)}\|_E$, car

$$\|r_{(i)}\|_2 \leqslant \|T^{(p)}\|_2 \, \|E^{(p)}\|_2 \, \|e_{(i)}\|_2 \leqslant \|E^{(p)}\|_2 \leqslant \|E^{(p)}\|_E \, .$$

Nous verrons en section 12 que cette propriété entraîne que $c_{(i)}$ approche le vecteur propre avec une approximation du même ordre.

Voyons à présent comment se calcule $T^{(p)}$. Il s'agit de multiplier à chaque pas $T^{(k)}$ par $S^{(k+1)}$. On a donc, si $j \neq p$, $j \neq q$,

$$\left.\begin{aligned} t_{ij}^{(k+1)} &= t_{ij}^{(k)} \\ t_{ip}^{(k+1)} &= \sum_s t_{is}^{(k)} \, S_{sp}^{(k+1)} = t_{ip}^{(k)} \cos\theta_{k+1} - t_{iq}^{(k)} \sin\theta_{k+1} \\ t_{iq}^{(k+1)} &= \sum_s t_{is}^{(k)} \, S_{sq}^{(k+1)} = t_{ip}^{(k)} \sin\theta_{k+1} + t_{iq}^{(k)} \cos\theta_{k+1} \end{aligned}\right\} \; i = 1, \ldots, n$$

Il suffit donc d'effectuer ces opérations simples à chaque pas pour obtenir en fin de calcul tous les vecteurs propres.

7. TECHNIQUES PARTICULIERES

La recherche de l'élément $a_{pq}^{(k)}$ de module maximal est très longue, car elle nécessite l'inspection de $\frac{n(n-1)}{2}$ termes, soit $\frac{n(n-1)}{2} - 1$ tests pour trouver le plus grand. Pour n = 50, on obtient

$$\frac{50.49}{2} - 1 = 1224 \text{ tests} \, ,$$

à faire à chaque pas. En admettant que 6 cycles de $\frac{n(n-1)}{2}$ annulations suffisent, il faudra effectuer tous ces tests

$$6 \times \frac{n(n-1)}{2} = 7350 \text{ fois},$$

ce qui fait un total de l'ordre de neuf millions de tests.

7.1 - On allège sensiblement le travail en cherchant non pas l'élément maximal, mais l'élément dit optimal, que l'on définit comme suit:

a) On commence par calculer les normes $\|l_{(1)}^{(k)}\|_2$,..., $\|l_{(n)}^{(k)}\|_2$ de chaque ligne de l'extradiagonale $E^{(k)}$, et on repère la ligne p dont la norme est maximale. Ceci implique

$$n \, \|l_{(p)}^{(k)}\|_2^2 \geqslant \sum_i \|l_{(i)}^{(k)}\|_2^2 = \|E^{(k)}\|_E^2 \, ,$$

donc

$$\|l_{(p)}^{(k)}\|_2^2 \geqslant \frac{\|E^{(k)}\|_E^2}{n} \, .$$

b) On cherche alors dans cette ligne l'élément $a_{pq}^{(k)}$ de module maximal. Comme

$$(n-1) \, a_{pq}^{(k)\,2} \geqslant \|l_{(p)}^{(k)}\|_2^2$$

du fait que la ligne p compte au plus (n-1) termes non nuls, on a encore

$$a_{pq}^{(k)\,2} \geqslant \frac{\|E^{(k)}\|_E^2}{n(n-1)} \quad ,$$

et toute la théorie précédente continue de s'appliquer comme avec l'élément maximal.

c) Après annulation du terme $a_{pq}^{(k)}$, seules les lignes p et q changent en norme, car on vérifie aisément que si $i \neq p$ et $i \neq q$,

$$a_{ip}^{(k+1)\,2} + a_{iq}^{(k+1)\,2} = a_{ip}^{(k)\,2} + a_{iq}^{(k)\,2} \quad .$$

Il suffit donc de recalculer $\|l_{(p)}^{(k+1)}\|_2^2$ et $\|l_{(q)}^{(k+1)}\|_2^2$ pour rechercher le nouvel élément optimal. Au total, une recherche nécessite donc 2(n-1) produits, plus les additions, (n-1) tests pour trouver la ligne maximale, (n-2) tests pour trouver l'élément optimal, soit

$$\begin{cases} 2(n-1) \text{ multiplications} \\ 2n-3 \text{ tests} \end{cases}$$

Pour une matrice 50 x 50, on obtient 98 multiplications et 97 tests.

7.2 - Une autre manière de procéder consiste à établir des barrières. Le principe est le suivant: on définit un nombre positif α, et on parcourt systématiquement tous les $\frac{n(n-1)}{2}$ termes différents de l'extradiagonale supérieure. On annule les termes de module supérieur à la barrière, et on laisse les autres.

Ici, on a évidemment

$$\|E^{(k+1)}\|_E^2 \leqslant \|E^{(k)}\|_E^2 - 2\alpha^2 \quad ,$$

ce qui assure la décroissance de l'extradiagonale. Lorsque tous les éléments extradiagonaux sont inférieurs à la barrière, on la remplace par une autre plus petite. Le choix des barrières est cependant assez délicat. En effet, une barrière trop petite fera perdre du temps par annulation de petits termes qui ne modifient presque pas $\|E^{(k}\|_E^2$; une barrière trop grande retarde la révision des termes de grandeur moyenne. Un mauvais choix de barrière peut enfin détruire la propriété de convergence quadratique dont nous parlerons en section 8. Pour maintenir toutes les propriétés importantes de l'algorithme, il faut garantir qu'un élément révisé vérifie

$$|a_{pq}^{(k)}| \geqslant \frac{\|E^{(k)}\|_E}{\sqrt{n(n-1)}} \quad .$$

Ceci suggère la méthode suivante:

a) On calcule $\|E^{(0)}\|_E^2 = 2 \sum_{j>i} a_{ij}^{(k)\,2}$,

ce qui nécessite $\left(\frac{n(n-1)}{2} + 1 \right)$ multiplications.

b) On prend pour barrière

$$\alpha = \frac{\|E^{(0)}\|_E}{\sqrt{n(n-1)}}$$

c) On parcourt tous les éléments de l'extradiagonale supérieure, et on les annule s'ils dépassent en module la barrière. Chaque révision garantit donc

$$|a_{pq}^{(k)}| \geqslant \alpha \geqslant \frac{\|E^{(k)}\|_E}{\sqrt{n(n-1)}} \quad ,$$

puisque la norme de l'extradiagonale ne peut que décroître.

d) Après avoir parcouru toute l'extradiagonale supérieure, on recalcule $\|E^{(k)}\|_E^2$ et on recommence avec la nouvelle barrière correspondante.

On peut s'attendre, lors d'un tel balayage, à annuler environ la moitié des termes extradiagonaux, si ceux-ci sont uniformément répartis entre le plus petit et le plus grand. Mais on n'a aucune garantie sérieuse sur ce point.

Une autre méthode, plus convaincante, consiste à utiliser une barrière flottante. On la fixe au départ à

$$\alpha^{(0)} = \frac{\|E^{(0)}\|_E}{\sqrt{n(n-1)}} \quad .$$

On parcourt toujours tous les éléments de l'extradiagonale supérieure dans un ordre déterminé. Mais dès qu'une révision est faite, on calcule

$$\|E^{(k+1)}\|_E^2 = \|E^{(k)}\|_E^2 - 2\, a_{pq}^{(k)\,2} \quad ,$$

$a_{pq}^{(k)}$ étant le terme extradiagonal annulé, et on définit la nouvelle barrière

$$\alpha^{(k+1)} = \frac{\|E^{(k+1)}\|_E}{\sqrt{n(n-1)}} \quad ,$$

qui servira pour les tests suivants. Il faut cependant prendre garde au fait que la mise à jour de $\|E^{(k)}\|_E^2$ par soustractions successives est instable numériquement. Dès que, par exemple,

$$\|E^{(k)}\|_E^2 \leqslant 4\, a_{pq}^{(k)\,2} \quad ,$$

il convient de recalculer $\|E^{(k+1)}\|_E^2$ par la voie directe.

8. CONVERGENCE QUADRATIQUE DE LA METHODE DE JACOBI

8.1 - L'évaluation

$$\|E^{(k+1)}\|_E^2 \leqq \left(1 - \frac{2}{n(n-1)}\right) \|E^{(k)}\|_E^2$$

pourrait faire croire que la convergence dela méthode de JACOBI est fort lente. En réalité, elle s'accélère rapidement lorsque la norme $\|E^{(k)}\|_E$ est devenue suffisamment petite. Pour montrer ce phénomène, appelons $A^{(k)}$ la matrice obtenue après k itérations et $A^{(k)} + H^{(k)}$ la matrice diagonale finale, dont tous les éléments sont des valeurs propres de A. Ces deux matrices ont les mêmes valeurs propres, cé qui permet de démontrer un certain nombre de propriétés intéressantes.

a) Soit A une matrice symétrique, et soient $z_{(1)}, \ldots, z_{(n)}$ une base de vecteurs propres orthonormés de A, correspondant aux valeurs propres $\lambda_1, \ldots, \lambda_n$. Soit (A + H) une matrice semblable à A, dont les vecteurs propres orthonormés, correspondant bien sûr aux mêmes valeurs propres $\lambda_1, \ldots, \lambda_n$, soient $x_{(1)}, \ldots, x_{(n)}$. L'ün quelconque de ceux-ci peut être mis sous la forme

$$x_{(i)} = \sum_k \alpha_k z_{(k)} ,$$

avec $\sum_k \alpha_k^2 = 1$. On a donc

$$x_{(i)}^T (A+H) x_{(i)} = \lambda_i ,$$

soit

$$x_{(i)}^T A x_{(i)} - \lambda_i = - x_{(i)}^T H x_{(i)} .$$

Le développement de $x_{(i)}$ donne

$$\begin{aligned} - x_{(i)}^T H x_{(i)} &= \sum_{kl} \alpha_k \alpha_l z_{(k)}^T A z_{(l)} - \lambda_i \\ &= \sum_k \alpha_k^2 \lambda_k - \lambda_i \sum_k \alpha_k^2 \\ &= \sum_k \alpha_k^2 (\lambda_k - \lambda_i) = \sum_{\lambda_k \neq \lambda_i} \alpha_k^2 (\lambda_k - \lambda_i) , \end{aligned}$$

d'où

$$x_{(i)}^T H x_{(i)} = \sum_{\lambda_k \neq \lambda_i} (\lambda_i - \lambda_k) \alpha_k^2 \qquad (1)$$

b) Dans les mêmes conditions, soient $x_{(i)}$ et $x_{(j)}$ deux vecteurs propres de (A + H) correspondant à <u>la même valeur propre</u> λ_i. Posant

$$x_{(i)} = \sum_k \alpha_k z_{(k)} \quad , \quad \sum_k \alpha_k^2 = 1 ,$$

$$x_{(j)} = \sum_k \beta_k z_{(k)} \quad , \quad \sum_k \beta_k^2 = 1 ,$$

avec la condition d'orthogonalité

$$\sum_K \alpha_k \beta_k = 0 ,$$

on obtient

$$0 = x_{(i)}^T (A+H) x_{(j)} = \sum_{kl} \alpha_k \beta_l z_{(k)}^T A z_{(l)} + x_{(i)}^T H x_{(j)} ,$$

soit

$$x_{(i)}^T H x_{(j)} = - \sum_k \alpha_k \beta_k \lambda_k = - \sum_k \alpha_k \beta_k \lambda_k + \lambda_i \sum_k \alpha_k \beta_k$$

$$= \sum_k (\lambda_i - \lambda_k) \alpha_k \beta_k$$

ou encore,

$$x_{(i)}^T H x_{(j)} = \sum_{\lambda_k \neq \lambda_j} (\lambda_i - \lambda_k) \alpha_k \beta_k . \qquad (2)$$

c) Toujours dans le cadre des mêmes hypothèses, on a encore

$$\| Hx_{(i)} \|_2^2 = \| (A+H) x_{(i)} - A x_{(i)} \|_2^2 = \| \lambda_i x_{(i)} - A x_{(i)} \|_2^2$$

et, pour

$$x_{(i)} = \sum_k \alpha_k z_{(k)} , \qquad \sum_k \alpha_k^2 = 1 ,$$

il vient

$$\| H x_{(i)} \|_2^2 = \| \lambda_i \sum_k \alpha_k z_{(k)} - \sum_k \alpha_k \lambda_k z_{(k)} \|_2^2$$

$$= \sum_k (\lambda_i - \lambda_k)^2 \alpha_k^2 ,$$

c'est-à-dire

$$\| H x_{(i)} \|_2^2 = \sum_{\lambda_k \neq \lambda_i} (\lambda_i - \lambda_k)^2 \alpha_k^2 . \qquad (3)$$

d) En combinant les relations (1) et (3), on obtient

$$x_{(i)}^T H x_{(i)} = \sum_{\lambda_k \neq \lambda_i} (\lambda_i - \lambda_k) \alpha_k^2 = \sum_{\lambda_k \neq \lambda_i} \frac{(\lambda_i - \lambda_k)^2}{(\lambda_i - \lambda_k)} \alpha_k^2 ,$$

ce qui entraîne

$$|x_{(i)}^T H x_{(i)}| \leq \frac{1}{\inf_{\lambda_i \neq \lambda_k} |\lambda_i - \lambda_k|} \sum_{\lambda_k \neq \lambda_i} (\lambda_i - \lambda_k)^2 \alpha_k^2 = \frac{\|H x_{(i)}\|_2^2}{\inf_{\lambda_i \neq \lambda_k} |\lambda_i - \lambda_k|} ,$$

soit

$$|x_{(i)}^T H x_{(i)}| \leq \frac{\|H\|_2^2}{\inf_{\lambda_i \neq \lambda_k} |\lambda_k - \lambda_i|} . \qquad (4)$$

e) A l'aide des relations (2) et (3), on déduit, pour deux vecteurs propres $x_{(i)}$ et $x_{(j)}$ correspondant à la même valeur propre i,

$$|x_{(i)}^T H x_{(j)}| \leq \sum_{\lambda_k \neq \lambda_i} |\alpha_k| |\beta_k| \frac{|\lambda_k - \lambda_i|^2}{|\lambda_k - \lambda_i|}$$

$$\leq \sum_{\lambda_k \neq \lambda_i} \frac{|\lambda_k - \lambda_i| |\alpha_k| |\lambda_k - \lambda_i| |\beta_k|}{|\lambda_k - \lambda_i|}$$

$$\leq \frac{1}{\inf_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i|} \left(\sum_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i|^2 |\alpha_k|^2 \right)^{\frac{1}{2}} \left(\sum_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i|^2 |\beta_k|^2 \right)^{\frac{1}{2}}$$

$$\frac{1}{\inf_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i|} \|H x_{(i)}\|_2 \|H x_{(j)}\|_2 \leq \frac{\|H\|_2^2}{\inf_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i|} \qquad (5)$$

Dans le problème qui nous concerne, $A + H = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, donc les vecteurs propres $x_{(i)}$ ne sont autres que les vecteurs $e_{(i)}$ de la base canonique de R^n. H est ici la différence entre la diagonale finale et la matrice $A^{(k)}$, soit explicitement:

$$H = H^{(k)} = \begin{bmatrix} \lambda_1 - a_{11}^{(k)} & - a_{12}^{(k)} & \cdots & - a_{1n}^{(k)} \\ - a_{21}^{(k)} & \lambda_2 - a_{22}^{(k)} & \cdots & - a_{2n}^{(k)} \\ \vdots & & & \vdots \\ - a_{n1}^{(k)} & \cdots & & \lambda_n - a_{nn}^{(k)} \end{bmatrix}$$

Les relations (4) et (5) ont donc la signification suivante:

$$|\lambda_i - a_{ii}^{(k)}| \quad \frac{\|H^{(k)}\|_2^2}{\inf_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i|} \qquad (4')$$

et

$$|a_{ij}^{(k)}| \leq \frac{\|H^{(k)}\|_2^2}{\inf\limits_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i|} \qquad \text{si} \quad \lambda_i = \lambda_j \quad . \qquad (5')$$

8.2 - Il est utile d'exprimer la norme de la matrice d'erreur $\|H^{(k)}\|_2$ en fonction de la norme de l'extradiagonale $\|E^{(k)}\|_E$. On a évidemment

$$\|H^{(k)}\|_2^2 \leq \|H^{(k)}\|_E^2 = \sum_i |\lambda_i - a_{ii}^{(k)}|^2 + \|E^{(k)}\|_E^2$$

et, en vertu de (4'),

$$\sum_i |\lambda_i - a_{ii}^{(k)}|^2 \leq \frac{n \|H^{(k)}\|_2^4}{\gamma^2} \quad ,$$

avec

$$\gamma = \inf_{\lambda_i \neq \lambda_k} |\lambda_k - \lambda_i| \quad . \qquad (6)$$

On en déduit l'inégalité

$$\frac{n}{\gamma^2} \|H^{(k)}\|_E^4 - \|H^{(k)}\|_E^2 + \|E^{(k)}\|_E^2 \geq 0 \ ,$$

impliquant que $\|H^{(k)}\|_E^2$ doit se trouver à l'extérieur des racines du trinôme

$$\frac{n}{\gamma^2} \xi^2 - \xi + \|E^{(k)}\|_E^2 \ .$$

Ces racines sont

$$\xi = \frac{1 \pm \sqrt{1 - \frac{4n}{\gamma^2} \|E^{(k)}\|_E^2}}{2n/\gamma^2} \quad .$$

Pour $\|E^{(k)}\|_E = 0$, on doit avoir $\|H^{(k)}\|_E = 0$, du fait de la convergence du processus. L'inégalité à prendre en considération est donc

$$\|H^{(k)}\|_E^2 \leq \frac{\gamma^2}{2n} \left(1 - \sqrt{1 - \frac{4n}{\gamma^2} \|E^{(k)}\|_E^2}\right) .$$

Elle n'a d'ailleurs de sens que si

$$\|E^{(k)}\| \leq \frac{\gamma}{2\sqrt{n}} \qquad (7)$$

Dans ce cas, on peut poser

$$\frac{2\sqrt{n}}{\gamma} \|E^{(k)}\|_E = \sin\psi \quad , \quad 0 \leq \psi \leq \pi/2 \ ,$$

ce qui implique

$$\|H^{(k)}\|_E^2 \leq \frac{\gamma^2}{2n}(1 - \cos\psi) = \frac{\gamma^2}{n}\sin^2\frac{\psi}{2} \ .$$

Mais comme $0 \leq \psi/2 \leq \pi/4$, on a

$$\sin\psi = 2\sin\frac{\psi}{2}\cos\frac{\psi}{2} \geq 2\sin\frac{\psi}{2}\frac{\sqrt{2}}{2} \ ,$$

soit

$$\sin\frac{\psi}{2} \leq \frac{\sin\psi}{\sqrt{2}} \ .$$

Il en découle

$$\|H^{(k)}\|_E^2 \leq \frac{\gamma^2}{2n}\sin^2\psi = \frac{\gamma^2}{2n} \cdot \frac{4n}{\gamma^2}\|E^{(k)}\|_E^2 \ ,$$

c'est-à-dire

$$\|H^{(k)}\|_E^2 \leq 2\|E^{(k)}\|_E^2 \qquad (8)$$

Ainsi, pour autant que la condition (7) soit vérifiée, les termes diagonaux de la matrice $A^{(k)}$ ne diffèrent des valeurs propres de la matrice A que de grandeurs $O(\|E^{(k)}\|_E^2)$.

De plus, les termes non diagonaux situés au croisement d'une ligne et d'une colonne dont les termes diagonaux convergent vers la même valeur propre sont également de grandeur $O(\|E^{(k)}\|_E^2)$.

8.3 - Examinons à présent l'ordre de grandeur des rotations θ_k , lorsque l'on révise un terme a_{pq} tel que $\lambda_p \neq \lambda_q$. On a

$$\operatorname{tg} 2\theta_k = \frac{2\,a_{pq}^{(k)}}{a_{qq}^{(k)} - a_{pp}^{(k)}}$$

Or, on sait que

$$2\,a_{pq}^{(k)\,2} \leq \|E^{(k)}\|_E^2 \ ,$$

ce qui entraîne

$$2|a_{pq}^{(k)}| \leq \sqrt{2}\,\|E^{(k)}\|_E \ .$$

On a par ailleurs

$$|a_{qq}^{(k)} - a_{pp}^{(k)}| = |a_{qq}^{(k)} - \lambda_q + \lambda_q - \lambda_p + \lambda_p - a_{pp}^{(k)}|$$

$$\geq |\lambda_q - \lambda_p| - |a_{qq}^{(k)} - \lambda_q| - |a_{pp}^{(k)} - \lambda_p|$$

$$\geq \gamma - \frac{4\,|||E^{(k)}|||_E^2}{\gamma} = \gamma\left(1 - \frac{4\,|||E^{(k)}|||_E^2}{\gamma^2}\right).$$

On en déduit

$$|\operatorname{tg} 2\theta_k| \frac{\sqrt{2}\,\frac{|||E^{(k)}|||_E}{\gamma}}{1 - \frac{4\,|||E^{(k)}|||_E^2}{\gamma^2}} = \frac{1}{2\sqrt{2}} \cdot \frac{2\,\frac{2|||E^{(k)}|||_E}{\gamma}}{1 - \frac{4\,|||E^{(k)}|||_E^2}{\gamma^2}}$$

Les majorations utilisées ci-dessus supposent (7), soit

$$\frac{2\,|||E^{(k)}|||_E}{\gamma} \leq \frac{1}{\sqrt{n}} .$$

Pour n=1, le problème aux valeurs propres n'a pas de sens; pour n=2, il se résout en une passe. <u>On peut donc supposer $n \geq 3$</u>, ce qui donne la condition

$$\frac{2\,|||E^{(k)}|||_E}{\gamma} \leq \frac{1}{\sqrt{3}} .$$

Posant alors

$$\operatorname{tg} \phi = \frac{2\,|||E^{(k)}|||_E}{\gamma} , \qquad 0 \leq \phi \leq \pi/6 ,$$

on obtient

$$|\operatorname{tg} 2\,\theta_k| \leq \frac{1}{2\sqrt{2}} \frac{2\operatorname{tg}\phi}{1 - \operatorname{tg}^2\phi} = \frac{1}{2\sqrt{2}} \operatorname{tg} 2\phi .$$

Comme $0 \leq \phi \leq \pi/6$ et comme la fonction

$$\frac{\operatorname{tg} 2\phi}{\operatorname{tg}\phi} = \frac{2}{1 - \operatorname{tg}^2\phi}$$

est croissante dans cet intervalle, on a

$$\frac{\operatorname{tg} 2\,\phi}{\operatorname{tg}\phi} \leq \frac{\operatorname{tg}\pi/3}{\operatorname{tg}\pi/6} = 3 ,$$

ce qui implique

$$|\operatorname{tg} 2\,\theta_k| \leq \frac{3}{2\sqrt{2}} \operatorname{tg}\phi$$

et

$$|\sin \theta_k| \leq |\operatorname{tg} \theta_k| \leq |\tfrac{1}{2} \operatorname{tg} 2\theta_k| \leq \frac{3}{4\sqrt{2}} \operatorname{tg} \phi = \frac{3}{2\sqrt{2}} \frac{\|E^{(k)}\|_E}{\gamma} . \quad (9)$$

Cette inégalité exprime que les angles de rotation décroissent en $O(\|E^{(k)}\|_E)$.

8.4 - Lors de la modification du terme $a_{pq}^{(k)}$ $(\lambda_p \neq \lambda_q)$, que deviennent les termes extradiagonaux situés sur sa ligne? On a

$$a_{ip}^{(k+1)} = a_{ip}^{(k)} \cos \theta_k - a_{iq}^{(k)} \sin \theta_k ,$$

d'où

$$|a_{ip}^{(k+1)} - a_{ip}^{(k)}| \leq |a_{ip}^{(k)}| (1 - \cos \theta_k) + |a_{iq}^{(k)}| |\sin \theta_k|$$

$$\sqrt{a_{ip}^{(k)\,2} + a_{iq}^{(k)\,2}} \quad \sqrt{1 - 2\cos \theta_k + \cos^2 \theta_k + \sin^2 \theta_k}$$

$$\frac{\|E^{(k)}\|_E}{\sqrt{2}} \cdot 2 \left|\sin \frac{\theta_k}{2}\right| .$$

Comme $0 \leq |\theta_k| \leq \pi/4$, on a

$$\sin |\theta_k| = 2 \sin (|\theta_k|/2) \cos (\theta_k/2) \geq 2 \sin (|\theta_k|/2) \cos (\pi/8) ,$$

d'où

$$\sin \frac{|\theta_k|}{2} \leq \frac{\sin |\theta_k|}{2 \cos \frac{\pi}{8}}$$

et

$$|a_{ip}^{(k+1)} - a_{ip}^{(k)}| \leq \frac{\|E^{(k)}\|_E |\sin \theta_k|}{\sqrt{2} \cos (\pi/8)} .$$

Tenant compte de la majoration (9), on obtient

$$|a_{ip}^{(k+1)} - a_{ip}^{(k)}| \leq \frac{3}{4 \cos \frac{\pi}{8}} \frac{\|E^{(k)}\|_E^2}{\gamma} = 0{,}8118\ldots \frac{\|E^{(k)}\|_E^2}{\gamma}$$

$$\leq \frac{\|E^{(k)}\|_E^2}{\gamma} \quad (10)$$

Ce raisonnement se transpose entièrement pour un terme situé sur la colonne du terme modifié. La majoration est la même. On costate donc que les termes de la ligne et de la colonne da l'élément révisé ne subissent que des modifications $O(\|E^{(k)}\|_E^2)$.

8.5 - Nous sommes à présent en mesure de déterminer le véritable ordre de convergence de la méthode de JACOBI.

Mais tout d'abord, une restriction quant à la recherche de l'élément non diagonal à réviser. Nous dirons que cette recherche est judicieuse si elle garantit constamment que l'élément révisé $a_{pq}^{(k)}$ vérifie

$$|a_{pq}^{(k)}| \geq \frac{||E^{(k)}||_E}{\sqrt{n(n-1)}} \quad .$$

C'est le cas de la technique de l'élément max mal, de celle de l'élément optimal, ainsi que de certaines techniques de barrières. Nous nous limiterons dans ce qui suit à ne considérer que des algorithmes à recherche judicieuse.

Cela étant, dès que

$$||E^{(k)}||_E \leq \frac{\gamma}{2\sqrt{n}} \quad ,$$

on a les faits suivants:

a) La révision d'un terme non diagonal $a_{pq}^{(k)}$ avec $\lambda_p \neq \lambda_q$ entraîne une modification des termes de sa ligne et de sa colonne qui vérifie

$$|a_{ir}^{(k+1)} - a_{ir}^{(k)}| \leq \frac{||E^{(k)}||_E^2}{\gamma} \quad , \quad r = p, q \; . \qquad (11)$$

b) Un terme non diagonal situé au croisement d'une ligne et d'une colonne dont les termes diagonaux convergent vers la même valeur propre est nécessairement inférieur à $2\,||E^{(k)}||_E^2 / \gamma$. Par conséquent, il ne peut plus jamais être révisé dès le moment où

$$\frac{||E^{(k)}||_E}{\sqrt{n(n-1)}} \geq \frac{2\,||E^{(k)}||_E^2}{\gamma} \quad ,$$

c'est-à-dire

$$||E^{(k)}||_E \leq \frac{\gamma}{2\sqrt{n(n-1)}} \qquad (12)$$

Venons-en à l'évaluation de l'ordre de convergence. L'extradiagonale ne contient que $N = \frac{n(n-1)}{2}$ termes différents. Par conséquent, ,dans un ensemble de $(N + 1)$ itérations, de numéros k, k+1, ..., k + N, un terme non diagonal au moins sera révisé plus d'une fois. Parmi les couples d'itérations où la révision porte sur le même élémént, choisissons le couple (k+r, k+s) le plus proche, c'est-à-dire tel que (s - r) soit le plus petit. Il ne peut donc y avoir, entre (k + r) et (k + s), de répétition.

A l'itération (k + s), pour autant que l'on utilise une recherche judicieuse, le terme $a_{pq}^{(k+s-1)}$ vérifie

$$|a_{pq}^{(k+s-1)}| \geq \frac{\|E^{(k+s-1)}\|_E}{\sqrt{n(n-1)}} .$$

Or, depuis l'annulation de a_{pq} à l'itération (k + r), on n'a pu réviser qu'une seule fois chaque terme de sa ligne et de sa colonne, ,ce qui fait un maximum de (2n-2) éléments, correspondant aux itérations $k+l_1$, ... $k+l_t$, $t \leq 2n-2$. En vertu de l'inégalité (11), $a_{pq}^{(k+s-1)}$ doit donc vérifier

$$|a_{pq}^{(k+s-1)}| \leq \frac{1}{\gamma} \left(\|E^{(k+l_1)}\|_E^2 + \ldots + \|E^{(k+l_t)}\|_E^2 \right) \leq \frac{2n-2}{\gamma} \|E^{(k+r)}\|_E^2 ,$$

puisque la norme de l'extradiagonale croît toujours. On a donc

$$\frac{\|E^{(k+s-1)}\|_E}{\sqrt{n(n-1)}} \leq \frac{2n-2}{\gamma} \|E^{(k+r)}\|_E^2$$

soit

$$\|E^{(k+s-1)}\|_E \leq \frac{(2n-2)\sqrt{n(n-1)}}{\gamma} \|E^{(k+r)}\|_E^2 \leq \frac{2n^2}{\gamma} \|E^{(k+r)}\|_E^2 .$$

mais alors, en vertu de la décroissance de la norme de l'extradiagonale,

$$\|E^{(k+N)}\|_E \leq \|E^{(k+s-1)}\|_E \leq \frac{2n^2}{\gamma} \|E^{(k+r)}\|_E^2 \leq \frac{2n^2}{\gamma} \|E^{(k)}\|_E^2 .$$

Il y a donc convergence quadratique, quand on considère des cycles de $\frac{n(n-1)}{2}$ itérations.

9. METHODE DE LA PUISSANCE

9.1 - La méthode de la puissance permet de calculer le vecteur propre relatif à la valeur propre de plus grand module. Rangeons donc les valeurs propres par ordre des modules décroissants:

$$|\lambda_1| \geq |\lambda_2| \geq \ldots\ldots \geq |\lambda_n| ,$$

et considérons un vecteur $x^{(0)}$ arbitraire. On peut le décomposer dans la base des vecteurs propres orthonormés $z_{(i)}$:

$$x^{(0)} = \sum_{i=1}^{n} \alpha_i z_{(i)} \quad , \quad \|x^{(0)}\|_2^2 = \sum_{i=1}^{n} \alpha_i^2 .$$

Calculons successivement

$$x^{(1)} = A\, x^{(0)} = \sum_{i=1}^{n} \alpha_i z_{(i)} = \sum_{i=1}^{n} \alpha_i \lambda_i z_{(i)} ,$$

$$x^{(2)} = A\,x^{(1)} = A^2\,x^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^2 z_{(i)} ,$$

.....

$$x^{(p)} = A^p\,x^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^p z_{(i)} .$$

On a alors

$$\frac{x^{(p)T}\,x^{(p)}}{x^{(p)T}\,x^{(p-1)}} = \frac{\sum_i \alpha_i^2 \lambda_i^{2p}}{\sum_i \alpha_i^2 \lambda_i^{2p-1}} = \lambda_1 \frac{\sum_i \alpha_i^2 (\lambda_i/\lambda_1)^{2p}}{\sum_i \alpha_i^2 (\lambda_i/\lambda_1)^{2p-1}} .$$

Si la première valeur propre est <u>strictement</u> supérieure à la seconde, on a

$$\frac{x^{(p)T}\,x^{(p)}}{x^{(p)T}\,x^{(p-1)}} = \lambda_1 \frac{\alpha_1^2 + \sum_{i>1} (\lambda_i/\lambda_1)^{2p} \alpha_i^2}{\alpha_1^2 + \sum_{i>1} (\lambda_i/\lambda_1)^{2p-1} \alpha_i^2}$$

$$= \lambda_1 \frac{1 + O(|\lambda_2/\lambda_1|^{2p})}{1 + O(|\lambda_2/\lambda_1|^{2p-1})} = \lambda_1 (1 + O(|\lambda_2/\lambda_1|^{2p-1})) .$$

Le rapport en question converge donc vers λ_1 avec une vitesse dépendant essentiellement du rapport $|\lambda_2/\lambda_1|$. Ceci suppose cependant $\lambda_1 \neq 0$. On pourrait en effet imaginer le cas malheureux où le vecteur de départ serait orthogonal à $z_{(1)}$. Dans ce cas, on convergerait théoriquement vers λ_2. Si c'est bien ainsi que démarre le processus, les erreurs d'arrondi finissent en général par faire apparaître une composante selon le premier vecteur propre. Après un début de convergence vers le second, le processus se met à diverger pour converger en définitive vers le premier. Cependant, un plus grand nombre d'itérations sont nécessaires et, en définitive, on n'est pas toujours sûr que cela se produira effectivement. C'est pourquoi le vecteur de départ doit être choisi avec bonheur. Une solution assez sûre consiste à prendre pour point de départ un <u>vecteur aléatoire</u>, c'est-à-dire dont les composantes sont des nombres aléatoires ou pseudo-aléatoires.

Lorsque la valeur propre est grande ou petite en module, la norme du vecteur croît ou décroît assez vite. Il faut donc normer le vecteur de temps à autre. On peut aussi le diviser systématiquement par la valeur propre approchée

$$\tilde{\lambda}_1 = \frac{x^{(p)T}\,x^{(p)}}{x^{(p)T}\,x^{(p-1)}} .$$

On notera qu'il y a <u>convergence du vecteur propre</u>. Ceci demande une explication. Un vecteur propre n'est défini qu'à une constante près, si

bien qu'il ne faut considérer que l'angle entre le vecteur approché et le véritable vecteur propre. Soit S_{λ_1} le sous-espace propre relatif à la valeur propre λ_1. Le vecteur $x^{(p)}$ peut (fig. 3) être décomposé en sa projection dand S_{λ_1} et sa projection orthogonale à S_{λ_1} :

$$x^{(p)} = pr_{S_{\lambda_1}} x^{(p)} + pr_{S_{\lambda_1}\perp} x^{(p)} .$$

L'angle θ entre les deux vecteurs $x^{(p)}$ et $pr_{S_{\lambda_1}} x^{(p)}$ peut être considéré comme une bonne mesure de l'erreur sur le vecteur propre. On considérera donc le sinus de cet angle:

$$\sin \theta = \frac{\| pr_{S_{\lambda_1}\perp} x^{(p)} \|_2}{\| x^{(p)} \|_2} .$$

ou encore, sa tangente:

$$\operatorname{tg} \theta = \frac{\| pr_{S_{\lambda_1}\perp} x^{(p)} \|_2}{\| pr_{S_{\lambda_1}} x^{(p)} \|_2} .$$

Ces deux mesures sont équivalentes, car $\left\{ \begin{array}{l} \sin\theta \to 0 \\ \operatorname{tg}\theta \to 0 \end{array} \right\}$ entraîne $\theta \to 0$, et donc $\left\{ \begin{array}{l} \operatorname{tg}\theta \to 0 \\ \sin\theta \to 0 \end{array} \right\}$.

Pour la méthode de la puissance, on a

$$\left\{ \begin{array}{l} \| pr_{S_{\lambda_1}\perp} x^{(p)} \|_2^2 = \sum_{i>1} \alpha_i^2 \lambda_i^{2p} \\ \\ \| pr_{S_{\lambda_1}} x^{(p)} \|_2^2 = \alpha_1^2 \lambda_1^{2p} \end{array} \right. ,$$

d'où

$$\operatorname{tg} \theta = \left(\frac{\sum_{i>1} \alpha_i^2 \lambda_i^{2p}}{\alpha_1^2 \lambda_1^{2p}} \right)^{\frac{1}{2}} \leq | \lambda_2 / \lambda_1 |^p \ \frac{\left[\| x^{(0)} \|^2 - \alpha_1^2 \right]^{\frac{1}{2}}}{| \alpha_1 |} ,$$

ce qui prouve la convergence angulaire.

9.2 - Que se passe-t-il si r valeurs propres sont confondues? Dans ce cas, on a donc

$$\lambda_1 = \lambda_2 = \ldots = \lambda_r \quad , \quad | \lambda_1 | > | \lambda_{r+1} | .$$

Il vient alors

$$x^{(0)} = \sum_{i=1}^{r} \alpha_i z_{(i)} + \sum_{i=r+1}^{n} \alpha_i z_{(i)}$$

et

$$x^{(p)} = \sum_{i=1}^{r} \alpha_i \lambda_i^p z_{(i)} + \sum_{i=r+1}^{n} \alpha_i \lambda_i^p z_{(i)}$$

$$= \lambda_1^p \sum_{i=1}^{r} \alpha_i z_{(i)} + \sum_{i=r+1}^{n} \alpha_i \lambda_i^p z_{(i)} .$$

Il y a donc en théorie convergence vers le vecteur propre $\sum_{i=1}^{r} \alpha_i z_{(i)}$,

puisque

$$\operatorname{tg} \theta = \frac{\| pr_{S_{\lambda_1 \perp}} x^{(p)} \|_2}{\| pr_{S_{\lambda_1}} x^{(p)} \|_2} = \frac{\| \sum_{i>r} \alpha_i \lambda_i^p z_{(i)} \|_2}{|\lambda_1|^p \| \sum_{i=1}^{r} \alpha_i z_{(i)} \|_2} = \left| \frac{\lambda_{r+1}}{\lambda_1} \right|^p \frac{| \sum_{i \le r} \alpha_i^2 |^{\frac{1}{2}}}{| \sum_{i=1}^{r} \alpha_i^2 |^{\frac{1}{2}}} .$$

En fait, le vecteur tourne constamment dans le sous-espace propre du fait des erreurs d'arrondi. On s'arrête donc quand la valeur propre

$$\lambda_1 = \frac{x^{(p)T} x^{(p)}}{x^{(p)T} x^{(p-1)}} = \lambda_1 \frac{\sum_{i=1}^{r} \alpha_i^2 + \sum_{i>r} (\lambda_i/\lambda_1)^{2p} \alpha_i^2}{\sum_{i=1}^{r} \alpha_i^2 + \sum_{i>r} (\lambda_i/\lambda_1)^{2p-1} \alpha_i^2}$$

$$= \lambda_1 \left(1 + O(|\lambda_{r+1}/\lambda_1|^{2p-1}) \right)$$

a convergé. Le vecteur propre obtenu est donc <u>un des vecteurs</u> du sous-espace propre.

9.3 - Les cas insolubles pratiquement par la méthode de la puissance sont:

a) Des valeurs propres trop voisines (convergence lente)

b) Des valeurs propres égales en grandeur, mais de signes opposés. Ce dernier problème ne se présente pas dans le cas de matrices définies non négatives.

9.4 - <u>Recherche des valeurs propres suivantes par déflation orthogonale</u>

Pour trouver le mode et la valeur propre qui suivent, il faut partir d'un vecteur orthogonal à $z_{(1)}$, et <u>le maintenir constamment orthogonal à $z_{(1)}$ tout au cours du processus</u>, car les erreurs d'arrondi risquent de ramener sans cesse une petite composante de ce mode. A cette fin, on

utilise un opérateur de projection

$$O_1 = I - \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2} .$$

Un vecteur quelconque, quand on lui applique O_1, se voit orthogonalisé à $z_{(1)}$:

$$O_1 x = I x - \frac{z_{(1)}\, z_{(1)}^T\, x}{\|z_{(1)}\|_2^2} = x - \frac{x^T z_{(1)}}{\|z_{(1)}\|_2^2}\, z_{(1)}$$

est orthogonal à $z_{(1)}$, comme on le vérifie aisément. On notera que

$$A\, O_1 = A\, (I - \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2}) = A - \lambda_1 \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2}$$

et

$$O_1\, A = (I - \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2})\, A = A - \lambda_1 \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|^2} ,$$

c'est-à-dire que $A\, O_1 = O_1\, A$.

On a encore

$$O_1\, O_1 = (I - \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2})\, (I - \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2})$$

$$= I - 2\, \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2} + \frac{z_{(1)}\, z_{(1)}^T\, z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^4} = O_1 .$$

(Idempotence). Enfin, $O_1^T = O_1$ (symétrie).

Pour notre problème, il s'agit de trouver un vecteur $z_{(2)}$ tel que

$$A\, z_{(2)} = z_{(2)} \quad \text{et} \quad z_{(2)}^T\, z_{(1)} = 0 .$$

On a donc

$$O_1\, z_{(2)} = z_{(2)} = \frac{z_{(2)}^T\, z_{(1)}}{\|z_{(1)}\|_2^2}\, z_{(2)} ,$$

et il est équivalent de chercher le vecteur $z_{(2)}$ tel que

$$A\, O_1\, z_{(2)} = z_{(2)} .$$

On notera que la matrice $A\, O_1$ vérifie

$$O_1^T\, A\, O_1 = O_1\, O_1\, A = O_1\, A = A\, O_1 ,$$

donc elle est symétrique. On itère donc sur la matrice

$$A_{(1)} = A\, O_1 ,$$

ce qui a pour effet de filtrer à chaque itération toute composante en $z_{(1)}$. La convergence est alors assurée si $|\lambda_2| > |\lambda_3|$.

Pour trouver le vecteur propre relatif à la troisième valeur propre, on itère avec la matrice $A\, O_1\, O_2$. On notera que

$$O_1\, O_2 = \left(I - \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2}\right)\left(I - \frac{z_{(2)}\, z_{(2)}^T}{\|z_{(2)}\|_2^2}\right)$$

$$= I - \frac{z_{(1)}\, z_{(1)}^T}{\|z_{(1)}\|_2^2} - \frac{z_{(2)}\, z_{(2)}^T}{\|z_{(2)}\|_2^2} = O_{12} .$$

Plus généralement, on cherche le $(r+1)^e$ vecteur propre en utilisant pour l'itération la matrice

$$A_{(r)} = A\, O_1 \cdots O_r ,$$

avec

$$O_1 \cdots O_r = I - \sum_{k=1}^{r} \frac{z_{(i)}\, z_{(i)}^T}{\|z_{(i)}\|_2^2}$$

10. ITERATION SUR UN SOUS-ESPACE

Une variante très intéressante de la méthode de la puissance a été introduite par BAUER [14]. L'idée fondamentale est d'itérer non plus sur un vecteur, mais sur plusieurs à la fois, tout en les forçant à rester orthogonaux. C'est pourquoi on parle également d'itération simultanée, qoique cette expression n'ait aucun sens précis.

On choisit donc $r \leqslant n$ vecteurs aléatoires, que l'on orthonorme entre eux. Soient $x_{(1)}, \ldots, x_{(r)}$ ces vecteurs orthonormés. A chacun d'eux, on applique un certain nombre de fois la matrice A; après chacune de ces opérations, on les réorthogonalise. Les vecteurs obtenus après p itérations sont donc des combinaisons linéaires de $A^p x_{(1)}, \ldots, A^p x_{(n)}$, orthogonales entre elles.

La théorie de l'itération sur un sous-espace repose sur le lemme suivant:

Supposons qu'il n'existe aucune combinaison linéaire des vecteurs de départ $x_{(1)}, \ldots, x_{(r)}$ qui soit orthogonale aux r premiers vecteurs propres $z_{(1)}, \ldots, z_{(r)}$ (on range les vecteurs propres par ordre de valeurs propres décroissantes en module) . Il existe alors r combinaisons linéaires $u_{(1)}, \ldots, u_{(r)}$ des vecteurs de départ qui vérifient

$$u_{(i)}^T z_{(j)} = \delta_{ij} .$$

Pour le montrer, écrivons les vecteurs $u_{(i)}$ sous la forme

$$u_{(i)} = \sum_{k=1}^{r} \beta_{ik} x_{(k)} .$$

Les conditions à vérifier par les vecteurs $u_{(i)}$ s'écrivent alors

$$u_{(i)}^T z_{(j)} = \sum_{k=1}^{r} \beta_{ik} x_{(k)}^T z_{(j)} = \delta_{ij} .$$

Introduisons les matrices

$$B = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1r} \\ \vdots & & \vdots \\ \beta_{r1} & \cdots & \beta_{rr} \end{bmatrix} , \quad C = \begin{bmatrix} x_{(1)}^T z_{(1)} & \cdots & x_{(1)}^T z_{(r)} \\ \vdots & & \vdots \\ x_{(r)}^T z_{(1)} & \cdots & x_{(r)}^T z_{(r)} \end{bmatrix} .$$

Les conditions ci-dessus s'écrivent alors $BC = I$, ce qui signifie que la matrice B est l'inverse de la matrice C. Il suffit donc de montrer que C est inversible, ce qui est évident, car dans le cas contraire, il existerait une combinaison linéaire nulle de ses lignes, ce qui s'écrirait

$$(\mu_1 x_{(1)}^T + \cdots + \mu_r x_{(r)}^T) z_{(j)} = 0 , \qquad j = 1, \ldots, r ,$$

en contradiction avec l'hypothèse.

Développons donc ces vecteurs $u_{(i)}$ dans la base des vecteurs propres: il vient, compte tenu du lemme que nous venons de démontrer, et après normation des $u_{(i)}$,

$$u_{(i)} = \gamma_i z_{(i)} + \sum_{k=r+1}^{n} \xi_{ik} z_{(k)} .$$

Après p applications de la matrice A, on obtient

$$A^p u_{(i)} = \gamma_i \lambda_i^p z_{(i)} + \sum_{k=r+1}^{n} \xi_{ik} \lambda_k^p z_{(k)} .$$

Un nombre p suffisant d'itérations donne un $A^p u_{(i)}$ approchant très bien $z_{(i)}$, car

$$\frac{\| pr_{z_{(i)}\perp} A^p u_{(i)} \|_2}{\| pr_{z_{(i)}} A^p u_{(i)} \|_2} = \frac{(\sum_{k=r+1}^{n} \xi_{ik}^2 \lambda_k^{2p})^{\frac{1}{2}}}{|\gamma_i| \, |\lambda_i|^p} \leq |\lambda_{r+1}/\lambda_i|^p \frac{(1-\gamma_i^2)^{\frac{1}{2}}}{|\gamma_i|} .$$

De plus, les vecteurs $A^p u_{(i)}$ <u>tendent à devenir perpendiculaires entre</u> eux, car, si $|\lambda_j| < |\lambda_i|$,

$$\cos\,(A^p u_{(i)},\, A^p u_{(j)}) = \frac{(A^p u_{(i)})^T A^p u_{(j)}}{\|A^p u_{(i)}\|_2 \, \|A^p u_{(j)}\|_2}$$

$$= \frac{\sum_{k=r+1}^{n} \xi_{ik}\, \xi_{jk}\, \lambda_k^{2p}}{(\gamma_i^2 \lambda_i^{2p} + \sum_{k=r+1}^{n} \xi_{ik}^2 \lambda_k^{2p})^{\frac{1}{2}} \, (\gamma_j^2 \lambda_j^{2p} + \sum_{k=r+1}^{n} \xi_{jk}^2 \lambda_k^{2p})^{\frac{1}{2}}}$$

$$\leqslant \frac{|\lambda_{r+1}|^{2p} \sum_{k=r+1}^{n} |\xi_{ik}|\, |\xi_{jk}|}{|\gamma_i|\, |\lambda_i|^p\, |\gamma_j|\, |\lambda_j|^p}$$

$$\leqslant \frac{|\lambda_{r+1}|^{2p}}{|\lambda_j|^{2p}} \quad \frac{(\sum_{k=r+1}^{n} \xi_{ik})^{\frac{1}{2}} \, (\sum_{k=r+1}^{n} \xi_{jk}^2)^{\frac{1}{2}}}{|\gamma_i|\, |\gamma_j|}$$

$$\left| \frac{\lambda_{r+1}}{\lambda_j} \right|^{2p} \cdot \frac{(1-\gamma_i^2)^{\frac{1}{2}} \, (1-\gamma_j^2)^{\frac{1}{2}}}{|\gamma_i|\, |\gamma_j|} \quad ,$$

soit une grandeur du second ordre en l'erreur sur le mode qui a le moins bien convergé.

En pratique, après un certain nombre p d'applications de A, on recherche les vecteurs propres par projection du problème aux valeurs propres dans le sous-espace (méthode de RAYLEIGH-RITZ), c'est-à-dire que si $y_{(1)}, \ldots, y_{(r)}$ en forment une base (les r vecteurs calculés), on cherche les modes propres sous la fomme

$$\tilde{z}_{(i)} = \sum_k \alpha_k\, y_{(k)} \,.$$

L'équation à résoudre,

$$A\, \tilde{z}_{(i)} = \tilde{\lambda}\, z_{(i)} \,,$$

se développe donc en

$$\sum_k \alpha_k\, A\, y_{(k)} = \tilde{\lambda} \sum_k \alpha_k\, y_{(k)} \,,$$

et, comme les vecteurs construits $y_{(k)}$ sont orthonormés, la multiplication de cette équation par $y_{(l)}^T$ donne

$$\sum_k y_{(l)}^T\, A\, y_{(k)}\, \alpha_l = \tilde{\lambda}\, \alpha_l \,.$$

Il s'agit d'un problème aux valeurs propres pour la matrice F, de dimension r x r, définie par

$$f_{lk} = y_{(l)}^T\, A\, y_{(k)} \,,$$

les vecteurs propres contenant les coefficients des $y_{(i)}$ dans le développement de chaque vecteur propre approché du problème initial. Généralement, la taille r de ce problème aux valeurs propres réduit est très inférieure à n. Pour le résoudre, on peut utiliser la méthode de JACOBI. Le premier mode propre obtenu est en fait la projection de $z_{(1)}$ dans le sous-espace. Les autres lui sont orthogonaux et diffèrent donc des $A^p u_{(i)}$ d'un angle $0(|\lambda_i / \lambda_{r+1}|^{2p})$, soit du second ordre par rapport à l'erreur principale.

Par rapport à la méthode de la puissance simple, l'itération sur un sous-espace possède les avantages suivants:

a) Les premiers modes convergent en $|\lambda_i / \lambda_{r+1}|^p$ et non plus $|\lambda_i / \lambda_{i+1}|^p$. (Accélération de la convergence).

b) Les cas de dégénérescence ne posent aucun problème, pour autant que la dimension du sous-espace r soit supérieure à la multiplicité de la valeur propre.

c) Un nombre $s < r$ de vecteurs propres à valeur propre proche se traite sans difficulté.

d) Le cas de deux valeurs propres égales en grandeur mais de signes différents peut être traité aisément.

11. PRINCIPE DE RAYLEIGH

Etant donné un vecteur x et une matrice symétrique A, on appelle quotient de RAYLEIGH de ce vecteur pour la matrice A, le nombre

$$\mathcal{R}_A(x) = \frac{x^T A x}{\|x\|_2^2} \quad .$$

Lorsque x est un vecteur propre de A, pour la valeur propre λ, on a

$$\mathcal{R}_A(x) = \frac{x^T \lambda x}{x^T x} = \lambda \quad .$$

Mais il y a plus. Appelons S_{λ_i} le sous-espace propre relatif à la valeur propre λ_i. Un vecteur quelconque x peut être développé en vecteurs propres:

$$x = \sum_{k=1}^{n} \alpha_k z_{(k)} \ ,$$

$z_{(k)}$ étant des vecteurs propres orthonormés. On a donc

$$x^T A x = \sum_{k=1}^{n} \alpha_k^2 z_{(k)}^T A z_{(k)} = \sum_{k=1}^{n} \alpha_k \lambda_k^2$$

et

$$\|x\|_2^2 = \sum_{k=1}^{n} \alpha_k^2 \ ,$$

ce qui entraîne

$$x^T A x - \lambda_i \|x\|_2^2 = \sum_{k=1}^{n} \alpha_k^2 (\lambda_k - \lambda_i) = \sum_{\lambda_k \neq \lambda_i} \alpha_k^2 (\lambda_k - \lambda_i)$$

et

$$|x^T A x - \lambda_i \|x\|_2^2| \leq \sup_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i| \sum_{\lambda_k \neq \lambda_i} \alpha_k^2$$

$$\sup_{\lambda_k \neq \lambda_i} |\lambda_k - \lambda_i| \, \|pr_{S_{\lambda_i}^{\perp}} x\|_2^2 ,$$

soit

$$|\mathcal{R}_A(x) - \lambda_i| \leq \sup_{\lambda_i \neq \lambda_k} |\lambda_k - \lambda_i| \, \frac{\|pr_{S_{\lambda_i}^{\perp}} x\|_2^2}{\|x\|_2^2} .$$

Ce résultat constitue le <u>principe de RAYLEIGH</u>: <u>Etant donné un vecteur x approchant le sous-espace propre</u> S_{λ_i} <u>avec une erreur angulaire du premier ordre, son quotient de Rayleigh approche</u> λ_i <u>au second ordre</u>.

Ce résultat signifie que l'on peut, à partir d'approximations relativement grossières des vecteurs propres, obtenir par le quotient de Rayleigh une approximation assez raisonnable de la valeur propre. Il sert de fondement à de nombreuses méthodes approchées.

12. EVALUATION A POSTERIORI DE L'ERREUR SUR UN VECTEUR PROPRE

12.1 - Etant donné un vecteur $\tilde{x}$ dont on pense qu'il approche le sous-espace S_{λ_i}, et la valeur propre approchée $\tilde{\lambda}_i$, comment mesurer l'erreur? On calcule le <u>résidu</u>

$$r = A \tilde{x} - \tilde{\lambda}_i \tilde{x}.$$

Le vecteur x se décompose dans la base des vecteurs propres orthonormés $z_{(i)}$ de A, sous la forme

$$\tilde{x} = \sum_{k=1}^{n} \alpha_k z_{(k)} ,$$

ce qui entraîne

$$r = \sum_{k=1}^{n} \alpha_k \lambda_k z_{(k)} - \tilde{\lambda}_i \sum_k \alpha_k z_{(k)} = \sum_k \alpha_k (\lambda_k - \tilde{\lambda}_i) z_{(k)} .$$

Il vient donc

$$\|r\|_2^2 = \sum_{k=1}^{n} \alpha_k^2 \, (\lambda_k - \tilde{\lambda}_i)^2 \geqslant \sum_{\lambda_k \neq \lambda_i} \alpha_k^2 \, (\lambda_k - \tilde{\lambda}_i)^2$$

$$\geqslant \inf_{\lambda_k \neq \lambda_i} |\lambda_k - \tilde{\lambda}_i|^2 \sum_{\lambda_k \neq \lambda_i} \alpha_k^2 \quad ,$$

c'est-à-dire

$$\|pr_{S_{\lambda_i \perp}} x\|_2 \leqslant \frac{\|r\|_2}{\inf\limits_{\lambda_k \neq \lambda_i} |\lambda_k - \tilde{\lambda}_i|}$$

L'erreur sur le vecteur vérifie donc

$$\frac{\|pr_{S_{\lambda_i \perp}} x\|_2}{\|x\|_2} \leqslant \frac{1}{\inf\limits_{\lambda_k \neq \lambda_i} |\lambda_k - \tilde{\lambda}_i|} \cdot \frac{\|r\|_2}{\|x\|_2} \, .$$

Elle est donc du même ordre de grandeur que le résidu relatif.

Le calcul effectif du coefficient est assez malaisé, car il nécessite la connaissance des valeurs propres ou de relations du type

$$\inf_{\lambda_k \neq \lambda_i} |\lambda_k - \tilde{\lambda}_i| \geqslant \beta \quad ,$$

permettant d'utiliser β au dénominateur.

12.2 - En ce qui concerne l'erreur sur la valeur propre, on peut en premier lieu noter que

$$\|r\|_2^2 = \|A\tilde{x} - \tilde{\lambda}_i \tilde{x}\|_2^2 = \sum_j (\lambda_j - \tilde{\lambda}_i)^2 \alpha_i^2 \geqslant \inf_j (\lambda_j - \tilde{\lambda}_i) \sum_j \alpha_j^2 \quad ,$$

soit

$$\inf_j |\lambda_j - \tilde{\lambda}_i| \leqslant \frac{\|r\|_2}{\|\tilde{x}\|_2} \, .$$

C'est la <u>borne d'erreur de KRYLOV et BOGOLIOUBOV</u> : On peut affirmer qu'<u>il existe une valeur propre λ_k telle que</u>

$$\tilde{\lambda}_i - \frac{\|r\|_2}{\|\tilde{x}\|_2} \leqslant \lambda_k \leqslant \tilde{\lambda}_i + \frac{\|r\|_2}{\|\tilde{x}\|_2} \, .$$

Si le vecteur propre $\tilde{x}$ est assez proche de $z_{(i)}$,la valeur propre λ_k la plus proche de $\tilde{\lambda}_i$ sera λ_i; mais dans le cas général, on ne peut nullement garantir que ce sera une valeur propre plutôt qu'une autre, spécialement quand elles sont peu espacées. L'avantage de cette borne

est qu'elle ne présuppose aucune connaissance du spectre de la matrice.

12.3 - Supposons un vecteur approché $\tilde{x}$ connu. Pour obtenir le meilleur encadrement de la valeur propre, c'est-à-dire le plus serré, il convient visiblement de trouver la valeur approchée $\tilde{\lambda}_{opt}$ qui minimise le résidu. Or,

$$\frac{d}{d\tilde{\lambda}} \|A\tilde{x} - \tilde{\lambda}\tilde{x}\|_2^2 = \frac{d}{d\tilde{\lambda}} \left\{ \|A\tilde{x}\|_2^2 - 2\tilde{\lambda}\, \tilde{x}^T A \tilde{x} + \tilde{\lambda}^2 \|\tilde{x}\|_2^2 \right\}$$

$$= 2\tilde{\lambda}\|\tilde{x}\|_2^2 - 2\,\tilde{x}^T A \tilde{x} \quad ,$$

ce qui donne la valeur optimale

$$\tilde{\lambda}_{opt} = \mathcal{R}_A(\tilde{x}) \quad .$$

La dérivée seconde de $\|r\|_2^2$ étant positive, il s'agit bien d'un minimum. C'est donc le quotient de RAYLEIGH qui constitue la valeur optimale de la valeur propre approchée, minimisant le résidu.

12.4 - Une borne d'erreur plus serrée peut être obtenue en remarquant que l'expression

$$(A\tilde{x} - \xi\tilde{x})^T (A\tilde{x} - \eta\tilde{x}) = \sum_i (\lambda_i - \xi)(\lambda_i - \eta)\alpha_i^2 \geq 0$$

pour autant qu'aucune valeur propre ne soit située entre les deux nombres ξ et η. Cette condition est sûrement vérifiée si l'on pose

$$\begin{cases} \xi = \lambda_k = \text{valeur propre la plus proche de } \mathcal{R}_A(\tilde{x}) \\ \eta = \mathcal{R}_A(\tilde{x}) - \inf\limits_{\lambda_i \neq \lambda_k} |\lambda_i - \mathcal{R}_A(x)| \operatorname{sign}(\lambda_k - \mathcal{R}_A(x)) \end{cases}$$

On a lors

$$0 \leq \|A\tilde{x}\|_2^2 - (\xi+\eta)\,\tilde{x}^T A \tilde{x} + \xi\eta \|\tilde{x}\|_2^2$$

$$\leq \|A\tilde{x}\|_2^2 - 2\mathcal{R}_A(\tilde{x})\,\tilde{x}^T A\tilde{x} + \mathcal{R}_A^2(\tilde{x})\|\tilde{x}\|_2^2 + (2\mathcal{R}_A(\tilde{x}) - \xi - \eta)\,\tilde{x}^T A\tilde{x}$$

$$+ (\xi\eta - \mathcal{R}_A^2(\tilde{x}))\|\tilde{x}\|_2^2$$

$$\leq \|r\|_2^2 + (\mathcal{R}_A^2(\tilde{x}) - (\xi+\eta)\mathcal{R}_A(x) + \xi\eta)\|\tilde{x}\|_2^2$$

$$\leq \|r\|_2^2 + (\mathcal{R}_A(\tilde{x}) - \xi)(\mathcal{R}_A(\tilde{x}) - \eta)\|\tilde{x}\|_2^2$$

$$\leq \|r\|_2^2 - |\mathcal{R}_A(\tilde{x}) - \lambda_k| \inf_{\lambda_i \neq \lambda_k} |\mathcal{R}_A(\tilde{x}) - \lambda_i| \|x\|_2^2 \quad ,$$

ce qui entraîne

$$|\mathcal{R}_A(\tilde{x}) - \lambda_k| \leq \frac{1}{\inf\limits_{\lambda_i \neq \lambda_k} |\mathcal{R}_A(\tilde{x}) - \lambda_i|} \quad \frac{\|r\|_2^2}{\|\tilde{x}\|_2^2} \quad .$$

Cette erreur est du second ordre en $\|r\|_2$, en accord avec le principe de Rayleigh, dont le présent résultat constitue un complément. La borne ainsi obtenue, due à TEMPLE et KATO, est cependant assez peu maniable, car il faut connaître λ_i . Mais si l'on connaît un nombre β tel que

$$\inf_{\lambda_i \neq \lambda_k} |\mathcal{R}_A(\tilde{x}) - \lambda_i| \geq \beta \quad ,$$

on a <u>a fortiori</u>

$$|\mathcal{R}_A(\tilde{x}) - \lambda_k| \leq \frac{1}{\beta} \quad \frac{\|r\|_2^2}{\|\tilde{x}\|_2^2} \quad .$$

Le nombre β peut s'obtenir à partir des évaluations approchées des valeurs propres voisines et des bornes de KRYLOV et BOGOLIOUBOV.

Exercice 1 - On désire vérifier si une matrice symétrique A est définie positive. Donner quatre méthodes numériques, dont trois au moins soient des conditions nécessaires et suffisantes, pour arriver à ce résultat et discuter de leurs avantages respectifs.

Suggestion : triangularisation de Gauss, décomposition de Choleski, Méthode de Jacobi de recherche des valeurs propres, cercles de Gershgorin.

Exercice 2 [12] - Démontrer que la méthode de JACOBI conserve sa convergence si les angles de rotation θ_k sont choisis comme suit dans $(-\frac{\pi}{4}, \frac{\pi}{4})$:

$$\operatorname{tg}\theta_k = \begin{cases} 1 \quad \text{si } |a_{pq}^{(k)}| > |a_{qq}^{(k)} - a_{pp}^{(k)}| \\ \dfrac{a_{pq}^{(k)}}{a_{qq}^{(k)} - a_{pp}^{(k)}} \quad \text{si } |a_{pq}^{(k)}| \leq |a_{qq}^{(k)} - a_{pp}^{(k)}| \end{cases}$$

Solution - On a alors

$$a_{pq}^{(k+1)} = \tfrac{1}{2}(a_{pp}^{(k)} - a_{qq}^{(k)}) \sin 2\theta_k + a_{pq}^{(k)} \cos 2\theta_k .$$

Dans le cas où $\operatorname{tg}\theta_k = 1$, $\theta_k = \frac{\pi}{4}$ et

$$a_{pq}^{(k+1)} = \tfrac{1}{2}(a_{pp}^{(k)} - a_{qq}^{(k)}),$$

d'où

$$|a_{pq}^{(k+1)}|^2 \leq \frac{1}{4} |a_{pq}^{(k)}|^2$$

et

$$\|E^{(k+1)}\|_E^2 = \|E^{(k)}\|_E^2 - 2\, a_{pq}^{(k)\,2} + 2\, a_{pq}^{(k+1)\,2} \leq \|E^{(k)}\|_E^2 - \frac{3}{2} |a_{pq}^{(k)}|^2$$

$$(1 - \frac{3/2}{n(n-1)}) \|E^{(k)}\|_E^2 .$$

Dans le cas où $\operatorname{tg}\theta_k \neq 1$,

$$a_{pq}^{(k+1)} = \cos 2\theta_k (a_{pq}^{(k)} + \tfrac{1}{2}(a_{pp}^{(k)} - a_{qq}^{(k)}) \operatorname{tg} 2\theta_k)$$

$$= \frac{\cos 2\theta_k \left[a_{pq}^{(k)} (1 - \operatorname{tg}^2\theta_k) + (a_{pp}^{(k)} - a_{qq}^{(k)}) \operatorname{tg}\theta_k \right]}{1 - \operatorname{tg}^2\theta_k}$$

$$= \frac{\cos 2\theta_k}{1 - \operatorname{tg}^2\theta_k} \left[a_{pq}^{(k)} (1 - \operatorname{tg}^2\theta_k) - a_{pq}^{(k)} \right]$$

$$= -\frac{\cos 2\theta_k}{1 - \operatorname{tg}^2\theta_k} \operatorname{tg}^2\theta_k \, a_{pq}^{(k)}$$

$$= - \frac{\cos^2\theta_k - \sin^2\theta_k}{1 - tg^2\theta_k} \; tg^2\theta_k \; a_{pq}^{(k)} = - \cos^2\theta_k \; tg^2\theta_k \; a_{pq}^{(k)} = - a_{pq}^{(k)} \sin^2\theta_k \; ,$$

donc

$$|a_{pq}^{(k+1)}|^2 \leq |a_{pq}^{(k)}|^2 \sin^4 \theta_k$$

et

$$\|E^{(k+1)}\|_E^2 \leq \|E^{(k)}\|_E^2 - 2 \, |a_{pq}^{(k)}|^2 (1 - \sin^4\theta_k) \leq \|E^{(k)}\|_E^2 - \frac{3}{2} |a_{pq}^{(k)}|^2$$

$$(1 - \frac{3/2}{n(n-1)}) \; \|E^{(k)}\|_E^2 \; .$$

* Remarque - La convergence quadratique est maintenue, car pour $\theta_k \to 0$, $tg\, 2\theta_k \longrightarrow \frac{1}{2} tg\, \theta_k$, et les formules deviennent exactes. Le résidu laissé à la place de 0 est en fait $0(\|E^{(k)}\|_E^6)$ lorsque l'angle est $0(\|E^{(k)}\|_E)$.

Exercice 3 - Soient K et M deux matrices symétriques définie positives. On considère le problème aux valeurs propres

$$K x = \lambda M x \; .$$

Montrer que

a) Deux vecteurs propres relatifs à des valeurs propres différentes vérifient

$$z_{(1)}^T K z_{(2)} = 0 \quad , \quad z_{(1)}^T M z_{(2)} = 0 \; .$$

b) Les valeurs propres sont toutes positives.

* c) Il existe exactement n vecteurs propres indépendants $z_{(i)}$ vérifiant les relations

$$\lambda_i = \frac{z_{(i)}^T K z_{(i)}}{z_{(i)}^T M z_{(i)}} = \inf_{P_i} \frac{x^T K x}{x^T M x} \; ,$$

avec

$$P_i = \{ x \in R^n \mid x^T M z_{(1)} = 0, \ldots, x^T M z_{(i-1)} = 0 \} \; .$$

(démonstration directe!)

d) On peut appliquer la méthode de la puissance à ce problème si les valeurs propres sont séparées.

Solution a) On a en effet

$$K z_{(1)} = \lambda_1 M z_{(1)} \quad , \quad K z_{(2)} = \lambda_2 M z_{(2)} \; ,$$

d'où

$$z_{(2)}^T K z_{(1)} = \lambda_1 z_{(2)}^T M z_{(1)} = \lambda_2 z_{(1)}^T M z_{(2)} ,$$

ce qui entraîne

$$(\lambda_1 - \lambda_2) z_{(2)}^T M z_{(1)} = 0$$

et

$$z_{(2)}^T M z_{(1)} = 0 .$$

Il en découle

$$z_{(2)}^T K z_{(1)} = \lambda_1 z_{(2)}^T M z_{(1)} = 0 .$$

b) En effet, si $K z_{(i)} = \lambda_i M z_{(i)}$, on a aussi

$$z_{(i)}^T K z_{(i)} = \lambda_i z_{(i)}^T M z_{(i)} ,$$

d'où

$$\lambda_i = \frac{z_{(i)}^T K z_{(i)}}{z_{(i)}^T M z_{(i)}} > 0 ,$$

puisque K et M sont définies positives.

* c) Considérons d'abord le problème de la minimisation du quotient de RAYLEIGH

$$\mathcal{R}(x) = \frac{x^T K x}{x^T M x}$$

dans $R^n - \{ 0 \}$. On peut évidemment supposer $x^T M x = 1$. La surface définie par cette condition est en fait la sphère relative à la norme

$$\|x\|_M = (x^T M x)^{\frac{1}{2}} ,$$

liée au produit scalaire

$$(x, y)_M = x^T M y .$$

La fonction $\mathcal{R}(x)$ étant continue sur cette sphère, elle y admet un minimum, puisque la sphère est compacte. Soit $z_{(1)}$ le point de cette sphère réalisant le minimum, dont la valeur sera notée λ_1. On a donc

$$z_{(1)}^T K z_{(1)} = \lambda_1 z_{(1)}^T M z_{(1)} .$$

Montrons que $z_{(1)}$ est vecteur propre du problème pour la valeur propre λ_1. Pour y tel que $y^T M z_{(1)} = 0$, avec $\|y\|_M = 1$, et pour ε suffisamment petit, on a

$$\frac{(z_{(1)}^T + \varepsilon\, y^T)\, K\, (z_{(1)} + \varepsilon\, y)}{(z_{(1)}^T + \varepsilon\, y^T)\, M\, (z_{(1)} + \varepsilon\, y)} \geqslant \lambda_1 \quad ,$$

soit

$$z_{(1)}^T K z_{(1)} + 2\,\varepsilon\, y^T K z_{(1)} + \varepsilon^2\, y^T K y \geqslant \lambda_1 (z_{(1)}^T M z_{(1)} + 2\,\varepsilon\, y^T M z_{(1)} + \varepsilon^2\, y^T M y) \quad ,$$

ce qui entraîne

$$2\,\varepsilon\, y^T (K z_{(1)} - \lambda_1 M z_{(1)}) \geqslant \lambda_1\, \varepsilon^2 (y^T M y - y^T K y) \quad .$$

Or, cette relation, vraie pour tout ε suffisamment petit, implique

$$y^T (K z_{(1)} - \lambda_1 M z_{(1)}) = 0 \quad .$$

En effet, on a toujours

$$|y^T M y| = 1 \quad ; \quad |y^T K y| \leqslant \|y\|_M \, \|K\|_M \, \|y\|_M = \|K\|_M \quad ,$$

avec

$$\|K\|_M = \sup_{x \neq 0} \frac{\|Kx\|_M}{\|x\|_M} \quad ,$$

d'où

$$\frac{2}{\varepsilon}\, y^T (K z_{(1)} - \lambda_1 M z_{(1)}) \geqslant \lambda_1 (1 - \|K\|_M) \quad .$$

Dès lors, si

$$y^T (K z_{(1)} - \lambda_1 M z_{(1)}) \neq 0 \quad ,$$

choisissons ε de signe contraire et tel que

$$\begin{cases} |\varepsilon| < \dfrac{|y^T (K z_{(1)} - \lambda_1 M z_{(1)})|}{|\lambda_1 (1 - \|K\|_M)|} & \text{si } \|K\|_M \neq 1 \\ |\varepsilon| < 1 \quad \text{si } \|K\|_M = 1 \quad . \end{cases}$$

Il vient alors

$$\begin{cases} -2\lambda_1 |1 - \|K\|_M| > \lambda_1 (1 - \|K\|_M) & \text{si } \|K\|_M \neq 1 \\ -2|y^T (K z_{(1)} - \lambda_1 M z_{(1)})| > 0 & \text{si } \|K_M\| = 1 \quad , \end{cases}$$

soit une absurdité dans chaque cas. Par conséquent, pour

$$x = \alpha\, z_{(1)} \quad , \quad z_{(1)}^T (K z_{(1)} - \lambda_1 M z_{(1)}) = 0$$

et pour x orthogonal à $z_{(1)}$, on a encore

$$x^T(K\, z_{(1)} - \lambda_1\, M\, z_{(1)}) = 0 \ .$$

On en déduit

$$K\, z_{(1)} = \lambda_1\, M\, z_{(1)} \ ,$$

c'est-à-dire que $z_{(1)}$ est bien un vecteur propre de valeur propre λ_1.

On recommence le même processus dans le sous-espace P_1, ,ce qui donne un vecteur propre $z_{(2)}$ et une valeur propre λ_2 , etc ... On obtient ainsi n vecteurs propres et n valeurs propres.

d) Développant

$$x = \sum_{i=1}^{n} \alpha_i\, z_{(i)} \quad ,$$

on a évidemment

$$M^{-1}\, K\, x = \sum_{i=1}^{n} M^{-1} \alpha_i\, \lambda_i\, M\, z_{(i)} = \sum_{i=1}^{n} \alpha_i\, \lambda_i\, z_{(i)}$$

.....

$$(M^{-1}\, K)^p\, x = \sum_{i=1}^{n} \alpha_i\, \lambda_i^p\, z_{(i)} \quad ,$$

ce qui conduit à trouver le premier vecteur propre, pour autant que $\lambda_1 > \lambda_2$. En itérant sur $K^{-1}M$, ontrouverait la plus petite valeur propre. C'est la procédure utilisée dans les calculs de vibrations des structures.

Exercice 4 - Pour le problème ci-dessus, montrer que l'on peut également itérer sur une matrice symétrique.

Solution - Faisant $M = LL^T$, on a

$$x^{(k+1)} = K^{-1}M\, x^{(k)} = K^{-1}LL^T x^{(k)} \ ,$$

d'où

$$L^T x^{(k+1)} = \underbrace{L^T K^{-1} L}_{\text{sym.}}\ L^T x^{(k)} \ .$$

A une transformation triangulaire près, c'est la même itération. On trouvera finalement $L^T z_{(1)}$, et il suffira d'effectuer la transformation inverse.

Exercice 5 - On considère la matrice

$$A = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 7 & 8 & 9 & 10 \\ 12 & 13 & 14 & 15 \\ 17 & 18 & 19 & 20 \end{bmatrix}$$

On demande d'évaluer dans quelle zone de l'espace complexe il est possible de trouver des valeurs propres de A (faire un dessin).

Solution: Par application directe des cercles de GERSHGORIN, tout point de $\mathbb{C}$ vérifiant une des inégalités

$$|\lambda - 2| \leq 12$$

ou

$$|\lambda - 8| \leq 26$$

ou

$$|\lambda - 14| \leq 40$$

ou

$$|\lambda - 20| \leq 54$$

pourrait être valeur propre. (Voir fig. 4)

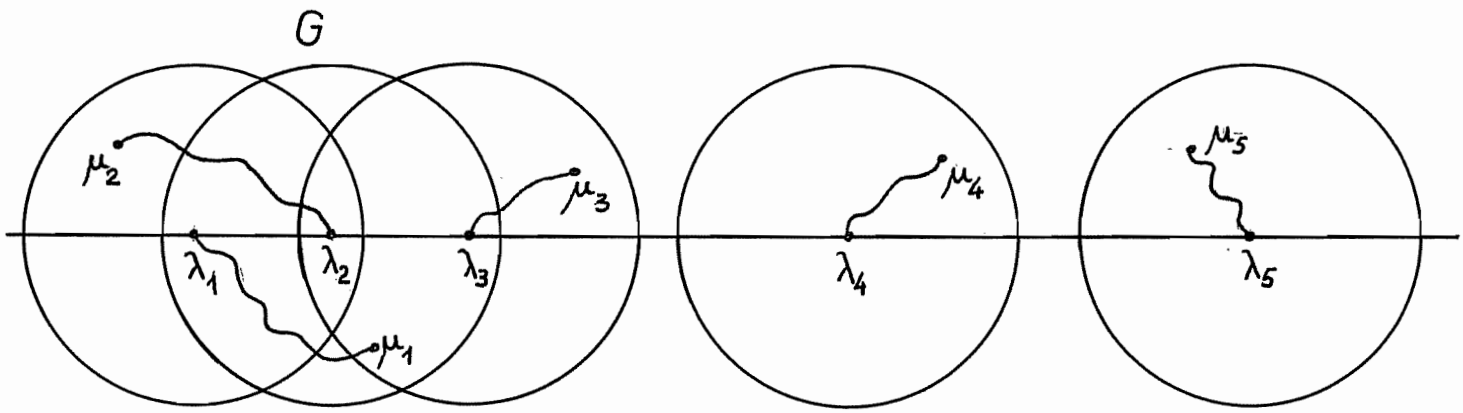

G
μ_2
μ_3
μ_4
μ_5
μ_1
λ_1
λ_2
λ_3
λ_4
λ_5

Fig. 1

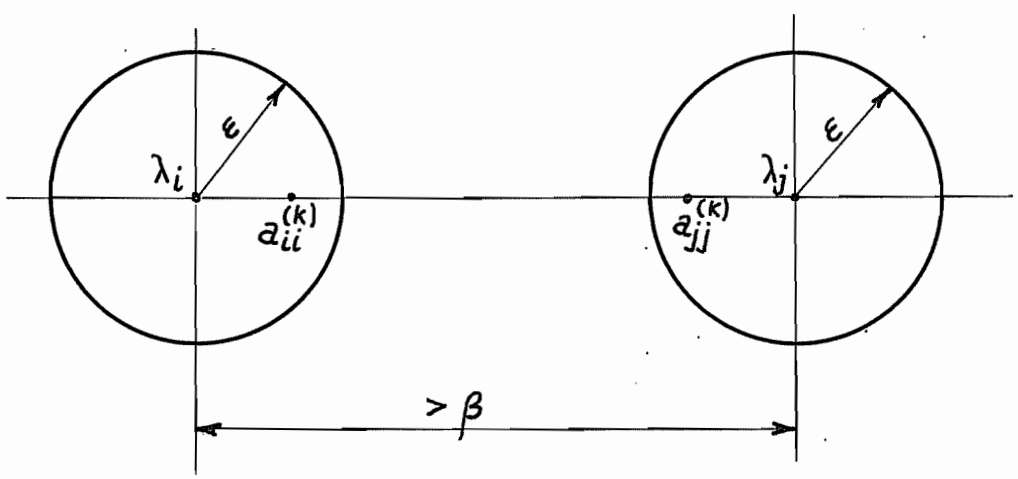

λ_i
ε
$a_{ii}^{(k)}$
λ_j
ε
$a_{jj}^{(k)}$
$> \beta$

Fig. 2

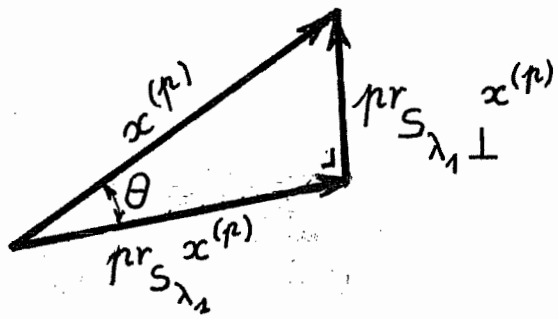

$x^{(p)}$
θ
$pr_{S_{\lambda_1}} x^{(p)}$
$pr_{S_{\lambda_1}\perp} x^{(p)}$

Fig. 3

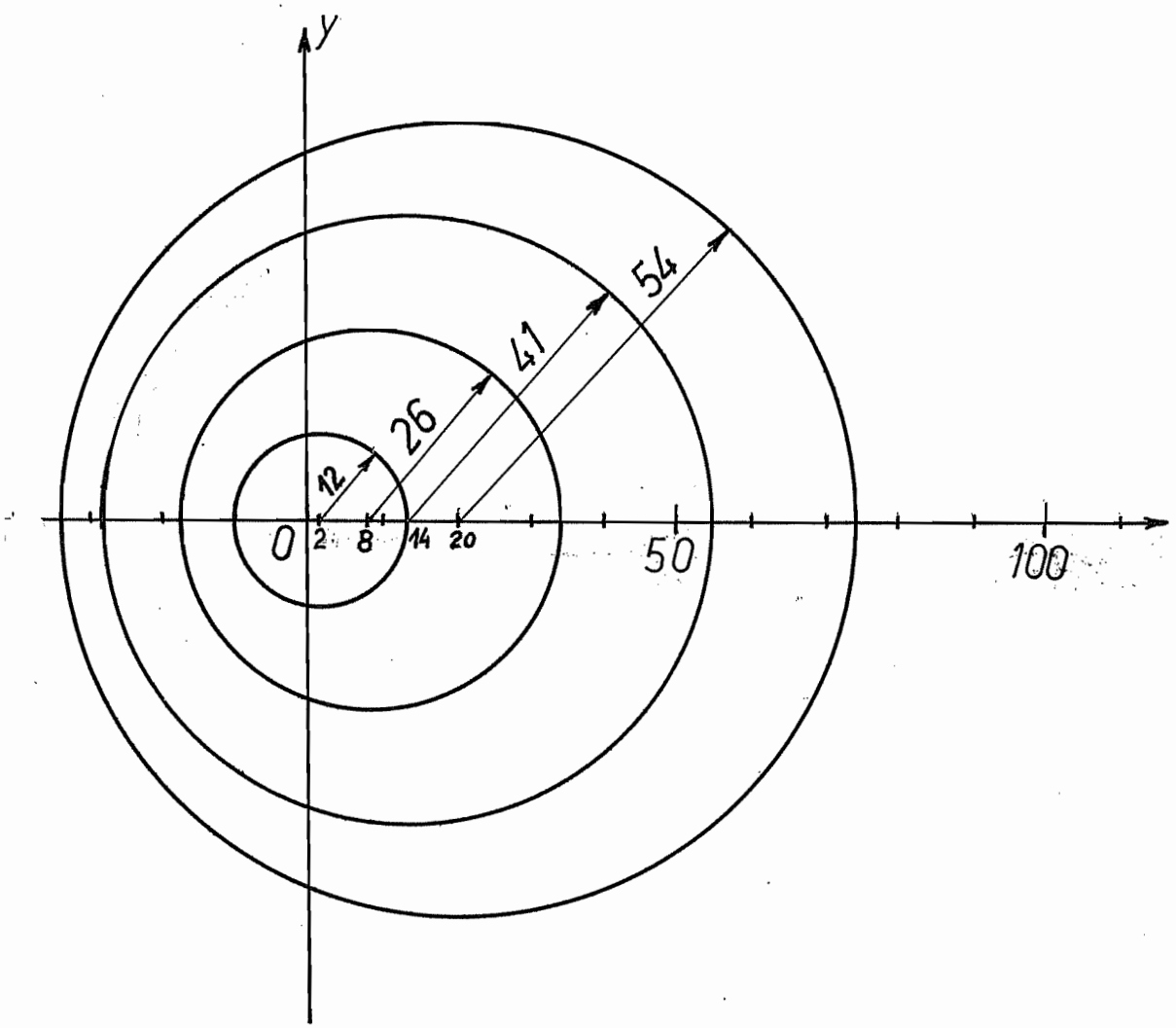

y
0
2
8
14
20
50
100
12
26
41
54

Fig. 4

P R O B L E M E S D I F F E R E N T I E L S

A V A L E U R S I N I T I A L E S

1. GENERALITES - THEOREME DE PICARD

1.1 - Le problème est de chercher une fonction y vérifiant l'équation différentielle

$$y'(t) = f(t, y(t)) ,$$

avec la condition initiale

$$y(t_o) = y_o .$$

1.2 - Une telle fonction existe-t-elle? Est-elle unique? PEANO a montré que, moyennant l'hypothèse de continuité de f(t, y), on peut garantir l'existence d'une solution au moins, mais non l'unicité [21] . Cette dernière nécessite des conditions un peu plus fortes. Le résultat le plus connu à ce sujet est dû à PICARD, qui a répondu à cette question par une voie constructive:

Théorème de PICARD - Si la fonction f(t, y) est à la fois

-continue par rapport à t

- uniformément lipschitzienne par rapport à y, c'est-à-dire qu'il existe une constante $L > 0$, indépendante de t, telle que

$$|f(t, y) - f(t, z)| \leq L\,|y - z| .$$

Alors, le problème

$$y' = f(t, y(t)) \quad , \quad y(t_o) = y_o$$

admet une solution unique pour $t \geq t_o$.

Considérons en effet l'intervalle $[\, t_o, t_1 \,)$, avec

$$t_1 = t_o + h , \qquad h < 1/L .$$

Le problème revient, sur cet intervalle, à chercher la fonction y(t) qui vérifie

$$y(t) = \int_{t_o}^{t} f(\tau , y(\tau)) \, d\tau .$$

On peut procéder par approximations successives, en partant d'une fonction y_1 arbitraire vérifiant $y_1(t_o) = y_o$ et en définissant la récurrence

$$y_{n+1}(t) = \int_{t_o}^{t} f(\tau , y_n(\tau)) \, d\tau .$$

Dans l'espace $C^o([t_o, t_1])$, muni de la norme

$$\|y\| = \sup_{[t_o, t_1]} |y(t)| ,$$

la relation de récurrence

$$y_{n+1}(t) = F_t(y_n)$$

définie par la relation ci-dessus est une <u>contraction</u>, car

$$|F_t(y) - F_t(z)| \leq \int_{t_o}^{t} |f(\tau, y(\tau)) - f(\tau, z(\tau))| \, d\tau$$

$$\leq L \sup_{[t_o, t_1]} |y(\tau) - z(\tau)| \cdot (t_1 - t_o)$$

$$\leq L\, h\, \|y - z\| .$$

Ceci étant vrai pour tout t dans l'intervalle, on a également

$$\|F_t(y) - F_t(z)\| \leq Lh \, \|y - z\| ,$$

avec, par construction, $Lh < 1$, ce qui implique l'existence et l'unicité de la limite, qui est visiblement la solution du problème.

Connaissant la solution en t_1, on peut recommencer le processus dans l'intervalle $[t_1, t_1 + h]$, etc..., ce qui mène finalement à une solution unique sur un intervalle aussi grand que l'on veut.

1.3 - Il existe des méthodes analytiques d'approximations successives fondées sur le théorème de PICARD. Voyons par exemple [22] comment s'intègre l'équation

$$y' = y \quad , \text{ avec } \quad y_o = 1 .$$

En prenant comme point de départ $y_1(t) = 1$, on obtient successivement

$$y_2(t) = 1 + \int_0^t d\tau = 1 + t$$

$$y_3(t) = 1 + \int_0^t (1 + \tau) \, d\tau = 1 + t + \frac{t^2}{2}$$

$$y_4(t) = 1 + \int_0^t (1 + \tau + \frac{\tau^2}{2}) \, d\tau = 1 + t + \frac{t^2}{2} + \frac{t^3}{6}$$

.....

On retrouve progressivement le développement en série de TAYLOR de l'exponentielle.

2. METHODES PAS A PAS

Les méthodes pas à pas consistent à chercher la solution sous forme d'une table numérique. L'intervalle d'intégration est subdivisé par des points t_i (fig. 1). Un pas consiste à passer de l'approximation $\tilde{y}_i \approx y(t_i)$ à l'approximation $\tilde{y}_{i+1} \approx y(t_{i+1})$. Il existe deux grandes espèces de méthodes pas à pas:

- Les méthodes à pas indépendants, où $\tilde{y}_{i+1}$ ne dépend que de $\tilde{y}_i$ et de la fonction f ;

- Les méthodes à pas liés, où $\tilde{y}_{i+1}$ dépend de la fonction f et d'un certain nombre de valeurs précédentes $\tilde{y}_j$, $j < i$.

3. INTRODUCTION AUX METHODES DE RUNGE-KUTTA

Les méthodes de RUNGE-KUTTA consistent à remplacer le problème exact

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt = F(y(t_n))$$

par une expression approchée

$$\tilde{y}_{n+1} = \tilde{F}(y_n) .$$

Pour un même point de départ y_n , la différence

$$F(y_n) - \tilde{F}(y_n)$$

est appelée erreur par pas. Si l'intervalle $[\, t_n, t_{n+1} \,]$ a une longueur h, et si

$$|F(y_n) - \tilde{F}(y_n)| = O(h^{p+1}),$$

on dit que l'algorithme est d'ordre p.

Les algorithmes les plus simples sont très faciles à construire. A l'ordre 1,

$$y_{n+1} - y_n = \int_{t_n}^{t_{n+1}} f(t,y(t))dt = \int_{t_n}^{t_{n+1}} [\, f(t_n, y_n) + O(h) \,]\, dt$$

$$= h\, f(t_n, y_n) + O(h^2) ,$$

ce qui mène à la méthode d'EULER

$$\tilde{y}_{n+1} = \tilde{y}_n + h\, f(t_n, \tilde{y}_n)$$

qui est donc du premier ordre. On peut obtenir tout aussi simplement

une méthode du second ordre en remarquant que

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(t, y)\, dt$$

$$= y_n + h\, f(t_n + \frac{h}{2},\ y(t_n + \frac{h}{2})) + O(h^3)$$

(formule d'intégration des rectangles). Il est vrai que l'on ne connaît pas $y(t_n + \frac{h}{2})$. Mais on a de toute évidence

$$y(t_n + \frac{h}{2}) = y_n + \frac{h}{2} f(t_n, y_n) + O(h^2) ,$$

ce qui, si l'on pose

$$k_1 = h\, f(t_n, y_n) ,$$

donne

$$y(t_n + \frac{h}{2}) = y_n + \frac{k_1}{2} + O(h^2) .$$

Il en découle, puisque f est lipschitzienne en y,

$$y_{n+1} = y_n + h\, f(t_n + \frac{h}{2} , y_n + \frac{k_1}{2}) + O(h^3) .$$

C'est le principe de <u>l'algorithme d'EULER modifié</u>, d'ordre 2, que l'on présente généralement comme suit:

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \frac{1}{2}h, \tilde{y}_n + \frac{1}{2} k_1) \\ \tilde{y}_{n+1} = \tilde{y}_n + k_2 . \end{cases}$$

4. ALGORITHMES GENERAUX DE RUNGE-KUTTA

Des approximations d'ordre plus élevé peuvent être obtenues à partir de l'algorithme général suivant:

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \alpha_2 h , \tilde{y}_n + \beta_{21} k_1) \\ k_3 = h\, f(t_n + \alpha_3 h , \tilde{y}_n + \beta_{31} k_1 + \beta_{32} k_2) \\ \cdots\cdots\cdots \\ k_q = h\, f(t_n + \alpha_q h , \tilde{y}_n + \beta_{q1} k_1 + \cdots + \beta_{q,q-1} k_{q-1}) \\ \tilde{y}_{n+1} = \tilde{y}_n + \mu_1 k_1 + \cdots + \mu_q k_q \end{cases}$$

Comme il y a q calculs $k_1, \ldots, k_q$, on dit qu'il s'agit d'une formule à q approximations. On choisira les coefficients $\alpha_2, \ldots, \alpha_q$, $\beta_{21}, \ldots, \beta_{q,q-1}$ et $\mu_1, \ldots, \mu_q$ pour obtenir un ordre p donné. une formule d'ordre p à q approximations se note RK_{pq}. Ainsi, la méthode d'EULER modifiée est un RK_{22}. La détermination de l'ordre se fait à partir des développements de TAYLOR de la solution et de son approximation.

4.1 - Formules du second ordre

Les calculs ne se font sans difficulté que si l'on adopte des notations suffisamment concises. Nous écrirons

$$f_{ij} = \left(\frac{\partial^{i+j} f}{\partial t^i \, \partial y^j} \right)_{t=t_n,\ y=y_n}$$

a) On a successivement

$$y'(t_n) = f(t_n, y_n) = f_{00}$$

$$y''(t_n) = f_{10} + f_{01}\, y'(t_n) = f_{10} + f_{01}\, f_{00}$$

et

$$y = y'(t_n)\Delta t + y''(t_n)\frac{\Delta t^2}{2} + 0(\Delta t^3)$$

$$= f_{00}\Delta t + (f_{10} + f_{01}\, f_{00})\frac{\Delta t^2}{2} + 0(\Delta t^3)\ . \qquad (1)$$

C'est la formule de TAYLOR limitée au second ordre.

b) Voyons à présent l'algorithme de RUNGE-KUTTA. Il s'écrit

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \alpha_2 h, \tilde{y}_n + \beta_{21} k_1) \\ \tilde{y}_{n+1} = \tilde{y}_n + \mu_1 k_1 + \mu_2 k_2 \end{cases}$$

Dans le calcul de l'erreur par pas, il faut supposer $\tilde{y}_n = y_n$. On développe alors k_1 et k_2:

$$k_1 = h\, f_{00}$$

$$k_2 = h(f_{00} + \alpha_2 h\, f_{10} + \beta_{21} k_1\, f_{01})$$

$$= h\, f_{00} + h^2(\alpha_2 f_{10} + \beta_{21}\, f_{01}\, f_{00})$$

et

$$y = \mu_1 k_1 + \mu_2 k_2 = h(\mu_1 + \mu_2) f_{00} + h^2(\mu_2 \alpha_2 f_{10} + \mu_2 \beta_{21} f_{01} f_{00})\ . \qquad (2)$$

On compare les développements (1) et (2) et on relève les conditions

pour qu'ils soient égaux jusqu'à l'ordre deux::

- Termes en h : $(\mu_1 + \mu_2) = 1$
- Termes en h^2 : $\begin{cases} \text{en } f_{10} : \frac{1}{2} = \mu_2 \alpha_2 \\ \text{en } f_{01} f_{00} : \frac{1}{2} = \mu_2 \beta_{21} \end{cases}$.

On obtient donc trois équations liant les quatres variables α_2, β_{21}, μ_1, μ_2. Il existe par conséquent une infinité simple d'algorithmes RK_{22}. Voici les plus courants:

- Algorithme d'EULER-CAUCHY: $\alpha_2 = 1$

  Il en découle $\mu_2 = \frac{1}{2}$, $\mu_1 = \frac{1}{2}$, $\beta_{21} = 1$, ce qui donne

$$\begin{cases} k_1 = h\, f(t_n, y_n) \\ k_2 = h\, f(t_n + h, \tilde{y}_n + k_1) \\ \tilde{y}_{n+1} = \tilde{y}_n + \frac{1}{2}(k_1 + k_2) \end{cases} .$$

- Algorithme d'EULER modifié: $\alpha_2 = \frac{1}{2}$

  Il en découle $\mu_2 = 1$, $\mu_1 = 0$, $\beta_{21} = \frac{1}{2}$, ce qui donne

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \frac{1}{2}h, \tilde{y}_n + \frac{1}{2} k_1) \\ \tilde{y}_{n+1} = \tilde{y}_n + k_2 \end{cases}$$

- Algorithme de HEUN: $\alpha_2 = 2/3$

  Il en découle $\mu_2 = 3/4$, $\mu_1 = 1/4$, $\beta_{21} = 2/3$, ce qui donne

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \frac{2}{3} h\ , \tilde{y}_n + \frac{2}{3} k_1) \\ \tilde{y}_{n+1} = \tilde{y}_n + \frac{1}{4} k_1 + \frac{3}{4} k_2 \end{cases}$$

4.2 - Formules du troisième ordre

On calcule d'abord

$$y'''(t_n) = f_{20} + f_{11} y'(t_n) + f_{02} y'^2(t_n) + f_{01}\, y''(t_n)$$

$$= f_{20} + 2 f_{11} f_{00} + f_{02} f_{00}^2 + f_{01}(f_{10} + f_{01} f_{00}) \quad ,$$

d'où

$$\Delta y = f_{00}\,\Delta t + (f_{10} + f_{01}f_{00})\,\frac{\Delta t^2}{2} + (f_{20} + 2\,f_{11}f_{00} + f_{02}f_{00}^2 + f_{01}(f_{10} + f_{01}f_{00}))\,\frac{\Delta t^3}{6} + 0(\Delta t^4)\ .$$

Quant à l'algorithme général de RUNGE-KUTTA à trois approximations, il s'écrit

$$\begin{cases} k_1 = h\,f(t_n, \tilde{y}_n) \\ k_2 = h\,f(t_n + \alpha_2\,h, \tilde{y}_n + \beta_{21}k_1) \\ k_3 = h\,f(t_n + \alpha_3\,h, \tilde{y}_n + \beta_{31}k_1 + \beta_{32}k_2) \\ \tilde{y}_{n+1} = \tilde{y}_n + \mu_1\,k_1 + \mu_2\,k_2 + \mu_3\,k_3 \end{cases}$$

Calculons les développements de TAYLOR de k_1, k_2, k_3. On a d'abord

$$k_1 = h\,f_{00}\ .$$

Ensuite,

$$k_2 = h(f_{00} + h\,\alpha_2 f_{10} + \beta_{21}k_1\,f_{01} + \tfrac{1}{2}\,h^2\alpha_2^2\,f_{20} + h\,\alpha_2\,\beta_{21}k_1\,f_{11} + \tfrac{1}{2}\,\beta_{21}^2\,k_1^2\,f_{02}) + 0(h^4)\ ,$$

soit

$$k_2 = h\,f_{00} + h^2(\alpha_2\,f_{10} + \beta_{21}\,f_{00}\,f_{01}) + h^3(\tfrac{1}{2}\alpha_2^2\,f_{20} + \alpha_2\,\beta_{21}\,f_{00}\,f_{11} + \tfrac{1}{2}\beta_{21}^2\,f_{00}^2\,f_{02}) + 0(h^4).$$

Enfin,

$$k_3 = h(f_{00} + h\,\alpha_3 f_{10} + (\beta_{31}k_1 + \beta_{32}k_2)\,f_{01} + \tfrac{1}{2}\,h^2\alpha_3^2\,f_{20} + h\,\alpha_3(\beta_{31}k_1 + \beta_{32}k_2)\,f_{11} + \tfrac{1}{2}(\beta_{31}k_1 + \beta_{32}k_2)^2\,f_{02} + 0(h^4).$$

En y introduisant les expressions de k_1 et k_2, tout en se limitant à l'ordre 3, on obtient

$$k_3 = h\, f_{00} + h^2(\alpha_3 f_{10} + (\beta_{31} + \beta_{32}) f_{01} f_{00})$$

$$+ h^3 [\alpha_2 \beta_{32} f_{10} f_{01} + \beta_{32} \beta_{21} f_{00} f_{01}^2 + \tfrac{1}{2} \alpha_3^2 f_{20}$$

$$+ \alpha_3(\beta_{31} + \beta_{32}) f_{11} f_{00} + \tfrac{1}{2}(\beta_{31} + \beta_{32})^2 f_{00}^2 f_{02}]$$

$$+ O(h^4) .$$

Les conditions d'identité des deux développements jusqu'à l'ordre 3 sont:

Ordre 1: $(\mu_1 + \mu_2 + \mu_3) = 1$ (1)

Ordre 2:

- Terme en f_{10}: $\frac{1}{2} = \mu_2 \alpha_2 + \mu_3 \alpha_3$ (2)
- Terme en $f_{01}\, f_{00}$: $\frac{1}{2} = \mu_2 \beta_{21} + \mu_3(\beta_{31} + \beta_{32})$ (3)

Ordre 3:

- Terme en f_{20}: $\frac{1}{6} = \frac{1}{2} \mu_2 \alpha_2^2 + \frac{1}{2} \mu_3 \alpha_3^2$ (4)
- Terme en $f_{11}\, f_{00}$: $\frac{1}{3} = \mu_2 \alpha_2 \beta_{21} + \mu_3 \alpha_3(\beta_{31} + \beta_{32})$ (5)
- Terme en $f_{02}\, f_{00}^2$: $\frac{1}{6} = \frac{1}{2} \mu_2 \beta_{21}^2 + \frac{1}{2} \mu_3(\beta_{31} + \beta_{32})^2$ (6)
- Terme en $f_{01}\, f_{10}$: $\frac{1}{6} = \frac{1}{2} \mu_3 \alpha_2 \beta_{32}$ (7)
- Terme en $f_{00}\, f_{01}^2$: $\frac{1}{6} = \mu_3 \beta_{32} \beta_{21}$ (8)

Les variables sont au nombre de huit: $\alpha_2, \alpha_3, \beta_{21}, \beta_{31}, \beta_{32}, \mu_1, \mu_2, \mu_3$, ce qui pourrait faire croire que la solution est unique. Il n'en est rien, comme nous allons le montrer. Ajoutons l'équation (6) à l'équation (4) et soustrayons l'équation (5) au résultat: on obtient

$$0 = \tfrac{1}{2} \mu_2 (\alpha_2 - \beta_{21})^2 + \tfrac{1}{2} \mu_3 (\alpha_3 - (\beta_{31} + \beta_{32}))^2 .$$

Dès lors, si l'on veut obtenir μ_2 et μ_3 positifs, il faudra que

$$\alpha_2 = \beta_{21} \quad (4')$$

$$\alpha_3 = (\beta_{31} + \beta_{32}) \quad (5')$$

Dès lors, les équations (2) et (3) deviennent équivalentes, de même que (7) et (8). On obtient en définitive le système simplifié

$$\begin{cases} \mu_1 + \mu_2 + \mu_3 = 1 & (1") \\ \mu_2 \alpha_2 + \mu_3 \alpha_3 = \frac{1}{2} & (2") \\ \mu_2 \alpha_2^2 + \mu_3 \alpha_3^2 = \frac{1}{3} & (3") \\ \beta_{21} = \alpha_2 & (4") \\ \beta_{31} + \beta_{32} = \alpha & (5") \\ \mu_3 \alpha_2 \beta_{32} = 1/6 & (6") \end{cases}$$

/3

qui admet une double infinité de solutions.

L'algorithme RK_{33} le plus classique correspond au choix suivant:

$\alpha_2 = 1/2$, $\alpha_3 = 1$.

On a alors

$\mu_2 = 2/3$, $\mu_3 = 1/6$, $\mu_1 = 1/6$

et

$\beta_{21} = 1/2$, $\beta_{32} = 2$, $_{31} = -1$,

ce qui donne le RK_{33} classique:

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \frac{1}{2}h\ ,\ \tilde{y}_n + \frac{1}{2} k_1) \\ k_3 = h\, f(t_n + h,\ \tilde{y}_n - k_1 + 2\, k_2) \\ \tilde{y}_{n+1} = \tilde{y}_n + \frac{1}{6}(k_1 + 4\, k_2 + k_3) \end{cases}$$

On remarquera [8] que si $\frac{\partial f}{\partial y} = 0$, cette formule équivaut à celle de SIMPSON, dont on sait qu'elle mène à une erreur $0(h^5)$. Le RK_{33} classique convient donc particulièrement pour les problèmes où la dépendance de f par rapport à y est faible.

4.3 - Formules du quatrième ordre

Nous ne développerons pas ici la discussion générale des formules à quatre approximations, qui est fort lourde. Deux algorithmes RK_{44} sont fort utilisés: ce sont

- L'algorithme de RUNGE, encore appelé RK_{44} classique:

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \frac{1}{2}h, \tilde{y}_n + \frac{1}{2} k_1) \\ k_3 = h\, f(t_n + \frac{1}{2}h, \tilde{y}_n + \frac{1}{2} k_2) \\ k_4 = h\, f(t_n + h, \tilde{y}_n + k_3) \\ \tilde{y}_{n+1} = \tilde{y}_n + \frac{1}{6}(k_1 + 2\, k_2 + 2\, k_3 + k_4) \end{cases}$$

Comme le RK_{33} classique, cet algorithme dégénère, si $\frac{\partial f}{\partial y} = 0$, en la formule de SIMPSON. C'est, de loin, le plus utilisé.

- L'algorithme de KUTTA

$$\begin{cases} k_1 = h\, f(t_n, \tilde{y}_n) \\ k_2 = h\, f(t_n + \frac{h}{3} , \tilde{y}_n + \frac{1}{3} k_1) \\ k_3 = h\, f(t_n + \frac{2}{3} h , \tilde{y}_n - \frac{1}{3} k_1 + k_2) \\ k_4 = h\, f(t_n + h , \tilde{y}_n + k_1 - k_2 + k_3) \\ \tilde{y}_{n+1} = \tilde{y}_n + \frac{1}{8} (k_1 + 3\, k_2 + 3\, k_3 + k_4) \end{cases}$$

Dans le cas où $\frac{\partial f}{\partial y} = 0$, cet algorithme dégénère en la formule des 3/8.

5. CONTROLE DE L'ERREUR PAR PAS

L'erreur par pas est une fonction de h: nous noterons donc

$$e(h) = y(t_n + h) - y(t_n) - \sum_{j=1}^{q} \mu_j\, k_j \ .$$

Si la solution est suffisamment régulière, on peut écrire

$$e(h) = e(0) + h\, e'(0) + \frac{h^2}{2} e''(0) + \ldots + \frac{h^{p+1}}{(p+1)!} e^{(p+1)}(0) + \frac{h^{p+2}}{(p+2)!} e^{(p+2)}(\theta h) \ .$$

Dans ce contexte, l'affirmation que l'algorithme est d'ordre p revient à dire que

$$e(0) = 0 \ , \quad e'(0) = 0 \ , \quad \ldots \ , \quad e^{(p)}(0) = 0,$$

soit

$$e(h) = \frac{h^{p+1}}{(p+1)!} e^{(p+1)}(0) + \frac{h^{p+2}}{((p+2)!} e^{(p+2)}(\theta h)$$

Le premier terme est la partie principale de l'erreur. On peut s'en faire une idée de la manière suivante (fig. 2) : après avoir fait deux pas, menant de t_n à $t_n + 2h$, on refait le même calcul en un seul pas de longueur 2h. Soient $\tilde{y}_{n+2}$ la valeur obtenue en deux pas et $\tilde{\tilde{y}}_{n+2}$ valeur obtenue en un pas. On suppose évidemment la valeur de départ exacte. Il vient alors

$$y_{n+2} - \tilde{y}_{n+2} = 2 \frac{h^{p+1}}{(p+1)!} e^{(p+1)}(0) + O(h^{p+2})$$

et

$$y_{n+2} - \tilde{\tilde{y}}_{n+2} = \frac{2^{p+1} h^{p+1}}{(p+1)!} e^{(p+1)}(0) + O(h^{p+2})$$

Soustrayons ces deux résultats: on obtient

$$\tilde{\tilde{y}}_{n+2} - \tilde{y}_{n+2} = 2(1 - 2^p) \frac{h^{p+1}}{(p+1)!} e^{(p)}(0) + O(h^{p+2})$$

soit

$$\frac{h^{p+1}}{(p+1)!} e^{(p+1)}(0) = \frac{y_{n+2} - y_{n+2}}{2(1 - 2^p)} + O(h^{p+2}) .$$

On en déduit

$$y_{n+2} = \tilde{y}_{n+2} + \frac{\tilde{y}_{n+2} - \tilde{\tilde{y}}_{n+2}}{2^p - 1} + O(h^{p+2}) .$$

Ceci fournit:

a) Une approximation à l'ordre (p+1) de y_{n+2}

b) Une évaluation de l'erreur par pas pour y_{n+2}:

$$\frac{1}{2} \frac{\tilde{y}_{n+2} - \tilde{\tilde{y}}_{n+2}}{2^p - 1}$$

Lorsque la solution varie brutalement, on observe souvent $\tilde{y}_{n+2} - \tilde{\tilde{y}}_{n+2}$ très grand. Il faut alors repartir en arrière, avec un pas plus petit (fig. 3)

6. ANALYSE DE L'ERREUR DES ALGORITHMES DE RUNGE-KUTTA

La solution exacte du problème différentiel vérifie

$$y_{n+1} = y_n + \int_0^h f(t, y(t))\, dt = F_h(y_n) .$$

La méthode de RUNGE-KUTTA revient à remplacer $F_h(y_n)$ par une expression approchée $\tilde{F}_h(\tilde{y}_n)$. L'erreur par pas correspond à la différence entre $F_h(y_n)$ et $\tilde{F}_h(y_n)$. L'algorithme ayant été construit pour être d'ordre p,

on aura

$$F_h(y_n) - \tilde{F}_h(y_n) = O(h^{p+1}) .$$

Nous admettrons l'existence d'une borne uniforme de l'erreur par pas, c'est-à-dire que, quel que soit n et quel que soit z_n, on a

$$| \tilde{F}_h(z_n) - F_h(z_n) | \leq M h^{p+1} .$$

En outre, à chaque pas, on commet des erreurs d'arrondi et d'évaluation de la fonction f, ce qui revient à dire qu'en lieu et place de F_h, on calcule une grandeur F_h^* légèrement différente. Ces erreurs sont très difficiles à chiffrer, car elles dépendent de la complication de la fonction f et du nombre d'approximations de l'algorithme. On peut cependant tenir le raisonnement suivant: si ε_o est l'erreur de calcul de $(\mu_1 k_1 + \ldots + \mu_q k_q)$, on calculera en fait, à la place de

$$\tilde{y}_{n+1} = \tilde{y}_n + h(\mu_1 k_1 + \ldots + \mu_q k_q) ,$$

le nombre

$$y^*_{n+1} = (\tilde{y}_n + h (\mu_1 k_1 + \ldots + \mu_q k_q + \varepsilon_o)(1 + \varepsilon_1))(1 + \varepsilon_2)$$

L'erreur a donc la forme

$$\varepsilon_2 \tilde{y}_{n+1} + h [\varepsilon_1(\mu_1 k_1 + \ldots + \mu_q k_q) + \varepsilon_o]$$

et son terme principal, pour h petit, est du type $\varepsilon_2 \tilde{y}_{n+1}$. Si l'on suppose les $\tilde{y}_n$ bornés, on peut donc condidérer que les erreurs d'arrondi et d'évaluation sont bornées:

$$| F_h^*(z_n) - \tilde{F}_h(z_n) | \leq C ,$$

avec C petit.

Etudions l'évolution de ces erreurs au cours du processus. La vraie solution vérifie

$$y_{n+1} = F_h(y_n) ,$$

tandis que la solution approchée vérifie

$$y^*_{n+1} = F_h^*(y^*_n) .$$

On a, en notant $e_n = y_n - y^*_n$,

$$e_{n+1} = y_{n+1} - y^*_{n+1} = F_h(y_n) - F_h^*(y^*_n)$$

$$= \underbrace{[F_h(y_n) - F_h(y^*_n)]}_{\text{I}} + \underbrace{[F_h(y^*_n) - \tilde{F}_h(y^*_n)]}_{\text{II}} + \underbrace{[\tilde{F}_h(y^*_n) - F_h^*(y^*_n)]}_{\text{III}}$$

Le groupe de termes III représente les erreurs d'évaluation et d'arrondi. Nous savons que

$$|F_h(y_n^*) - F_h^*(y_n^*)| \leq C .$$

Le groupe de termes II constitue l'erreur par pas, qui vérifie

$$|F_h(y_n^*) - F_h(y_n^*)| \leq M\, h^{p+1} .$$

Enfin, l'expression I constitue l'erreur propagée, c'est-à-dire l'erreur provenant des pas précédents, en supposant qu'au présent pas, l'intégration est exacte. Il s'agit de la différence entre les solutions des deux problèmes suivants:

$$y' = f(t, y) \quad , \; y(t_n) = y_n$$

et

$$z' = f(t, z) \quad , \; z(t_n) = y_n^* .$$

La soustraction de ces deux équations donne

$$y' - z' = f(t, y) - f(t, z)$$

et, en supposant f différentiable,

$$y' - z' = \left(\frac{\partial f}{\partial y} \right)_u (y - z) ,$$

où u est strictement compris entre y et z. Il en découle

$$\frac{y' - z'}{y - z} = \left(\frac{\partial f}{\partial y} \right)_u$$

et, après intégration entre t_n et t_{n+1} ,

$$\ln|y(t_{n+1}) - z(t_{n+1})| = \ln|F_h(y_n) - F_h(y_n^*)|$$

$$= \ln|y_n - y_n^*| + \int_{t_n}^{t_{n+1}} \left(\frac{\partial f}{\partial y} \right)_u dt .$$

Soit A un nombre tel que l'on ait toujours

$$\left(\frac{\partial f}{\partial y} \right) \leq A .$$

On obtient

$$\ln | F_h(y_n) - F_h(y_n^*) | \leq \ln |y_n - y_n^*| + A\, h ,$$

soit

$$|F_h(y_n) - F_h(y_n^*)| \leq |y_n - y_n^*|\, e^{Ah} .$$

Si la dérivée $\frac{\partial f}{\partial y}$ est toujours négative, on aura $A < 0$, et l'erreur se propagera en décroissant: le problème différentiel est alors stable

par rapport aux conditions initiales. Dans le cas contraire, on a $A > 0$ et on peut utiliser pour A la constante de LIPSCHITZ L du problème. L'erreur propagée croît exponentiellement, ce qui ne fait que traduire l'instabilité du problème différentiel par rapport aux conditions initiales.

Au total, on a donc

$$|e_{n+1}| \leq |e_n| \, e^{Ah} + M h^{p+1} + C .$$

Soit alors un point $T = t_o + N h$. On a successivement

$$|e_N| \leq e^{Ah} |e_{N-1}| + M h^{p+1} + C$$

$$\leq e^{Ah} \left[e^{Ah} |e_{N-2}| + M h^{p+1} + C \right] + Mh^{p+1} + C$$

$$= e^{2Ah} |e_{N-2}| + (M h^{p+1} + C)(e^{Ah} + 1)$$

.....

$$\leq e^{NAh} |e_o| + (Mh^{p+1} + C)(1 + ... + e^{(N-1)Ah})$$

et, comme $e_o = 0$,

$$|e_N| \leq (Mh^{p+1} + C) \frac{e^{NAh} - 1}{e^{Ah} - 1} = (M h^{p+1} + C) \frac{e^{AT} - 1}{e^{Ah} - 1}$$

Soit d'abord $A < 0$, c'est-à-dire que $\frac{\partial f}{\partial y}$ est toujours négative.

Alors,

$$|e_N| \leq (M h^{p+1} + C) \frac{1}{1 - e^{Ah}}$$

et, pour h suffisamment petit,

$$1 - e^{Ah} = - A h + O(h^2) ,$$

ce qui donne l'estimation asymptotique pour $h \longrightarrow 0$

$$|e_N| \lesssim \frac{1}{|A|} \left| M h^p + \frac{C}{h} \right| .$$

On constate que pour h donné, lorsque l'on s'éloigne de l'origine t_o, l'erreur reste bornée.

Au contraire, si $A > 0$, on a

$$|e_N| \leq (M h^{p+1} + C) \frac{e^{AT}}{e^{Ah} - 1} \lesssim \frac{1}{A} \left| M h^p + \frac{C}{h} \right| e^{AT}$$

A mesure que l'on s'éloigne de l'origine t_o, l'erreur croît exponentiellement, selon une loi qui ne dépend que de la fonction f, à l'exclusion de tout paramètre dépendant du pas ou de l'algorithme utilisé.

Cependant, dans les deux cas, en l'absence d'erreurs d'arrondi et d'évaluation, la solution calculée converge uniformément sur [0 , T] vers la solution exacte, l'erreur décroissant comme h^p, si p est l'ordre de l'algorithme utilisé.

Ce dernier point s'explique aisément par le fait qu'il y a N = T/h pas, ce qui fait perdre un ordre à grande distance.

Mais pour les très petits pas, la solution calculée finit par diverger du fait des erreurs d'arrondi et d'évaluation de f.

Il existe en fait (fig. 4) un pas optimal, correspondant au minimum de l'expression $(M h^p + \frac{C}{h})$, que l'on trouve en annulant sa dérivée:

$$p M h_{opt}^{p-1} - \frac{C}{h_{opt}^2} = 0 ,$$

ce qui donne

$$p M h_{opt}^{p+1} = C .$$

Comme le montre la figure 5, le pas optimal croît généralement avec l'ordre de l'algorithme. On peut en effet admettre que C varie peu avec cet ordre, et les courbes représentant $p Mh^{p+1}$ pour les ordres successifs épousent de plus en plus l'axe des h. Au pas optimal, on a

$$M h_{opt}^p + \frac{C}{h_{opt}} = \frac{p+1}{p} \frac{C}{h_{opt}}$$

et cette valeur décroît lorsque p croît. Ainsi, les algorithmes d'ordre élevé permettent d'obtenir une précision meilleure et ce, pour un pas plus grand que les algorithmes d'ordre plus bas.

7. METHODES A PAS LIES

7.1 - Le principe des méthodes à pas liés est de tenir compte, à chaque pas, de l'allure de la solution obtenue antérieurement. La formule générale sera, pour une formule à r pas,

$$\tilde{y}_k = \alpha_1 \tilde{y}_{k-1} + ... + \alpha_r \tilde{y}_{k-r} = h(\beta_0 \tilde{f}_k + \beta_1 \tilde{f}_{k-1} + ... + \beta_r \tilde{f}_{k-r})$$

où les α_i et β_i sont indépendants de h.

Deux cas sont à distinguer:

a) $\beta_0 = 0$: y_k n'est déterminé que par les valeurs précédentes et par f. On dit que l'algorithme est explicite.

b) $\beta_0 \neq 0$: y_k dépend aussi de $\tilde{f}_k = f(t_k, \tilde{y}_k)$: l'algorithme est implicite, car il ne peut être résolu directement sous

cette forme.

Evidemment, on s'arrangera pour que l'ordre soit le plus élevé possible. Notant que le développement de TAYLOR à l'ordre p est un polynôme de degré p, la formule sera d'ordre p si et seulement si elle est exacte pour tout polynôme de degré p.

Considérant donc les (p+1) fonctions

$$\begin{aligned} &y(t) = 1 \quad \ldots\ldots \quad y'(t) = 0 \\ &y(t) = t \quad \ldots\ldots \quad y'(t) = 1 \\ &\ldots\ldots\ldots \\ &y(t) = t^p \quad \ldots\ldots \quad y'(t) = p\, t^{p-1} \quad , \end{aligned}$$

cette condition mène au système d'équations linéaires (on prend h = 1)

$$\left\{ \begin{aligned} &1 = \alpha_1 + \ldots + \alpha_r \\ &r = \alpha_1 (r-1) + \ldots + \alpha_{r-1} + \beta_0 + \ldots + \beta_r \\ &\ldots\ldots\ldots \\ &r^p = \alpha_1 (r-1)^p + \alpha_2 (r-2)^p + \ldots + \alpha_{r-1} + p(r^{p-1} \beta_0 + (r-1)^{p-1} \beta_1 + \ldots \\ &\qquad\qquad \ldots + \beta_{r-1}) \end{aligned} \right.$$

Pour une formule à r pas, on a (2r+1) paramètres à déterminer. Dès lors, une formule à r pas sera au plus de degré (2r).

8. ALGORITHMES EXPLICITES D'ADAMS

Ils correspondent à $\alpha_1 = 1$, $\alpha_2 = \ldots = \alpha_r = 0$, ce qui donne

$$\tilde{y}_k = \tilde{y}_{k-1} + h \sum_{j=1}^{r} \beta_j \tilde{f}_{k-j} \, .$$

Les coefficients sont donnés par le système

$$\left\{ \begin{aligned} &1 = 1 \\ &r = (r-1) + \sum_{j=1}^{r} \beta_j \quad , \text{ soit} \quad \sum_{j=1}^{r} \beta_j = 1 \\ &r^s = (r-1)^s + s \sum_{j=1}^{r} \beta_j \, (r-j)^{s-1} \, , \end{aligned} \right.$$

$$\text{soit} \quad \sum_{j=1}^{r} (r-j)^{s-1} \beta_j = \frac{r^s - (r-1)^s}{s}$$

On constate qu'il y a en fait p conditions pour l'ordre p. Puisqu'il y a r paramètres, on aura p = r.

Voici les 5 premières formules d'ADAMS explicites:

r	Formule	ordre
1	$\tilde{y}_k = \tilde{y}_{k-1} + h\,\tilde{f}_{k-1}$ (EULER)	1
2	$\tilde{y}_k = \tilde{y}_{k-1} + \frac{h}{2}(3\,\tilde{f}_{k-1} - \tilde{f}_{k-2})$	2
3	$\tilde{y}_k = \tilde{y}_{k-1} + \frac{h}{12}(23\,\tilde{f}_{k-1} - 16\,\tilde{f}_{k-2} + 5\,\tilde{f}_{k-3})$	3
4	$\tilde{y}_k = \tilde{y}_{k-1} + \frac{h}{24}(55\,\tilde{f}_{k-1} - 59\,\tilde{f}_{k-2} + 37\,\tilde{f}_{k-3} - 9\,\tilde{f}_{k-4})$	4
5	$\tilde{y}_k = \tilde{y}_{k-1} + \frac{h}{720}(1901\,\tilde{f}_{k-1} - 2774\,\tilde{f}_{k-2} + 2616\,\tilde{f}_{k-3} - 1274\,\tilde{f}_{k-4} + 251\,\tilde{f}_{k-5})$	5

9. ALGORITHMES IMPLICITES D'ADAMS

Ils correspondent également à $\alpha_1 = 1$, $\alpha_2 = \ldots = \alpha_r = 0$ et sont donc de la forme

$$\tilde{y}_k = \tilde{y}_{k-1} + h \sum_{j=0}^{r} \beta_j \tilde{f}_{k-j} \quad .$$

On détermine les coefficients β_j à partir du système

$$\begin{cases} 1 = 1 \\ r = (r-1) + \sum_{j=0}^{r} \beta_j \quad , \text{ soit} \quad \sum_{j=0}^{r} \beta_j = 1 \\ \cdots\cdots\cdots \\ r^s = (r-1)^s + s \sum_{j=0}^{r} \beta_j (r-j)^{s-1} \end{cases}$$

$$\text{soit} \quad \sum_{j=0}^{r} (r-j)^{s-1} \beta_j = \frac{r^s - (r-1)^s}{s}$$

Il y a ici (r+1) paramètres libres et p conditions pour l'ordre p, puisque la condition d'ordre 0 est triviale. L'ordre p de l'algorithme est donc donné par

$$p = r+1$$

Voici les quatre premières formules d'ADAMS implicites:

r	FORMULE	Ordre
1	$y_k = y_{k-1} + \frac{h}{2}(f_k + f_{k+1})$ (EULER-CAUCHY)	2
2	$y_k = y_{k-1} + \frac{h}{12}(5f_k + 8 f_{k-1} - f_{k-2})$	3
3	$y_k = y_{k-1} + \frac{h}{24}(9 f_k + 19 f_{k-1} - 5 f_{k-2} + f_{k-3})$	4
4	$y_k = y_{k-1} + \frac{h}{720}(251 f_k + 646 f_{k-1} - 264 f_{k-2} + 106 f_{k-3} - 19 f_{k-4})$	5

10. DEMARRAGE D'UN ALGORITHME A PAS LIES (fig. 6)

On ne peut évidemment faire démarrer un algorithme à r pas que si r valeurs de y sont connues. A priori, on ne connaît que y_o, et il faut donc compléter les données en donnant $y_1, \ldots, y_{r-1}$. Le plus souvent, on calcule ces (r-1) valeurs par une méthode de RUNGE-KUTTA. Il faut prendre garde au fait que les erreurs consenties dans cette phase de démarrage se propageront tout au long du calcul. Les calculs de démarrage par RUNGE-KUTTA doivent donc être aussi précis que possible.

11. MISE EN OEUVRE D'UNE METHODE IMPLICITE

Dans une méthode implicite, il se pose le problème de l'évaluation de $\tilde{f}_k$. A cette fin, on peut utiliser:

a) Une méthode itérative (HAMMINGS)

Pour un algorithme d'ADAMS, on aura

$$\tilde{y}_k = \tilde{y}_{k-1} + h\ \beta_o\ f(t_k, \tilde{y}_k) + h \sum_{i=1}^{r} \beta_i\ \tilde{f}_{k-i}\ .$$

Le dernier terme, calculable une fois pour toutes, sera noté B. On a alors

$$\tilde{y}_k = \tilde{y}_{k-1} + h\,\beta_o\ f(t_k, \tilde{y}_k) + B\ .$$

Partant d'une valeur d'essai $\tilde{y}_k^{(0)}$, on pose

$$\tilde{y}_k^{(n+1)} = \tilde{y}_{k-1} + B + h\,\beta_o\ f(t_k, \tilde{y}_k^{(n-1)})\ .$$

Il s'agit d'une méthode itérative classique: il y aura convergence si

$$h|\beta_o|L < 1\ ,$$

soit si

$$h < \frac{1}{|\beta_0| L} \quad .$$

b) Une méthode de prédiction-correction (MILNE)

Au lieu d'itérer, on remplace $\tilde{y}_k$ dans la formule implicite par une valeur y_k^* approchée par une formule explicite. La question se pose alors de choisir l'ordre de la formule explicite donnant y_k^* . Si l'ordre de la formule implicite est p, en calculant y_k^* à l'ordre (p-1), on aura une erreur $O(\ ^{p-1})$ sur cette dernière valeur. Alors, l'erreur sur f_k sera $O(h^{(k-1)})$ également, du fait de la condition de LIPSCHITZ sur f. Cette erreur sera, dans l'algorithme implicite, multipliée par h, si bien que l'algorithme final sera bien d'ordre p. Ainsi, l'étape de prédiction de $\tilde{y}_k$ peut être faite par un algorithme explicite d'ordre (p-1) sans altérer l'ordre final de l'algorithme. Dans le cadre des algorithmes d'ADAMS, cette conclusion est particulièrement utile, car la formule explicite d'ordre (p-1) et la formule implicite d'ordre p comportent le même nombre de pas, ce qui simplifie notamment le démarrage de l'algorithme. Si β_i^* sont les coefficients de l'algorithme explicite, on effectue donc les deux étapes suivantes:

- Prédiction : $y_k^* = \tilde{y}_{k-1} + h \sum_{i=1}^{r} \beta_i^* \tilde{f}_{k-i}$

- Correction : $\tilde{y}_k = \tilde{y}_{k-1} + h \left[\beta_0 f(t_k, y_k^*) + \sum_{i=1}^{r} \beta_i \tilde{f}_{k-i} \right]$.

12. CONDITION DE CONSISTANCE

Considérons un algorithme à pas liés, de la forme

$$\sum_{j=0}^{r} \alpha_j \tilde{y}_{k-j} = h \sum_{j=0}^{r} \beta_j \tilde{f}_{k-j} \quad , \qquad \alpha_0 \neq 0 \qquad (1)$$

Sans faire a priori d'hypothèses sur la manière dont les coefficients α_j et β_j ont été obtenus, examinons dans quelles conditions cet algorithme approche effectivement le problème différentiel

$$y'(t) = f(t, y) \qquad (2)$$

pour $t \in [\,0, T\,]$, en admettant, bien sûr que f est continue de t et lipschitzienne de y sur $[\,0,T\,]$. La solution y est donc continûment dérivable. Définissant l'erreur d'approximation de l'équation (2) par le schéma (1) comme

$$r(t_k) = \frac{1}{h} \sum_{j=0}^{r} \alpha_j y(t_k - jh) - \sum_{j=0}^{r} \beta_j f(t_k - jh, y(t_k - jh)) \ ,$$

on dira que l'algorithme (1) est <u>consistant</u> sur $[\,0,\,T\,]$ si

$$\sup_{t_k \in [\,0,\,T\,]} \| r(t_k) \| \longrightarrow 0$$

pour $h \longrightarrow 0$.

D'après les propriétés imposées à f, on peut développer $y(t_k - jh)$ en

$$y(t_k - jh) = y(t_k) - jh\, y'(t_k - \theta_j\, jh)$$

$$= y(t_k) - jh\, y'(t_k) - jh \left[y'(t_k - j\,\theta_j\, h) - y'(t_k) \right] ,$$

le terme entre crochets convergeant uniformément vers zéro sur le compact $[\,0,\,T\,]$, ce qui donne

$$\frac{1}{h} \sum_{j=0}^{r} \alpha_j\, y(t_k - jh) = \frac{1}{h}\, y(t_k) \sum_{j=0}^{r} \alpha_j + y'(t_k) \left(- \sum_{j=0}^{r} j\,\alpha_j \right) + \frac{o(h)}{h} ;$$

par ailleurs, $f(t, y(t))$ est uniformément continue sur le compact $[\,0,\,T\,]$, si bien que

$$\sum_{j=0}^{r} \beta_j\, f(t_k - jh,\, y(t_k - jh)) = \left(\sum_{j=0}^{r} \beta_j \right) f(t_k,\, y(t_k))$$

$$+ \left[\sum_{j=0}^{r} \beta_j \left(f(t_k - jh,\, y(t_k - jh)) - f(t_k,\, y(t_k)) \right) \right]$$

la quantité entre crochets tendant vers zéro pour $h \longrightarrow 0$. On a donc

$$r(t_k) = \frac{y(t_k)}{h} \sum_{j=0}^{r} \alpha_j + y'(t_k) \left(- \sum_{j=0}^{r} j\,\alpha_j \right) + f(t_k,\, y(t_k)) \sum_{j=0}^{r} \beta_j$$

$$+ \psi(t_k,\, h) ,$$

avec $\psi(t_k,\, h) \longrightarrow 0$. Comme $y'(t_k) = f(t_k,\, y(t_k))$, <u>les conditions de consistance sont</u>:

$$\boxed{\sum_{j=0}^{r} \alpha_j = 0 \quad , \qquad \sum_{j=0}^{r} j\,\alpha_j = - \sum_{j=0}^{r} \beta_j}$$

Ces conditions reviennent à dire que l'algorithme doit être d'ordre 1 au moins. On y adjoint généralement la <u>condition de normalisation</u>

$$\sum_{j=0}^{r} \beta_j = 1 ,$$

qu'il est toujours possible de vérifier si $\sum_{j=0}^{r} \beta_j \neq 0$. Le cas de

nullité de la somme des β_j est en effet à rejeter, car il rendrait impossible l'intégration d'une fonction à dérivée constante. La condition de normalisation entraîne

$$\sum_{j=1}^{r} j\,\alpha_j = -1 \ ,$$

d'où

$$\frac{1}{h} \sum_{j=0}^{r} \alpha_j\, y(t_k - jh) \longrightarrow y'(t_k)$$

et

$$\sum_{j=0}^{r} \beta_j\, f(t_k - jh,\ y(t_k - jh)) \longrightarrow f(t_k,\ y(t_k)).$$

13. CONDITION DE STABILITE

Imaginons qu'à un moment donné, la solution subisse une légère perturbation δy_k, due par exemple aux erreurs d'arrondi et d'évaluation de f ou encore, à un choix imprécis des valeurs de démarrage. Comment se propagera cette erreur? On aura, au premier ordre,

$$\sum_{j=0}^{r} \alpha_j\, \delta y_{k-j} = h \sum_{j=0}^{r} \beta_j \frac{\partial f}{\partial y_{k-j}}\, \delta y_{k-j} \ .$$

Il s'agit d'une équation récurrente en les δy_k :

$$\sum_{j=0}^{r} \left(\alpha_j - h\,\beta_j \frac{\partial f}{\partial y_{k-j}}\right) \delta y_{k-j} = 0.$$

Pour un pas h petit, on peut considérer cette équation comme une perturbation de l'équation

$$\sum_{j=0}^{r} \alpha_j\, \delta y_{k-j} = 0$$

par addition de termes $0(h)$ aux α_j. Afin de cerner aussi simplement que possible le comportement des erreurs au cours des itérations, nous étudierons le problème idéalisé

$$\sum_{j=0}^{r} (\alpha_j - M\,h\,\beta_j)\, \delta y_{k-j} = 0$$

correspondant à l'équation simple

$$y' = M\,y \quad , \quad M = c^{te}.$$

Posons

$$\phi(z) = \sum_{j=0}^{r} \alpha_j\, z^{r-j}$$

et

$$\psi(z) = \sum_{j=0}^{r} \beta_j \, z^{r-j} \ .$$

Il suffit donc d'étudier les racines du polynôme perturbé

$$\phi_h(z) = \phi(z) - M\, h\, \psi(z) \ .$$

Remarquons tout de suite les faits suivants:

a) Si ϕ_h admet une racine z_{1h} telle que $|z_{1h}| \geqslant A > 1$ lorque $h \longrightarrow 0$, il existera une solution de l'équation récurrente de la forme

$$|\delta y_k| = |z_{1h}|^k \geqslant A^k \ .$$

Lorsque $h \longrightarrow 0$, la valeur de l'erreur en $t = T$ fixé correspondra à $k = T/h$, et vaudra

$$|\delta y(T)| = A^{T/h} \longrightarrow \infty \ ,$$

c'est-à-dire que la perturbation croîtra indéfiniment.

b) Les racines de module inférieur à l'unité donnent des solutions décroissantes: si $|z_{ih}| \leqslant \alpha < 1$, on a

$$|z_{ih}|^k \leqslant \alpha^k = \alpha^{T/h} \longrightarrow 0 \ .$$

c) Des racines de module 1 donnent une perturbation stable, qui reste égale à elle-même.

d) Une racine z_{ih} vérifiant

$$|z_{ih}| \leqslant 1 + C\, h \qquad , \ C > 0 \ ,$$

conduit à

$$|\delta y(T)| \leqslant (1 + Ch)^{T/h} \ .$$

Or,

$$\lim_{h \to 0} (1 + Ch)^{T/h} = \left[\lim_{h \to 0} (1 + Ch)^{1/(Ch)} \right]^{CT} = \exp(CT) \ ,$$

soit une valeur bornée indépendamment de h: la perturbation reste stable lorsque $h \longrightarrow 0$.

e) Une racine z_{ih} vérifiant

$$|z_{ih}| \leqslant 1 + C\, h^{1/p} \quad , \quad p > 1 \ ,$$

mène à

$$|\delta y(T)| \geq (1 + C\,h^{1/p})^{T/h} = \left[(1 + C\,h)^{\frac{1}{Ch^{1/p}}} \right]^{\frac{CT}{h^{1-1/p}}}$$

qui tend vers l'infini comme $\exp\left(\frac{CT}{h^{1-1/p}} \right)$.

Il nous reste à examiner comment les racines de $\phi_h(z)$ et celles de $\phi(z)$ sont liées. Soit z_i une racine de ϕ, de multiplicité p. On a donc

$$\phi(z) = (z - z_i)^p\, Q(z) .$$

Trois cas sont possibles:

a) z_i <u>n'est pas racine de ψ</u> - Alors, les racines correspondantes z_{ih} de ϕ_h vérifient

$$0 = \phi_h(z_{ih}) = (z_{ih} - z_i)^p\, Q(z_{ih}) - Mh\,\psi(z_{ih}) ,$$

soit

$$z_{ih} - z_i = \left(\frac{Mh\,\psi(z_{ih})}{Q(z_{ih})} \right)^{1/p} = O(h^{1/p}) .$$

Il existe donc p racines distinctes z_{ih} voisines de z_i de $O(h^{1/p})$.

b) z_i <u>est racine de ψ, de multiplicité $q < p$</u> - Alors,

$$\psi(z) = (z - z_i)^q\, R(z)$$

et

$$0 = \phi_h(z_{ih}) = (z_{ih} - z_i)^q \left[(z_{ih} - z_i)^{p-q}\, Q(z_{ih}) - Mh\, R(z_{ih}) \right] .$$

Dans ce cas, z_i est encore une racine de ϕ_h , de multiplicité q; il apparaît en outre (p-q) racines $z_{ih} \neq z_i$, qui vérifient

$$z_{ih} - z_i = \left(\frac{Mh\, R(z_{ih})}{Q(z_{ih})} \right)^{1/(p-q)} = O\left(h^{1/(p-q)} \right) .$$

c) z_i <u>est racine de ψ, de multiplicité $q \geq p$</u>. - Alors, z_i est encore racine de ϕ_h, de multiplicité p.

Cela étant, nous possédons tous les éléments pour discuter le problème. Tout d'abord, pour une racine z_i de ϕ située <u>à l'intérieur du disque unité</u>, les racines correspondantes de ϕ_h finiront toujours par se retrouver à l'intérieur de ce disque, et le problème de stabilité ne se pose pas. Si ϕ possède une racine de module supérieur à l'unité, pour h suffisamment petit, les racines correspondantes

de ϕ_h seront également de module supérieur à 1, et l'algorithme est instable. Il reste à examiner le cas de racines de ϕ situées sur le cercle unité. Si ces racines sont simples, les racines correspondantes de ϕ_h vérifieront

$$|z_{ih}| \leq 1 + O(h)$$

et la stabilité est assurée. S'il existe une racine z_i de module 1 et de multiplicité $p > 1$, la stabilité n'est assurée que si z_i est également racine de ψ, de multiplicité au moins égale à $(p-1)$. Cependant, cette dernière conclusion n'est vraie que pour le problème modèle idéalisé. Dans la pratique, les dérivées par rapport à y de f changent constamment et tout se passe comme si ψ était modifié à chaque pas. On ne peut donc pas compter sur les zéros de ψ, et il faut exiger que tous les zéros de ϕ de module 1 soient simples.

En résumé, on peut énoncer la condition de stabilité suivante:

Pour que l'algorithme

$$\sum_{j=0}^{r} \alpha_j \tilde{y}_{k-j} = h \sum_{j=0}^{r} \beta_j \tilde{f}_{k-j}$$

soit stable, c'est-à-dire que pour $h \to 0$, les perturbations se propagent en restant bornées, il faut que l'équation caractéristique

$$\phi(z) = \sum_{j=0}^{r} \beta_j z^{r-j}$$

ait toutes ses racines z_i de module inférieur ou égal à l'unité et que les racines de module 1 soient simples.

On remarquera que tout algorithme consistant vérifie $\phi(1) = 0$. Mais s'il est normalisé, on a

$$\phi'(1) = \sum_{j=0}^{r} j \alpha_j = -1 ,$$

ce qui signifie que z=1 est une racine simple.

14. THEOREME DE CONVERGENCE DES METHODES A PAS LIES [8]

La convergence des algorithmes à pas liés est régie par le théorème suivant:

Lorsque $h \to 0$, un algorithme à pas liés consistant, normalisé et stable converge si l'on fait abstraction de l'accumulation des erreurs d'arrondi et d'évaluation de f et si les erreurs sur les valeurs de démarrage tendent vers zéro pour $h \to 0$.

* Démonstration

Notons d'abord que l'équation effectivement résolue est

$$\sum_{j=0}^{r} \alpha_j \tilde{y}_{k-j} - h \sum_{j=0}^{r} \beta_j \tilde{f}_{k-j} = e'_k + e''_k$$

où

e'_k représente l'erreur d'arrondi et d'évaluation de f,

e''_k représente l'erreur éventuelle de non-résolution exacte de l'algorithme, notamment dans le cas d'une méthode de prédiction-correction. Cette erreur vérifie $|e''_k| \leq C\ h^q$.

La solution exacte vérifie quant à elle

$$\sum_{j=0}^{r} \alpha_j\ y(t_{k-j}) - h \sum_{j=0}^{r} \beta_j\ f(t_{k-j},\ y(t_{k-j})) = h\ r_k\ ,$$

avec $r_k \to 0$ pour $h \to 0$. Notant

$$\delta y_k = \tilde{y}(t_k) - y(t_k)\ ,$$

on obtient donc

$$\sum_{j=0}^{r} \alpha_j\ \delta y_{k-j} - h \sum_{j=0}^{r} \beta_j \frac{\partial f}{\partial y}(t_{k-j},\ y^*_{k-j})\ \delta y_{k-j} = e'_k + e''_k - h\ r_k = g_k\ ,$$

où y^*_{k-j} est une valeur intermédiaire entre $y(t_{k-j})$ et $\tilde{y}_{k-j}$. Pour abréger, nous noterons f_{yj} les grandeurs $\frac{\partial f}{\partial y}(t_{k-j},\ y^*_{k-j})$. On suppose naturellement

$$|f_{yk}| \leq L\ .$$

Dans ces notations, on a donc

$$\sum_{j=0}^{r} (\alpha_j - h\ \beta_j\ f_{yj})\ \delta y_{k-j} = g_k\ .$$

On suppose bien entendu que $\alpha_o \neq 0$, et nous nous placerons dans le cas où h est suffisamment petit pour que

$$|\ h\ \beta_o\ L\ | < \alpha_o/2\ .$$

On peut alors expliciter δy_k , ce qui donne

$$\delta y_k = - \sum_{j=1}^{r} \frac{\alpha_j - h\ \beta_j\ f_{yj}}{\alpha_o - h\ \beta_o\ f_{yo}}\ \delta y_{k-j} + \frac{g_k}{\alpha_o - h\ \beta_o\ f_{yo}}$$

On notera que

$$\frac{\alpha_j - h\ \beta_j\ f_{yj}}{\alpha_o - h\ \beta_o\ f_{yo}} = \frac{\alpha_j}{\alpha_o} + h\ v_{kj}\ ,$$

avec

$$v_{kj} = \frac{\alpha_j \beta_o f_{yo} - \alpha_o \beta_j f_{yj}}{\alpha_o (\alpha_o - h \beta_o f_{yo})}$$

et

$$|v_{kj}| \leq 2 \frac{|\alpha_j \beta_o| + |\alpha_o \beta_j|}{\alpha_o^2} L ,$$

ce qui permet d'écrire

$$\delta y_k = - \sum_{j=1}^{r} \frac{\alpha_j}{\alpha_o} \delta y_{k-j} - h \sum_{j=1}^{r} v_{kj} \delta y_{k-j} + w_k , \qquad (1)$$

où

$$w_k = \frac{g_k}{\alpha_o - h \beta_o f_{yo}} , \quad |w_k| \leq \frac{|2 g_k|}{|\alpha_o|} .$$

Les notations matricielles suivantes:

$$z_{(k)} = \begin{bmatrix} \delta y_k \\ \delta y_{k-1} \\ \vdots \\ \delta y_{k-r+1} \end{bmatrix} , \quad A = \begin{bmatrix} -\frac{\alpha_1}{\alpha_o} & -\frac{\alpha_2}{\alpha_o} & \cdots\cdots & -\frac{\alpha_r}{\alpha_o} \\ 1 & 0 & \cdots\cdots & 0 \\ 0 & 1 & \cdots\cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots\cdots & 1 & 0 \end{bmatrix} ,$$

$$V_{(k)} = \begin{bmatrix} v_{k1} & \cdots\cdots & v_{kr} \\ 0 & \cdots\cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots\cdots & 0 \end{bmatrix} , \quad w_{(k)} = \begin{bmatrix} w_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

permettent d'écrire l'équation (1) sous la forme

$$z_{(k)} = A z_{(k-1)} + h V_{(k)} z_{(k-1)} + w_{(k)} \qquad (1')$$

Quelles sont les valeurs propres de A? On vérifie aisément que l'équation

$$A x = \lambda x$$

implique d'une part

$$x_1 = \lambda x_2 , \ \ldots , \ x_{r-1} = \lambda x_r ,$$

soit

$$x_1 = \lambda^{r-1} x_r , \ \ldots , \ x_{r-1} = \lambda x_r$$

et, d'autre part,

$$\lambda \alpha_o x_1 + \alpha_1 x_1 + \ldots + \alpha_r x_r = 0 ,$$

soit

$$\alpha_o \lambda^r + \alpha_1 \lambda^{r-1} + \ldots + \alpha_r = 0 ,$$

ou encore,

$$\phi(\lambda) = 0 ,$$

où $\phi(\lambda) = 0$ est l'équation caractéristique du schéma aux différences. En vertu de la condition de stabilité, toutes les racines de cette équation ont leur module inférieur ou égal à l'unité, et les racines de module 1 sont simples. Nous avons vu, au cours de l'étude des méthodes itératives de résolution des systèmes matriciels, que, dans ces conditions, il existe une matrice inversible S telle que la matrice

$$\hat{A} = S^{-1} A S$$

vérifie

$$\|\hat{A}\|_\infty \leq 1 .$$

Multiplions l'équation (1') par S^{-1}: on obtient

$$\hat{z}_{(k)} = \hat{A}\, \hat{z}_{(k-1)} + h\, \hat{V}_{(k)}\, \hat{z}_{(k-1)} + \hat{w}_{(k)} \qquad (2)$$

avec les notations

$$\hat{z}_{(k)} = S^{-1} z_{(k)} \quad , \quad \hat{z}_{(k-1)} = S^{-1} z_{(k-1)} \quad , \quad \hat{V}_{(k)} = S^{-1} V_{(k)} S ,$$

$$\hat{w}_{(k)} = S^{-1} w_{(k)} .$$

En ce qui concerne les normes, on a donc $\|\hat{A}\|_\infty \leq 1$. Pour $\|\hat{V}_{(k)}\|_\infty$, on notera que la matrice $V_{(k)}$ n'a qu'une ligne non nulle, d'où

$$\|V_{(k)}\|_\infty = \sum_{j=1}^{r} |v_{kj}| \leq \frac{2L}{a_o^2} \sum_{j=1}^{k} (|\alpha_j \beta_o| + |\alpha_o \beta_j|) = \mathcal{V} L ,$$

$\mathcal{V}$ ne dépendant que des coefficients de l'algorithme. Il vient donc

$$\|\hat{V}_{(k)}\|_\infty \leq \|S^{-1}\|_\infty \|S\|_\infty \mathcal{V} L = \gamma L .$$

La constante ne dépend également que des coefficients de l'algorithme, car tel est le cas de la matrice S. Enfin, pour $\|\hat{w}\|_\infty$, on a

$$\|\hat{w}_{(k)}\|_\infty \leq \|S^{-1}\|_\infty \|w_{(k)}\|_\infty \leq \|S^{-1}\|_\infty |w_k|$$

$$\leq 2 \|S^{-1}\| \frac{|g_k|}{|\alpha_o|} = \beta |g_k| ,$$

β ne dépendant que des coefficients de l'algorithme.

Il résulte de tout cela que

$$\|\hat{z}_{(k)}\|_\infty \leq \|\hat{z}_{(k-1)}\|_\infty (1 + h\gamma L) + \beta |g_k| \qquad (3)$$

De proche en proche, et en notant que le premier vecteur $\hat{z}_{(1)}$ qui soit défini est $\hat{z}_{(r-1)}$, on obtient

$$\|\hat{z}_{(k)}\|_\infty \leq \sum_{l=0}^{k-r} |g_{k-l}| (1 + h\gamma L)^l + (1 + h\gamma L)^{k-r+1} \|\hat{z}_{(r-1)}\|_\infty . \qquad (4)$$

On a encore

$$\sum_{l=0}^{k-r} |g_{k-l}| (1 + h\gamma L)^l \leq \sup_{l = r,\ldots,k} |g_l| \sum_{l=0}^{k-r} (1 + h\gamma L)^{k-l}$$

$$\sup_{l=r,\ldots,k} |g_l| \frac{(1 + h\gamma L)^{k-r+1} - 1}{h\gamma L}$$

et, en vertu de la relation générale pour $x > 0$

$$(1 + x) \leq e^x = 1 + x + x^2/2 + \ldots ,$$

on a

$$(1 + h\gamma L)^{k-r+1} \leq \exp\left[(k-r+1)h\gamma L \right] \leq \exp(kh\gamma L) = \exp(\gamma LT),$$

en notant $T = kh$. Il vient donc, en substituant la notation $\hat{z}(T)$ à la notation $\hat{z}_{(k)}$,

$$\|\hat{z}(T)\|_\infty \leq \beta \sup_{l=r,\ldots,k} |g_l| \frac{\exp(\gamma LT) - 1}{h\gamma L} + \exp(\gamma LT) \|\hat{z}_{(r-1)}\|_\infty .$$

Il reste à évaluer

$$\|\hat{z}_{(r-1)}\|_\infty \leq \|S^{-1}\|_\infty \|z_{(r-1)}\|_\infty = \|S^{-1}\|_\infty \sup_{l=1,\ldots,r-1} |\delta y_l|$$

Notant finalement que

$$\|z(T)\|_\infty \leq \|S\|_\infty \|\hat{z}(T)\|_\infty ,$$

on obtient

$$\|z(T)\|_\infty \leq C_1 \sup_{l=r,\ldots,k} |g_l| \cdot \frac{\exp(\gamma LT)}{h\gamma L} + C_2 \exp(\gamma LT) \sup_{l=1,\ldots,r-1} |\delta y_l| ,$$

les constantes C_1 et C_2 ne dépendant que des coefficients de l'algorithme. En réintroduisant les grandeurs composant $|g_l|$, on obtient

$$|\delta y(T)| \leq \|z(T)\|_\infty$$

$$\leq C_1 \sup_{l=r,\ldots,k} |r_l| + \sup_{l=r,\ldots,k} \frac{|e'_l|}{h} + \sup_{l=r,\ldots,k} \frac{|e''_l|}{h} \frac{\exp(\gamma LT)-1}{\gamma L}$$

$$+ C_2 \exp(\gamma LT) \sup_{l=1,\ldots,r-1} |\delta y_l| .$$

La condition de consistance implique

$$\lim_{h \to 0} \sup_{l=r,\ldots,k} |r_l| = 0 .$$

Les erreurs de résolution de l'algorithme devront en outre vérifier

$$\lim_{h \to 0} \sup_{l=r,\ldots,k} \frac{|e''_k|}{h} = 0 , \quad \text{soit} \quad |e''_k| = o(h) .$$

Les termes $|\delta y_l|$, $l = 1, \ldots, r-1$, représentent les <u>erreurs sur les conditions de démarrage</u>, qui doivent également tendre vers zéro pour $h \to 0$.

La seule source de divergence réside alors dans les erreurs d'arrondi et d'évaluation qui entraînent une erreur $O(1/h)$.

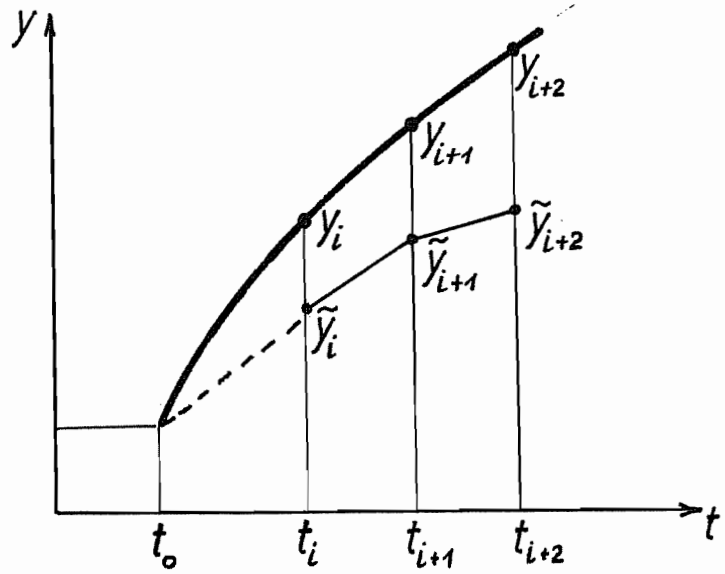


Fig. 1

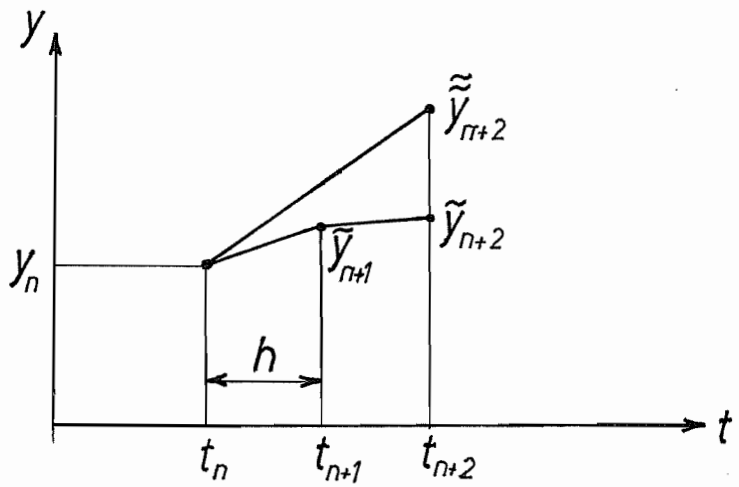


Fig. 2

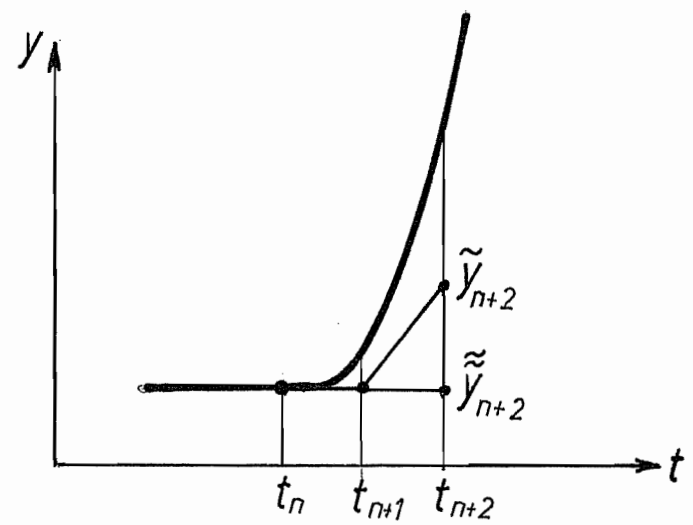


Fig. 3

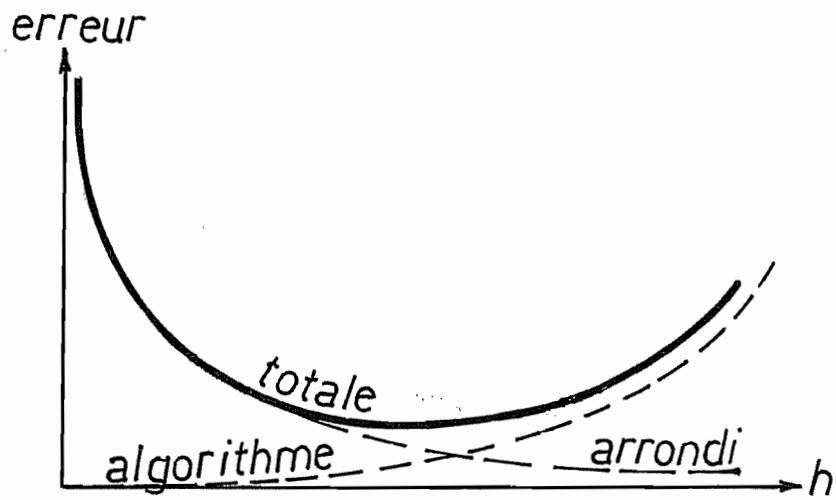


Fig. 4

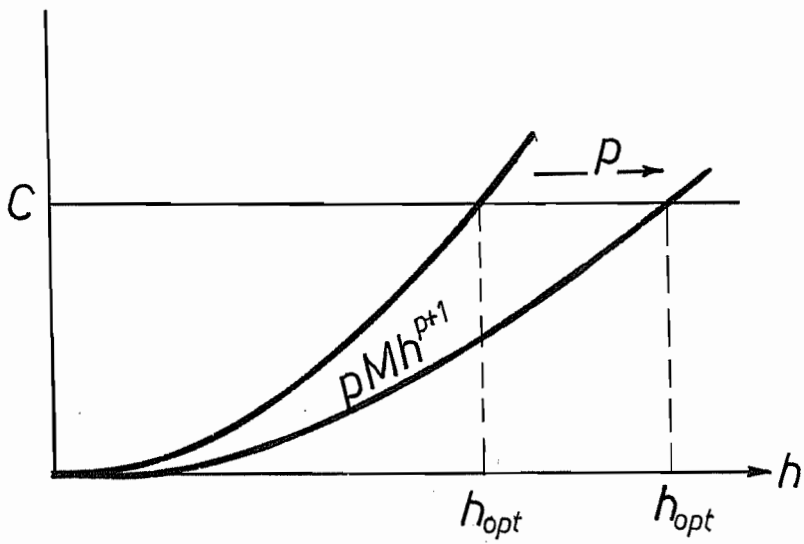


Fig. 5

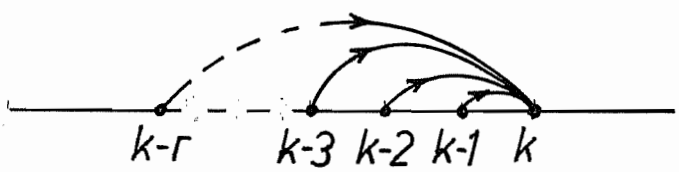


Fig. 6

S O M M A T I O N D E S S E R I E S

1. La sommation des séries n'est pas toujours une opération aussi simple qu'on pourrait le penser. Si certaines séries, comme

$$\sum_{k=0}^{\infty} \frac{1}{(k+1)k!} = 1{,}718281830$$

s'obtiennent aisément par simple sommation, (treize termes ont été suffisants pour stabiliser la neuvième décimale), il n'en est pas de même de certaines séries convergeant lentement comme

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}$$

pour laquelle il faudrait <u>deux milliards</u> de termes pour garantir la neuvième décimale. Il existe cependant un certain nombre de méthodes permettant d'accélérer la convergence des séries.

2. EVALUATION DU RESTE D'UNE SERIE

Pour déterminer le nombre de termes à prendre en considération pour obtenir une précision donnée sur une série, on peut le plus souvent faire appel aux deux critères suivants:

2.1 - Reste d'une série alternée

Soit la série $\sum (-1)^k a_k$, $a_k > 0$, $a_k \downarrow 0$. On a, si M a la même parité que (N+1),

$$(-1)^{N+1} \sum_{k=N+1}^{M} (-1)^k a_k = \left[a_{N+1} - (a_{N+2} - a_{N+3}) - (a_{N+4} - a_{N+5}) - \dots \right.$$
$$\left. \dots - (a_{M+1} - a_M) \right] \quad \begin{cases} \leq a_{N+1} \\ \geqslant 0 \end{cases}$$

Si M a la parité de N ,

$$(-1)^{N+1} \sum_{k=N+1}^{M} (-1)^k a_k = \left[a_{N+1} - (a_{N+2} - a_{N+3}) - \dots \right.$$
$$\left. \dots - (a_{M-2} - a_{M-1}) - a_M \right] \quad \begin{cases} \leq a_{N+1} \\ \geqslant 0 \end{cases}$$

Ceci étant vrai pour tout N, les sommes de CAUCHY tendent vers zéro (ce qui garantit la convergence) et, pour $M \longrightarrow \infty$, on obtient

$$|\mathcal{R}_N| = \left| \sum_{k=N+1}^{\infty} (-1)^k a_k \right| \leq a_{N+1} \quad , \quad \text{sign } \mathcal{R}_N = (-1)^{N+1}$$

2.2 - Reste des séries sommables à termes positifs décroissants

Considérons la série $\sum_{n=0}^{\infty} f(n)$, sommable, avec f mesurable,

$f \geqslant 0$, $f(x) \not\to 0$ pour $x \longrightarrow \infty$.

La fonction dénombrablement étagée

$g(x) = f(\text{ent}(x))$, ent(x) = partie entière de x,

est intégrable, car

$$\int_0^{\infty} g(x)\, dx = \sum_{n=0}^{\infty} f(n) .$$

Comme on a partout (fig. 1)

$g(x) = f(\text{ent}(x)) \geqslant f(x)$,

on en déduit que f est intégrable. Il en est de même de la fonction

$h(x) = f(x-1)$ définie sur $\rbrack +1, \infty \lbrack$, car

$$\int_1^{\infty} h(x)\, dx = \int_1^{\infty} f(x-1)\, dx = \int_0^{\infty} f(x)\, dx.$$

Par ailleurs, on a partout

$$f(x) \leqslant g(x) \leqslant h(x) ,$$

ce qui implique que, sur tout intervalle $\lbrack N+1, \infty \lbrack$,

$$\int_{N+1}^{\infty} f(x)\, dx \leqslant \int_{N+1}^{\infty} g(x)\, dx \leqslant \int_{N+1}^{\infty} h(x)\, dx ,$$

soit

$$\boxed{\int_{N+1}^{\infty} f(x)\, dx \leqslant \mathcal{R}_N \leqslant \int_{N}^{\infty} f(x)\, dx}$$

2.3 - Exemples

a) La série

$$-\sum_{k=2} \left(\frac{1}{k} + \ln \frac{k-1}{k} \right)$$

est à termes positifs, car

$$\ln\left(1 - \frac{1}{k}\right) = -\frac{1}{k} - \frac{1}{2k^2} - \frac{1}{3k^3} - \ldots < -\frac{1}{k}$$

Ses termes sont décroissants. Après N termes, on a

$$-\int_{N+1}^{\infty} \left[\frac{1}{x} + \ln\left(1 - \frac{1}{x}\right) \right] dx = -\left[\ln x + x \ln\left(1 - \frac{1}{x}\right) - \ln(x-1) \right]_{N+1}^{\infty} =$$

$$= \left[(x - 1) \ln(1 - \frac{1}{x}) \right]_{N+1}^{\infty} .$$

A l'infini, on a, par le théorème de l'HOSPITAL,

$$\lim_{x \to \infty} \frac{\ln(1 - \frac{1}{x})}{\frac{1}{x - 1}} = \lim \frac{\frac{1}{x^2 - x}}{-\frac{1}{(x-1)^2}} = - 1 ,$$

d'où

$$- \int_{N+1}^{\infty} \left[\frac{1}{x} + \ln(1 - \frac{1}{x}) \right] dx = - 1 - N \ln \frac{N}{N-1}$$

(Cette valeur tend vers zéro pour $N \to \infty$) , et

$$- 1 - N \ln \frac{N}{N+1} \leq \mathcal{R}_N \leq - 1 - (N-1) \ln \frac{N-1}{N} .$$

Comme

$$1 - N \ln(1 - \frac{1}{N}) = -1 - N(- \frac{1}{N} - \frac{1}{2N^2} \ldots) \approx \frac{1}{2N} ,$$

on constate que la convergence de cette série est <u>très lente</u>

b) La série

$$\sum_{k=1}^{\infty} \frac{(-1)^{n+1}}{n^3}$$

donne un reste

$$|\mathcal{R}_N| \leq \frac{1}{(N+1)^3} .$$

Il faudra, pour obtenir 3 chiffres significatifs exacts, calculer N tel que

$$\frac{1}{(N+1)^3} \leq \tfrac{1}{2} \cdot 10^{-3} ,$$

soit

$$N+1 \geq 2^{1/3} \cdot 10 = 12,60 .$$

Douze termes feront l'affaire. Pour obtenir six chiffres exacts, il en faudra 125.

3. ACCUMULATION DES ERREURS

A quoi bon, se dira-t-on peut-être, accélérer la convergence des séries, alors que les ordinateurs actuels peuvent effectuer des opérations nombreuses en si peu de temps? Nous avons, dans le chapitre relatif aux erreurs, montré que les arrondis peuvent détruire la con-

vergence et qu'il est de la plus grande importance que les séries convergent vite, sans quoi les erreurs numériques fausseront complètement le résultat. L'accélération de la convergence des séries n'est donc pas un problème d'ordre esthétique, mais une nécessité absolue si l'on veut que les calculs soient couronnés de succès.

4. POLYNOMES DE BERNOULLI

4.1 - Les polynômes de Bernoulli permettent d'obtenir la somme d'un certain nombre de séries particulières, dont nous nous servirons. Pour les introduire, considérons le développement

$$\frac{t\, e^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n(x)}{n!}\, t^n$$

au voisinage de $t = 0$. Pour obtenir les coefficients $B_n(x)$, on écrit

$$t\, e^{xt} = (e^t - 1) \sum_{n=0}^{\infty} \frac{B_n(x)}{n!}\, t^n = \left(\sum_{k=1}^{\infty} \frac{t^k}{k!} \right) \left(\sum_{n=0}^{\infty} \frac{B_n(x)}{n!}\, t^n \right)$$

soit encore

$$\sum_{n=0}^{\infty} \frac{x^n\, t^{n+1}}{n!} = \left(\sum_{k=1}^{\infty} \frac{t^k}{k!} \right) \left(\sum_{n=0}^{\infty} \frac{B_n(x)}{n!}\, t^n \right) .$$

En identifiant les coefficients des différentes puissances de t, on trouve, pour $n = 0$,

$$1 = B_0(x)$$

et, pour $n \quad 1$,

$$\frac{x^{n-1}\, t^n}{(n-1)!} = \sum_{k=1}^{\infty} \frac{t^k}{k!}\, \frac{B_{n-k}(x)}{(n-k)!}\, t^{n-k} \quad ,$$

soit

$$n\, x^{n-1} = \sum_{k=1}^{\infty} \frac{n!}{k!\,(n-k)!}\, B_{n-k}(x)$$

ou encore,

$$\boxed{\sum_{k=1}^{n} C_n^k\, B_{n-k}(x) = n\, x^{n-1} \qquad , \qquad B_0(x) = 1}$$

Ces relations permettent de déterminer les $B_n(x)$ par récurrence. Elles montrent en outre qu'il s'agit de polynômes. On les appelle polynômes de Bernoulli. Il est facile de vérifier que leur coefficient de tête vaut 1. Leurs valeurs à l'origine,

$$B_n = B_n(0) \quad ,$$

sont appelées nombres de Bernoulli. Il est aisé de voir que ce sont des nombres rationnels.

4.2 - Montrons la relation

$$\boxed{B_n(x+1) - B_n(x) = n\, x^{n-1}}$$

Elle est évidente pour n = 0. Si elle est vraie jusqu'à l'ordre (n-1), montrons-la pour B_n. On a, par la relation de récurrence,

$$\sum_{k=1}^{n+1} C_{n+1}^{k} \left[B_{n+1-k}(x+1) - B_{n+1-k}(x) \right] = (n+1) \left[(x+1)^n - x^n \right] ,$$

soit

$$(n+1) \left[B_n(x+1) - B_n(x) \right] + \sum_{k=2}^{n+1} C_{n+1}^{k} \left[B_{n+1-k}(x+1) - B_{n+1-k}(x) \right]$$

$$= (n+1) \left[(x+1)^n - x^n \right] .$$

La proposition étant vraie jusqu'à l'ordre (n-1), la somme du premier membre vaut

$$\sum_{k=2}^{n+1} C_{n+1}^{k} (n-k+1)\, x^{n-k} = \sum_{k=2}^{n} C_{n+1}^{k} (n-k+1)\, x^{n-k} =$$

$$= \sum_{k=2}^{n} \frac{(n+1)!\,(n-k+1)}{k!\,(n-k+1)!}\, x^{n-k}$$

$$= (n+1) \sum_{k=2}^{n} C_n^k\, x^{n-k}$$

$$= (n+1) \left[(1+x)^n - x^n - C_n^1\, x^{n-1} \right] ,$$

ce qui donne finalement

$$(n+1) \left[B_n(x+1) - B_n(x) \right] = (n+1)\, n\, x^{n-1} ,$$

comme annoncé.

On en déduit immédiatement que, pour n > 1,

$$B_n(1) = B_n(0) + 0 ,$$

soit

$$\boxed{B_n(1) = B_n \qquad \text{pour} \quad n > 1}$$

4.3 - Montrons que

$$\boxed{B'_n(x) = n\, B_{n-1}(x)}$$

On a évidemment $B'_0(x) = 0$. Supposant la relation ci-dessus vraie jusqu'en (n-1), on obtient, en dérivant la relation fondamentale de récurrence,

$$\sum_{k=1}^{n+1} C_{n+1}^{k}\, B'_{n-k+1}(x) = (n+1)n\, x^{n-1}$$

et, comme la proposition est vraie jusqu'à l'ordre (n-1),

$$(n+1)\, B'_n(x) = (n+1)n\, x^{n-1} - \sum_{k=2}^{n+1} (n-k+1)\, C_{n+1}^{k}\, B_{n-k}(x)$$

La somme s'écrit encore

$$(n+1) \sum_{k=2}^{n} \frac{n!}{k!\,(n-k)!}\, B_{n-k}(x) = (n+1) \sum_{k=2}^{n} C_n^k\, B_{n-k}(x)$$

$$= (n+1)\, [\, n\, x^{n-1} - C_n^1\, B_{n-1}(x)\,] \ ,$$

ce qui donne

$$(n+1)\, B'_n(x) = (n+1)n\, B_{n-1}(x) \ ,$$

ce qu'il fallait démontrer.

4.4 - <u>Les nombres de Bernoulli impairs, sauf</u> B_1, <u>sont nuls</u>.

En effet, faisant x = 0 dans la relation de définition des polynômes de Bernoulli, on obtient

$$\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n\, t^n}{n!} = 1 + B_1\, t + \sum_{n=2}^{\infty} \frac{B_n}{n!}\, t^n \ .$$

Il est aisé de voir, à partir de la relation de récurrence, que $B_1 = -1/2$, d'où

$$1 + \sum_{n=2}^{\infty} \frac{B_n}{n!}\, t^n = \frac{t}{e^t - 1} + \frac{t}{2} = \frac{t}{2}\, \frac{e^t + 1}{e^t - 1} = \frac{t/2}{\mathrm{th}(t/2)} \ .$$

La fonction du dernier membre étant paire, son développement en série de puissances ne peut contenir que des puissances paires de t, ce qui démontre la proposition.

4.5 - On a la relation

$$\boxed{B_n(1-x) = (-1)^n\, B_n(x)}$$

Pour n = 0 , c'est évident. Si c'est vrai pour $B_{n-1}(x)$, on note que

$$B_n(1-x) = B_n(1) + \int_1^{1-x} n\, B_{n-1}(t)\, dt = B_n(1) + \int_0^x n\, B_{n-1}(1-t)\, d(1-t)$$

$$= B_n(1) + (-1)^n \int_0^x n\, B_{n-1}(t)\, dt = B_n(1) + (-1)^n \int_0^x B'_n(x)\, dx$$

$$= B_n(1) - (-1)^n\, B_n(0) + (-1)^n\, B_n(x) \ .$$

Or, on a toujours

$$B_n(1) = (-1)^n\, B_n(0) \ .$$

C'est évident pour n pair, et pour n impair, $B_n = 0$, sauf dans le cas $n = 1$, pour lequel $B_1 = -1/2$ et $B_1(1) = 1/2$. Ceci achève la démonstration.

4.6 - On a également

$$\boxed{(-1)^n\, B_n(-x) = B_n(x) + n\, x^{n-1}}$$

En effet,

$$B_n(-x+1) = B_n(-x) + n\, (-x)^{n-1} \ ,$$

d'où

$$B_n(-x) = B_n(1-x) - (-1)^{n-1}\, n\, x^{n-1} = (-1)^n\, B_n(x) - (-1)^{n-1}\, x^{n-1}$$

$$= (-1)^n \left[B_n(x) + n\, x^{n-1} \right] .$$

4.7 - Les polynômes de Bernoulli admettent le développement de TAYLOR

$$\boxed{B_n(x+h) = \sum_{k=0}^{n} C_n^k\, B_k(x)\, h^{n-k}}$$

En effet,

$$B_n(x+h) = \sum_{k=0}^{n} \frac{B_n^{(k)}(x)}{k!}\, h^k = \sum_{k=0}^{n} \frac{n \ldots (n-k+1)}{k!}\, B_{n-k}(x)\, h^k$$

$$= \sum_{k=0}^{n} C_n^k\, B_{n-k}(x)\, h^k = \sum_{l=0}^{n} C_n^{n-l}\, B_l(x)\, h^{n-l}$$

$$= \sum_{l=0}^{n} C_n^l\, B_l(x)\, h^{n-l} \ .$$

En particulier, en faisant $x = 0$, $h = x$, on obtient

$$\boxed{B_n(x) = \sum_{k=0}^{n} C_n^k\, B_k\, x^{n-k}}$$

4.8 - Les polynômes de Bernoulli vérifient la relation

$$\sum_{k=1}^{m} k^n = \frac{B_{n+1}(m+1) - B_{n+1}}{n+1} \qquad n,m = 1, 2, \ldots$$

En effet,

$$\sum_{k=1}^{m} k^n = \sum_{k=1}^{m} \frac{B_{n+1}(k+1) - B_{n+1}(k)}{n+1} = \frac{B_{n+1}(m+1) - B_{n+1}(1)}{n+1}$$

4.9 - Développement des polynômes de Bernoulli en séries de FOURIER sur $]0, 1[$.

Commençons par chercher les coefficients de Fourier de $B_1(x)$. On a

$$c_k^{(1)} = 2\int_0^1 B_1(x)\, e^{2i\pi kx}\, dx = 2\left[B_1(x)\, \frac{e^{2i\pi kx}}{2i\pi k} \right]_0^1 - 2\int_0^1 B_o(x) \frac{e^{2i\pi kx}}{2i\pi k}\, dx$$

et, comme $B_o(x) = 1$, $B_1(x) = x - 1/2$,

$$c_k^{(1)} = \frac{2}{2i\pi k} = -\frac{2i}{2\pi k} = a_k^{(1)} + i\, b_k^{(1)} ,$$

c'est-à-dire

$$a_k^{(1)} = 0 \quad , \quad b_k^{(1)} = -\frac{2}{2\pi k} .$$

On a donc

$$B_1(x) = -\frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\sin 2\pi kx}{k\pi} = x - \tfrac{1}{2}$$

Passons à $n > 1$. On a, pour $k \neq 0$,

$$c_k^{(n)} = 2\int_0^1 B_n(x)\, e^{2i\pi kx}\, dx = 2\left[B_n(x)\, \frac{e^{2i\pi kx}}{2i\pi k} \right]_0^1 - 2n \int_0^1 B_{n-1}(x) \frac{e^{2i\pi kx}}{2i\pi k}\, dx$$

et, comme $B_n(0) = B_n(1)$, le terme intégré s'annule. De proche en proche, on arrive à

$$c_k^{(n)} = (-1)^{n-1}\, n(n-1)\ldots 2 \int_0^1 B_1(x)\, \frac{e^{2i\pi kx}}{(2i\pi k)^{n-1}}\, dx = \frac{(-1)^{n-1}\, n!}{(2i\pi k)^{n-1}}\, c_k^{(1)} ,$$

soit

$$c_k^{(n)} = 2\, \frac{(-1)^n\, n!}{(2i\pi k)^n} \quad , \quad k \neq 0 .$$

Pour $k = 0$,

$$c_o^{(n)} = \int_0^1 B_n(x)\,dx = \frac{1}{n+1}\int_0^1 B'_{n+1}(x)\,dx = \frac{1}{n+1}\,(B_{n+1}(1) - B_{n+1}) = 0 .$$

Dès lors, <u>pour n pair</u>, $n = 2m$ et $i^{2m} = (-1)^{m-1}$, ce qui donne

$$c_k^{(2m)} = \frac{(-1)^{m-1}.2.(2m)!}{(2\pi)^{2m}} \cdot \frac{1}{k^{2m}} = a_k^{(2m)} ,$$

et

$$B_{2m}(x) = \frac{(-1)^{m-1}.2.(2m)!}{(2\pi)^{2m}} \sum_{k=1}^{\infty} \frac{\cos 2k\pi x}{k^{2m}} \qquad m = 1, 2, \ldots$$

<u>Pour n impair</u>, $n = 2m-1$ et

$$\left(\frac{1}{i}\right)^{2m-1} = \frac{i}{i^{2m}} = (-1)^m\, i ,$$

d'où

$$c_k^{(2m-1)} = 2i\,\frac{(-1)^m\,(2m-1)!}{(2\pi)^{2m-1}} = i\,b_k^{(2m-1)} ,$$

ce qui donne

$$B_{2m-1}(x) = \frac{(-1)^m.2.(2m-1)!}{(2\pi)^{2m-1}} \sum_{k=1}^{\infty} \frac{\sin 2k\pi x}{k^{2m-1}} \qquad m = 1, 2, \ldots$$

4.6 - <u>Séries de puissances inverses</u>

Pour $n = 2m$ et $x=0$, on obtient directement

$$B_{2m} = \frac{(-1)^{m-1}.2.(2m)!}{(2\pi)^{2m}} \sum_{k=1}^{\infty} \frac{1}{k^{2m}}$$

soit

$$\sum_{k=1}^{\infty} \frac{1}{k^{2m}} = (-1)^{m-1}.\, B_{2m}\, \frac{(2\pi)^{2m}}{2.(2m)!} \qquad m = 1,2,\ldots$$

Pour $n = 2m-1$, le problème est un peu plus complexe, car les sinus que contient le développement de $B_{2m-1}(x)$ n'ont pas la valeur 1 simultanément. Calculons l'intégrale

$$\mathcal{J}_k = \int_0^1 \sin 2k\pi x \,\operatorname{cotg}\pi x\,dx .$$

On a d'une part

$$\int_0^1 \frac{\sin 2k\pi x \ \cos\pi x}{\sin\pi x}\, dx =$$

$$= \int_0^1 \frac{\sin(2k-1)\pi x \cos^2\pi x + \cos(2k-1)\pi x \sin\pi x \cos\pi x}{\sin\pi x}\, dx =$$

$$= \int_0^1 \frac{\sin(2k-1)\pi x}{\sin\pi x}\, dx + \int_0^1 [\cos(2k-1)\pi x \cos\pi x - \sin(2k-1)\pi x \sin\pi x]\, dx$$

soit

$$\mathcal{J}_k = \int_0^1 \frac{\sin(2k-1)\pi x}{\sin\pi x}\, dx + \int_0^1 \cos 2k\pi x\, dx = \int_0^1 \frac{\sin(2k-1)\pi x}{\sin\pi x}\, dx .$$

D'autre part,

$$\int_0^1 \frac{\sin 2k\pi x \cos\pi x}{\sin\pi x}\, dx = \tfrac{1}{2}\left\{ \int_0^1 \frac{\sin(2k+1)\pi x}{\sin\pi x}\, dx + \int_0^1 \frac{\sin(2k-1)\pi x}{\sin\pi x}\, dx \right\},$$

ce qui entraîne

$$\mathcal{J}_k = \int_0^1 \frac{\sin(2k-1)\pi x}{\sin\pi x}\, dx = \int_0^1 \frac{\sin(2k+1)\pi x}{\sin\pi x}\, dx = \mathcal{J}_{k+1}$$

Il en découle

$$\mathcal{J}_k = \mathcal{J}_{k-1} = \ldots = \mathcal{J}_1 = \int_0^1 \frac{\sin\pi x}{\sin\pi x}\, dx = 1 .$$

Cette propriété nous permet de traiter le cas $n = 2m-1$, car elle entraîne

$$\int_0^1 B_{2m+1}(x)\, \mathrm{cotg}\,\pi x\, dx = \frac{(-1)^{m+1} \cdot 2 \cdot (2m+1)!}{(2\pi)^{2m+1}} \sum_{k=1}^{\infty} \frac{1}{k^{2m+1}} ,$$

soit

$$\boxed{\sum_{k=1}^{\infty} \frac{1}{k^{2m+1}} = \frac{(-1)^{m+1}\,(2\pi)^{2m+1}}{2 \cdot (2m+1)!} \int_0^1 B_{2m+1}(x)\, \mathrm{cotg}\,\pi x\, dx}$$

ce qui ramène le calcul de la série à celui d'une intégrale. Comme $B_{2m+1}(x)$ s'annule en x=0 et x=1, l'intégrand ne comporte pas de singularité et on peut utiliser une formule de GAUSS.

Exemple: calcul de $\Sigma(1/n^3)$. On a

$$B_3(x) = x^3 - \tfrac{3}{2} x^2 + \tfrac{1}{2} x .$$

Utilisons une formule de GAUSS à 9 points, en tenant compte de la symétrie de la fonction par rapport au milieu de l'intervalle:

x_i	H_i	$B_3(x_i)\ \mathrm{cotg}\ \pi x_i$
0,015 919 881	0,081 274 388	0,151 507 980
0,081 984 447	0,180 648 161	0,119 436 677
0,193 314 284	0,260 610 964	0,068 821 036
0,337 873 289	0,312 347 077	0,020 256 430
0,5	-	0

$I = 0,058\ 152\ 288$

On a alors $\Sigma(1/n^3) = \frac{(2\pi)^3}{2.3!}\ I = 1,202\ 057\ 281$.

La vraie valeur est 1,202 056 903. L'erreur est donc de 4.10^{-7} . Rien n'empêche, bien sûr, d'utiliser des formules de degré plus élevé. Pour obtenir la même précision par sommation directe, il aurait fallu additionner 1118 termes.

5. POLYNOMES D'EULER

5.1 - Les polynômes d'Euler sont définis par le développement

$$\frac{2\,e^{xt}}{e^t + 1} = \sum_{n=0}^{\infty} \frac{E_n(x)}{n!}\ t^n\ ,$$

au voisinage de t=0. Pour obtenir les coefficients des $E_n(x)$, on écrit

$$2\,e^{xt} = (e^t + 1) \sum_{n=0}^{\infty} \frac{E_n(x)}{n!}\ t^n = (2 + \sum_{k=1}^{\infty} \frac{t^k}{k!})(\sum_{n=0}^{\infty} \frac{E_n(x)}{n!}\ t^n)\ ,$$

soit encore

$$2 \sum_{n=0}^{\infty} \frac{t^n x^n}{n!} = (2 + \sum_{k=1}^{\infty} \frac{t^k}{k!})\ (\sum_{n=0}^{\infty} \frac{E_n(x)}{n!}\ t^n)\ .$$

En identifiant les coefficients des différentes puissances de t, on trouve

$$E_o(x) = 1$$

et, pour $n \geq 1$,

$$2\,\frac{x^n t^n}{n!} = 2\,\frac{E_n(x)}{n!}\ t^n + \sum_{k=1}^{n} \frac{t^k}{k!}\ \frac{E_{n-k}(x)}{(n-k)!}\ t^{n-k}\ ,$$

soit

$$2\,x^n = 2\,E_n(x) + \sum_{k=1}^{n} \frac{n!}{k!\ (n-k)!}\ E_{n-k}(x)$$

ou encore,

$$\boxed{E_n(x) = x^n - \tfrac{1}{2} \sum_{k=1}^{n} C_n^k E_{n-k}(x) \qquad ; \quad E_0(x) = 1}$$

On appelle <u>nombres d'Euler</u> les nombres

$$E_n = 2^n E_n(\tfrac{1}{2}) .$$

5.2 - Montrons la relation

$$\boxed{E_n'(x) = n E_{n-1}(x)}$$

Pour n = 1, on a

$$E_1(x) = x - \tfrac{1}{2} E_0(x) = x - \tfrac{1}{2} ,$$

donc

$$E_1'(x) = 1 = E_0(x) .$$

Supposons donc la relation vraie jusqu'à l'ordre (n-1). Alors,

$$E_n'(x) = n x^{n-1} - \tfrac{1}{2} \sum_{k=1}^{n} C_n^k E_{n-k}'(x) = n x^{n-1} - \tfrac{1}{2} \sum_{k=1}^{n-1} C_n^k E_{n-k}'(x) ,$$

car $E_0'(x) = 0$, d'où

$$E_n'(x) = n x^{n-1} - \frac{n}{2} \sum_{k=1}^{n-1} \frac{(n-1)!}{k!\,(n-k+1)!} E_{n-k+1}(x)$$

$$= n\Big(x^{n-1} - \tfrac{1}{2} \sum_{k=1}^{n-1} C_{n-1}^k E_{n-k-1}(x) \Big) = n E_{n-1}(x) .$$

5.3 - On a également

$$\boxed{E_n(x+1) + E_n(x) = 2 x^n}$$

Cette relation est évidente pour n = 0 . Supposons-la vraie jusqu'à l'ordre (n-1). Alors,

$$E_n(x+1) = (1+x)^n - \tfrac{1}{2} \sum_{k=1}^{n} C_n^k E_{n-k}(x+1)$$

$$= (1+x)^n + \tfrac{1}{2} \sum_{k=1}^{n} C_n^k E_{n-k}(x) - \sum_{k=1}^{n} C_n^k x^{n-k}$$

$$= x^n + \tfrac{1}{2} \sum_{k=1}^{n} C_n^k E_{n-k}(x).$$

Ajoutant la valeur

$$E_n(x) = x^n - \tfrac{1}{2} \sum_{k=1}^{n} C_n^k E_{n-k}(x) ,$$

on obtient visiblement la relation annoncée.

En corollaire, on a toujours, pour $n \neq 0$,

$$E_n(1) = - E_n(0) .$$

5.4 - Les nombres d'Euler d'ordre impair sont nuls

En effet, si l'on fait $x = \frac{1}{2}$ dans la relation de définition des polynômes d Euler, on trouve

$$\sum_{n=0}^{\infty} \frac{E_n}{n!} t^n = \frac{2\, e^{t/2}}{e^t + 1} = \frac{2}{e^{t/2} + e^{-t/2}} = \frac{1}{\operatorname{ch} \frac{t}{2}} ,$$

fonction paire dont le développement ne peut contenir que des puissances paires de t.

5.5 - On a la relation

$$\boxed{E_n(1-x) = (-1)^n E_n(x)}$$

On note d'abord que pour $x = \frac{1}{2}$,

$$E_n = (-1)^n E_n .$$

D'autre part, l'assertion est évidente pour n = 0. Si elle est vraie pour $E_{n-1}(x)$, on a

$$E_n(1-x) = E_n(\tfrac{1}{2}) + \int_{\frac{1}{2}}^{1-x} n\, E_{n-1}(t)\, dt = \frac{E_n}{2^n} + \int_{\frac{1}{2}}^{x} n\, E_{n-1}(1-t)\, d(1-t)$$

$$= \frac{E_n}{2^n} + (-1)^n \int_{\frac{1}{2}}^{x} n\, E_{n-1}(t)\, dt$$

$$= (-1)^n \left[\frac{E_n}{2^n} + \int_{\frac{1}{2}}^{x} n\, E_{n-1}(t)\, dt \right] = (-1)^n E_n(x) .$$

5.6 - On a encore

$$\boxed{(-1)^{n+1} E_n(-x) = E_n(x) - 2\, x^n}$$

C'est une combinaison des deux relations précédentes:

$$E_n(-x+1) = - E_n(-x) + (-1)^n\, 2\, x^n = (-1)^n E_n(x) ,$$

d'où la relation annoncée.

5.7 - Les polynômes d'Euler admettent le développement de TAYLOR

$$\boxed{E_n(x+h) = \sum_{k=0}^{n} C_n^k \, E_n(x) \, h^{n-k}}$$

La démonstration est la même que pour les polynômes de Bernoulli. En remplaçant dans cette formule x par $\frac{1}{2}$ et h par $(x-\frac{1}{2})$, on obtient

$$E_n(x) = \sum_{k=0}^{n} C_n^k \, \frac{E_k}{2^k} \, (x - \tfrac{1}{2})^{n-k} \quad .$$

5.8 - Les polynômes d'Euler vérifient la relation

$$\boxed{\sum_{k=1}^{m} (-1)^{m-k} \, k^n = \frac{E_n(m+1) + (-1)^m \, E_n(0)}{2} \qquad m,n=1,2,\ldots}$$

En effet,

$$\sum_{k=1}^{m} (-1)^{m-k} \, k^n = \sum_{k=1}^{m} \frac{E_n(k+1) + E_n(k)}{2} \, (-1)^{m-k}$$

$$= \tfrac{1}{2} \sum_{k=1}^{m} (-1)^{m-k} \, E_n(k+1) + \tfrac{1}{2} \sum_{k=0}^{m-1} (-1)^{m-k+1} E_n(k+1)$$

$$= \frac{E_n(m+1) + (-1)^{m-1} \, E_n(1)}{2}$$

Comme

$$E_n(1) = 2.0^n - E_n(0) = - E_n(0) \quad ,$$

on a encore

$$(-1)^{m-1} \, E_n(1) = (-1)^m \, E_n(0) \quad ,$$

ce qui démontre l'assertion.

5.9 - <u>Les nombres d'Euler sont entiers</u>

C'est évidemment vrai pour E_o . Si c'est vrai jusqu'à l'ordre (n-1), on a

$$E_n = 2^n \, E_n(\tfrac{1}{2}) = 2^n \left\{ (\tfrac{1}{2})^n - \tfrac{1}{2} \sum_{k=1}^{n} C_n^k \, \frac{E_{n-k}}{2^{n-k}} \right\}$$

$$= 1 - \sum_{k=1}^{n} C_n^k \, 2^{k-1} \, E_{n-k} \, ,$$

avec, dans chaque terme, E_{n-k} entier, 2^{k-1} entier, C_n^k entier.

5.10 - Valeurs des polynômes d'Euler en 0 et 1

On a, pour x = 0,

$$\frac{2}{e^t + 1} = \sum_{n=0}^{\infty} \frac{E_n(0)}{n!} t^n \quad .$$

Par ailleurs,

$$\frac{1}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} t^{n-1} \quad ,$$

d'où

$$\frac{1}{e^t - 1} - \frac{1}{e^t + 1} = \frac{1}{t} + \sum_{n=0}^{\infty} \left(\frac{B_{n+1}}{(n+1)!} - \frac{E_n(0)}{2.n!} \right) t^n \quad ,$$

soit

$$\frac{2}{e^{2t} - 1} = \frac{1}{t} + \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{B_{n+1}}{n+1} - \frac{E_n(0)}{2} \right) t^n .$$

Le premier membre vaut encore

$$\frac{2}{e^{2t} - 1} = 2\left(\frac{1}{2t} + \sum_{n=0}^{\infty} \frac{B_{n+1}}{(n+1)!} (2t)^n \right) ,$$

d'où, par identification des coefficients des différentes puissances de t,

$$\frac{1}{(n+1)!} B_{n+1} = \frac{E_n(0)}{2.n!} + \frac{2^{n+1} B_{n+1}}{(n+1)!}$$

soit

$$\boxed{E_n(0) = -\frac{2}{n+1} \left(2^{n+1} - 1 \right) B_{n+1} \quad , \quad E_n(1) = E_n(0)}$$

En particulier, tous les polynômes d'Euler de degré pair 0 s'annulent à l'origine et en x=1.

5.11 - Développement des polynômes d'Euler en séries de Fourier sur]0, 1[.

Au lieu de la base classique $\{ e^{2i\pi kx} \}$, on peut tout aussi bien utiliser la base des $e^{i(2k+1)\pi x}$, soit

$$e^{i\pi x} , e^{3i\pi x} , e^{5i\pi x} , \ldots \quad ,$$

qui est orthogonale, puisque

$$\int_0^1 e^{i(2k+1)\pi x} \, e^{-i(2l+1)\pi x} \, dx = \int_0^1 e^{i(k-l)2\pi x} \, dx = 0 \quad \text{pour } k \neq l$$

et complète dans $L^2(\,]0,\,1[\,)$, car si

$$\int_0^1 f(x)\, e^{i(2k+1)\pi x}\, dx = 0$$

pour tout k, on a, par les séries de Fourier classiques,

$$f(x)\, e^{i\pi x} = 0 \quad \text{pp dans }]0,\,1[\ ,$$

soit

$$f(x) = 0 \text{ pp dans }]0,\,1[\ .$$

Cela étant, pour $E_o(x)$, on a

$$c_k^{(0)} = 2\int_0^1 E_o(x)\, e^{i(2k+1)\pi x}\, dx = \frac{2}{i(2k+1)\pi}\left[\, e^{i(2k+1)} - 1\,\right]$$

$$= \frac{4i}{(2k+1)\pi}\ ,$$

soit

$$a_k^{(0)} = 0\ , \qquad b_k^{(0)} = \frac{4}{(2k+1)\pi}\ ,$$

d'où

$$\boxed{E_o(x) = \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{\sin(2k+1)\pi x}{2k+1}}$$

Passons à $n>0$. On a

$$c_k^{(n)} = 2\int_0^1 E_n(x)\, e^{i(2k+1)\pi x}\, dx =$$

$$= \frac{2}{i(2k+1)\pi}\left[\, E_n(1)\, e^{i(2k+1)} - E_n(0)\,\right]$$

$$- \frac{2n}{i(2k+1)\pi}\int_0^1 E_{n-1}(x)\, e^{i(2k+1)\pi x}\, dx,$$

soit, comme $E_n(0) = -E_n(1)$,

$$c_k^{(n)} = -\frac{n}{i(2k+1)\pi}\, c_k^{(n-1)}\ .$$

De proche en proche, on obtient

$$c_k^{(n)} = \frac{(-1)^n\, n!}{i^n(2k+1)^n\pi^n}\, c_k^{(0)} = \frac{4\,(-1)^n\, n!}{i^{n-1}\,(2k+1)^{n+1}\,\pi^{n+1}}\ .$$

Dès lors, <u>pour n pair</u>, $n = 2m$ et $\dfrac{1}{i^{2m-1}} = (-1)^m\, i$,

d'où

$$c_k^{(2m)} = \frac{4\,(-1)^m\,(2m)!\;i}{(2k+1)^{2m+1}\,\pi^{2m+1}} = i\,b_k^{(2m)} \quad , \quad a_k^{(2m)} = 0$$

et

$$\boxed{E_{2m}(x) = \frac{(-1)^m\,4\,(2m)!}{\pi^{2m+1}} \sum_{k=0}^{\infty} \frac{\sin(2k+1)\pi x}{(2k+1)^{2m+1}}}$$

Pour n impair, $n = 2m-1$ et $\dfrac{(-1)^{2m-1}}{i^{2m-2}} = (-1)^m$,

d'où

$$c_k^{(2m-1)} = \frac{4\,(-1)^m\,(2m-1)!}{(2k+1)^{2m}\,\pi^{2m}} = a_k^{(2m-1)} \quad , \quad b_k^{(2m-1)} = 0$$

et

$$\boxed{E_{2m-1}(x) = \frac{(-1)^m\,4\,(2m-1)!}{\pi^{2m}} \sum_{k=0}^{\infty} \frac{\cos(2k+1)\pi x}{(2k+1)^{2m}}}$$

5.12 - Application aux séries de puissances inverses

Pour $n = 2m$ et $x = \frac{1}{2}$, on a

$$\sin(2k+1)\frac{\pi}{2} = (-1)^k$$

et

$$E_{2m}(\tfrac{1}{2}) = \frac{E_{2m}}{2^{2m}} = \frac{(-1)^m\,4\,(2m)!}{2m+1} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)^{2m+1}} \quad ,$$

soit

$$\boxed{\sum_{k=0}^{\infty} (-1)^k\,(2k+1)^{-(2m+1)} = (-1)^m\,\frac{(\pi/2)^{2m+1}}{2\,(2m)!}\,E_{2m}}$$

Pour $n = (2m+1)$, les cosinus du développement de $E_{2m-1}(x)$ ne prennent pas la valeur ± 1 simultanément, ce qui oblige à faire une transformation. A cette fin, observons que

$$\int_0^1 \frac{\cos(2k+1)\pi x + \cos(2k-1)\pi x}{\cos\ x}\,dx = 2\int_0^1 \cos\ 2k\pi x\,dx = 0,$$

d'où

$$\int_0^1 \frac{\cos(2k+1)\pi x}{\cos \pi x}\, dx = -\int_0^1 \frac{\cos(2k-1)\pi x}{\cos \pi x}\, dx \ .$$

Par applications successives de ce résultat, on obtient

$$\int_0^1 \frac{\cos(2k+1)\pi x}{\cos \pi x}\, dx = (-1)^k \int_0^1 \frac{\cos \pi x}{\cos \pi x}\, dx = (-1)^k \ ,$$

d'où

$$\int_0^1 \frac{E_{2m-1}(x)}{\cos \pi x}\, dx = \frac{(-1)^m\, 4\, (2m-1)!}{\pi^{2m}} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)^{2m}}$$

soit

$$\boxed{\sum_{k=0}^{\infty} (-1)^k\, (2k+1)^{-2m} = \frac{(-1)^m\, \pi^{2m}}{4(2m-1)!} \int_0^1 \frac{E_{2m-1}(x)}{\cos \pi x}\, dx}$$

ce qui ramène le calcul de la série à celui d'une intégrale. Comme $E_{2m-1}(x)$ s'annule en $x = \frac{1}{2}$, l'intégrand ne comporte pas de singularité, et on peut utiliser une formule de Gauss, à condition de noter, si la formule en question utilise le point central $x = \frac{1}{2}$, que

$$\lim_{x \to \frac{1}{2}} \frac{E_{2m-1}(x)}{\cos \pi x} = \lim_{x \to \frac{1}{2}} \frac{(2m-1)\, E_{2m-2}(x)}{-\pi \sin \pi x} = -\frac{(2m-1)\, E_{2m-2}}{2^{2m-2}\, \pi} \ ,$$

par application du théorème de l'HOSPITAL. Par ailleurs, on peut tenir compte de la symétrie de l'intégrand par rapport au centre de l'intervalle.

6. ENCADREMENT DES NOMBRES D'EULER ET DE BERNOULLI

6.1 - Introduisons les notations

$$\zeta(n) = \sum_{k=1}^{\infty} k^{-n} \qquad \text{(fonction } \zeta \text{ de RIEMANN)}$$

$$\eta(n) = \sum_{k=1}^{\infty} (-1)^{k-1}\, k^{-n}$$

$$\lambda(n) = \sum_{k=0}^{\infty} (2k+1)^{-n} \ .$$

Ces trois séries sont intimement liées pour $n > 1$ (cas de sommabilité), car

$$\zeta(n) = \sum_{k=0}^{\infty} (2k+1)^{n} + \sum_{k=1}^{\infty} (2k)^{-n} = \lambda(n) + 2^{-n}\zeta(n) \ ,$$

soit

$$\lambda(n) = (1 - 2^{-n})\ \zeta(n)\ ;$$

de plus,

$$\eta(n) = \sum_{k=0}^{\infty} (2k+1)^{-n} - \sum_{k=1}^{\infty} (2k)^{-n} = \lambda(n) - 2^{-n}\ \zeta(n)\ ,$$

d'où

$$\eta(n) = (1 - 2^{1-n})\ \zeta(n)\ .$$

6.2 - Les relations précédentes permettent d'encadrer les nombres de Bernoulli. En effet,

$$(-1)^{m+1}\ B_{2m} = \frac{2\ (2m)!}{(2\ )^{2m}}\ \zeta(2m)\ .$$

On a, d'une part,

$$\zeta(2m) = 1 + \frac{1}{2^{2m}} + \ldots \quad > 1$$

et, d'autre part,

$$\zeta(2m) = \frac{1}{1 - 2^{1-2m}}\quad \eta(2m) = \frac{1}{1 - 2^{1-2m}}\ (1 - \frac{1}{2^{2m}}\ ) \leqslant \frac{1}{1 - 2^{1-2m}}\ ,$$

par la majoration classique du reste d'une série alternée. Dès lors,

$$\boxed{\frac{2\ (2m)!}{(2\pi)^{2m}} < (-1)^{m+1}\ B_{2m} < \frac{1}{1 - 2^{1-2m}}\ \frac{2\ (2m)!}{(2\pi)^{2m}}}$$

En particulier, <u>pour m très grand</u>, on a

$$B_{2m} \approx \frac{(-1)^{m+1}\ 2\ (2m)!}{(2\pi)^{2m}}$$

en ce sens que la limite du rapport entre les deux membres vaut 1.

6.3 - Pour les nombres d'Euler, on a directement

$$(-1)^{m}\ E_{2m} = \frac{2\ (2m)!}{(\pi/2)^{2m+1}}\ \beta(2m+1)$$

en utilisant la notation

$$\beta(n) = \sum_{k=0}^{\infty} (-1)^{k}\ (2k+1)^{n}\ .$$

Il est clair que

$$1 - 3^{-n} < \beta(n) = 1 - 3^{-n} + 5^{-n} - \ldots < 1 ,$$

d'où

$$\boxed{(1 - 3^{-1-2m}) \frac{2\,(2m)!}{(\pi/2)^{2m+1}} < (-1)^m E_{2m} < \frac{2\,(2m)!}{(\pi/2)^{2m+1}}}$$

et, pour m très grand,

$$E_{2m} \approx \frac{(-1)^m\, 2\,(2m)!}{(\pi/2)^{2m+1}}$$

7. SOMMES DES PUISSANCES INVERSES

7.1 - Considérons, dans le cas général, les séries

$$\zeta(p) = \sum_{n=1}^{\infty} n^{-p} \quad , \quad p > 1$$

$$\eta(p) = \sum_{n=1}^{\infty} (-1)^{n+1} n^{-p} \quad , \quad p > 0$$

$$\lambda(p) = \sum_{n=0}^{\infty} (2n+1)^{-p} \quad , \quad p > 1$$

$$\beta(p) = \sum_{n=0}^{\infty} (-1)^n (2n+1)^{-p} \quad , \quad p > 0 .$$

Nous avons déjà vu que, pour $p > 1$, les trois premières séries sont liées. Leur calcul pour p entier a déjà été envisagé. Dans les autres cas, il est également possible de ramener ces séries à des intégrales.

a) Montrons que l'on a

$$\zeta(p) = \frac{1}{\Gamma(p)} \int_0^{\infty} \frac{x^{p-1}}{e^x - 1}\, dx.$$

Calculons

$$I = \int_0^{\infty} \frac{x^{p-1}}{e^x - 1}\, dx = \int_0^{\infty} \frac{x^{p-1}\, e^{-x}}{1 - e^{-x}}\, dx = \int_0^{\infty} x^{p-1}\, e^{-x} \sum_{n=0}^{\infty} e^{-nx}\, dx.$$

Les fonctions

$$f_N(x) = x^{p-1}\, e^{-x} \sum_{n=0}^{N} e^{-nx}$$

sont intégrables et vérifient

$$|f_N| \leq \frac{x^{p-1}\, e^{-x}}{1 - e^{-x}} \in L^1(\,]0, \infty[\,) \quad \text{pour} \quad p > 1 ,$$

ce qui permet d'affirmer, en vertu du théorème de LEBESGUE, que

$$I = \lim_{N \to \infty} \sum_{n=0}^{N} \int_0^\infty e^{-(n+1)x}\, x^{p-1}\, dx = \sum_{n=0}^{\infty} \int_0^\infty e^{-(n+1)x}\, x^{p-1}\, dx$$

$$= \sum_{n=0} \frac{1}{(n+1)^p} \int_0^\infty e^{-y}\, y^{p-1}\, dy = \Gamma(p) \sum_{n=0} \frac{1}{(n+1)^p} .$$

b) Démontrons la relation

$$\eta(p) = \frac{1}{\Gamma(p)} \int_0^\infty \frac{x^{p-1}}{e^x + 1}\, dx .$$

On a

$$I = \int_0^\infty \frac{x^{p-1}}{e^x + 1}\, dx = \int_0^\infty e^{-x}\, x^{p-1} \sum_{n=0}^{\infty} (-1)^n\, e^{-nx}\, dx.$$

Les fonctions

$$f_N(x) = e^{-x}\, x^{p-1} \sum_{n=0}^{N} (-1)^n\, e^{-nx}$$

sont intégrables, vérifient (reste des séries altermées)

$$|f_N| \leq x^{p-1}\, e^{-x} \in L^1(\,[0, \infty[\,) \text{ pour } \quad p > 0,$$

ce qui permet, par le théorème de LEBESGUE, d'écrire

$$I = \lim_{N \to \infty} \sum_{n=0}^{N} \int_0^\infty e^{-(n+1)x}\, x^{p-1}\, (-1)^n\, dx$$

$$= \sum_{n=0}^{\infty} \int_0^\infty e^{-(n+1)x}\, x^{p-1}\, (-1)^n\, dx$$

$$= \sum_{n=0}^{\infty} \int_0^\infty e^{-(n+1)x}\, x^{p-1}\, (-1)^n\, dx$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)^p} \int_0^\infty e^{-y}\, y^{p-1}\, dy = \Gamma(p) \sum_{n=0}^{\infty} \frac{1}{(n+1)^p}$$

c) On a d'autre part

$$\lambda(p) = \frac{1}{2\,\Gamma(p)} \int_0^\infty \frac{x^{p-1}}{\operatorname{sh} x}\, dx .$$

En effet,

$$\int_0^\infty \frac{x^{p-1}}{\operatorname{sh} x}\,dx = 2\int_0^\infty \frac{x^{p-1}}{e^x - e^{-x}}\,dx = 2\int_0^\infty \frac{x^{p-1}\,e^{-x}}{1 - e^{-2x}}\,dx$$

$$= 2\int_0^\infty e^{-x}\,x^{p-1}\sum_{n=0}^{\infty} e^{-2nx}\,dx = 2\sum_{n=0}^{\infty}\int_0^\infty e^{-(2n+1)x}\,x^{p-1}\,dx$$

$$= 2\sum_{n=0}^{\infty}\frac{1}{(2n+1)^p}\int_0^\infty e^{-y}\,y^{p-1}\,dy = 2\Gamma(p)\sum_{n=0}^{\infty}\frac{1}{(2n+1)^p}\ ,$$

les justifications se faisant comme en a).

d) On a enfin

$$\beta(p) = \frac{1}{2\Gamma(p)}\int_0^\infty \frac{x^{p-1}}{\operatorname{ch} x}\,dx\ ,$$

car

$$\int_0^\infty \frac{x^{p-1}}{\operatorname{ch} x}\,dx = 2\int_0^\infty \frac{x^{p-1}}{e^x + e^{-x}}\,dx = 2\int_0^\infty \frac{x^{p-1}\,e^{-x}}{1 + e^{-2x}}\,dx$$

$$= 2\int_0^\infty e^{-x}\,x^{p-1}\sum_{n=0}^{\infty} e^{-2nx}\,(-1)^n\,dx =$$

$$= 2\sum_{n=0}^{\infty}\int_0^\infty e^{-(2n+1)x}\,(-1)^n\,x^{p-1}\,dx$$

$$= 2\sum_{n=0}^{\infty}(-1)^n\frac{1}{(2n+1)^p}\int_0^\infty e^{-y}\,y^{p-1}\,dy = 2\Gamma(p)\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)^p}\ ,$$

les justifications se faisant comme en b).

7.2 - On peut calculer les intégrales en question, à savoir

$$\zeta(p) = \frac{1}{\Gamma(p)}\int_0^\infty e^{-x}\left(\frac{x^{p-1}}{1 - e^{-x}}\right)dx\ ,$$

$$\eta(p) = \frac{1}{\Gamma(p)}\int_0^\infty e^{-x}\left(\frac{x^{p-1}}{1 + e^{-x}}\right)dx\ ,$$

$$\lambda(p) = \frac{1}{\Gamma(p)}\int_0^\infty e^{-x}\left(\frac{x^{p-1}}{1 - e^{-2x}}\right)dx\ ,$$

$$\beta(p) = \frac{1}{\Gamma(p)}\int_0^\infty e^{-x}\left(\frac{x^{p-1}}{1 + e^{-2x}}\right)dx$$

par les formules de LAGUERRE, pour autant que les fonctions entre

parenthèses soient continues, ce qui a lieu

$$\begin{cases} - \text{ si } p \geqslant 2 \quad \text{pour} \quad \zeta \text{ et } \lambda \\ - \text{ si } p \geqslant 1 \quad \text{pour} \quad \eta \text{ et } \beta. \end{cases}$$

Cependant, la convergence est généralement médiocre, en raison du comportement des fonctions à l'origine, assez mal approché par des polynômes. Il est plus judicieux de scinder les intégrales comme suit:

a) On a

$$\Gamma(p)\,\zeta(p) = \int_0^1 \frac{x^{p-1}}{e^x - 1}\,dx + \int_1^\infty e^{-x}\left(\frac{x^{p-1}}{1 - e^{-x}}\right) dx = I_1 + I_2 ,$$

avec

$$I_2 = \int_0^\infty e^{-(y+1)}\,\frac{(1+y)^{p-1}}{1 - e^{-(y+1)}}\,dy = \frac{1}{e}\int_0^\infty e^{-y}\,\frac{(1+y)^{p-1}}{1 - e^{-(y+1)}}\,dy ,$$

qui se calcule assez bien par les formules de LAGUERRE. Pour I_1, on note que

$$\frac{1}{e^x - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!}\,x^{n-1} ,$$

d'où

$$I_1 = \int_0^1 \sum_{n=0}^{\infty} \frac{B_n}{n!}\,x^{n+p-2}\,dx .$$

Il résulte des encadrements que les B_n non nuls sont alternés. Dès lors, les fonctions

$$f_N(x) = \sum_{n=0}^{N} \frac{B_n}{n!}\,x^{n+p-2}$$

vérifient, pour $0 \leqslant x \leqslant 1$, la relation

$$|f_N(x)| \leqslant B_o\,x^{p-2} \in L^1([0, 1]) \quad \text{pour } p > 1 ,$$

ce qui permet, par le théorème de LEBESGUE, de permuter la série et l'intégrale:

$$I_1 = \sum_{n=0}^{\infty} \frac{B_n}{n!}\int_0^1 x^{n+p-2}\,dx = \sum_{n=0}^{\infty} \frac{B_n}{n!\,(n+p-1)}$$

Cette série converge vite, car pour r assez grand,

$$\left| \frac{B_{2r}}{(2r)!\ (p+2r-1)} \right| \approx \left| \frac{(-1)^{r+1}.\ 2}{(2\ )^{2r}\ (p+2r-1)} \right| \leqslant \frac{1}{(2\pi)^{2r}} \leqslant 10^{-K}$$

pour

$$2r \geqslant \frac{K}{\log(2\pi)} = 1{,}25\ K\ .$$

Pour $K = 10$, on obtient $2r = 12{,}5 < 14$, soit $r+2 = 9$ termes à sommer.

b) Ecrivons

$$\Gamma(p)\,\eta(p) = \int_0^{\frac{1}{2}} \frac{x^{p-1}}{e^x + 1}\,dx + \int_{\frac{1}{2}}^{\infty} e^{-x}\,\frac{x^{p-1}}{1 + e^{-x}}\,dx = I_1 + I_2\ ,$$

avec

$$I_2 = \frac{1}{\sqrt{e}} \int_0^{\infty} e^{-y}\ \frac{(y+\frac{1}{2})^{p-1}}{1 + e^{-(y+\frac{1}{2})}}\,dy\ ,$$

calculable par les formules de LAGUERRE. Pour I_1, on note que

$$\frac{1}{e^x + 1} = \sum_{n=0}^{\infty} E_n(0)\,\frac{x^n}{n!} = -\sum_{n=0}^{\infty} \frac{2}{n+1}\,(2^{n+1} - 1)\,B_{n+1}\,\frac{x^n}{n!}\ ,$$

d'où

$$I_1 = -\sum_{n=0}^{\infty} \frac{2}{n+1}\,(2^{n+1} - 1)\,B_{n+1}\ \frac{1}{n!}\ \frac{1}{2^{n+p}} =$$

$$= -\frac{1}{2^{p-2}} \sum_{n=0}^{\infty} \left(1 - \frac{1}{2^{n+1}}\right) \frac{B_{n+1}}{(n+1)!}\ ,$$

la justification de la permutation de la série et de l'intégrale se faisant comme ci-dessus, pour $p > 0$. Après $(r+2)$ termes, le reste est inférieur à

$$\left| \frac{B_{2r}}{(2r)!} \right| \approx \frac{2}{(2\pi)^{2r}}\ ,$$

ce qui assure une convergence très rapide.

c) $\lambda(p)$ se calcule à partir de $\zeta(p)$.

d) Pour $\beta(p)$, on effectue la décomposition

$$\Gamma(p)\,\beta(p) = \int_0^{\frac{1}{2}} \frac{x^{p-1}\,e^{x}}{e^{2x}+1}\,dx + \int_{\frac{1}{2}}^{\infty} e^{-x}\left(\frac{x^{p-1}}{1+e^{-2x}}\right)dx = I_1 + I_2,$$

avec

$$I_2 = \frac{1}{\sqrt{e}} \int_0^{\infty} e^{-y}\,\frac{(y+\frac{1}{2})^{p-1}}{1+e^{-(2y+1)}}\,dy\ .$$

Posant dans I_1, $x = y/2$, on obtient

$$I_1 = 2^{-p-1}\int_0^1 \frac{2\,y^{p-1}\,e^{y/2}}{e^{y}+1}\,dy = 2^{-p-1}\int_0^1 \sum_{n=0}^{\infty} \frac{E_n}{2^n}\,\frac{y^{n+p-1}}{n!}\,dy$$

$$= 2^{-p-1}\sum_{n=0}^{\infty} \frac{E_n}{2^n}\,\frac{1}{n!\,(n+p)}\ ,$$

la permutation de la série et de l'intégrale se justifiant par le théorème de LEBESGUE. En ce qui concerne la convergence, la série est alternée et son terme général vérifie, pour n suffisant,

$$\left|\frac{E_{2r}}{2^{2r}\,(2r)!\,(2r+n)}\right| \approx \left|\frac{2}{(2^{2r})\,(\pi/2)^{2r+1}\,(2r+n)}\right| \leq \frac{1}{(2\pi)^{2r+1}}\ ,$$

ce qui garantit une convergence rapide.

7.3 - <u>Exemple</u> - Calculons par cette méthode (1). On a d'abord

$$I_1 = \frac{1}{2^{p+1}} \sum_{n=0}^{\infty} \frac{E_n}{2^n}\,\frac{1}{n!\,(n+p)} = \frac{1}{4}\sum_{n=0}^{\infty} \frac{E_n}{2^n\,n!\,(n+1)}$$

r	E_{2r}	$\frac{E_{2r}}{2^{2r}\,(2r)!\,(2r+1)}$
0	1	1
1	-1	-0,041 666 667
2	5	0,002 604 167
3	-61	-0,000 189 112
4	1385	0,000 014 909
5	50 521	-0,000 001 236
6	2 702 765	0,000 000 106
7	-199 360 981	-0,000 000 009
8	19 391 512 145	0,000 000 000
		Σ = 0,960 762 158

$$I_1 = \Sigma/4 = 0{,}240\ 190\ 540\ .$$

Calculons I_2 par la formule de LAGUERRE à 8 points:

$$I_2 = \frac{1}{\sqrt{e}} \int_0^{\infty} e^{-y} \frac{1}{1 + e^{-(2y+1)}} \, dy \approx 0{,}545\ 190\ 768 \ .$$

Il vient

$$\beta(1) = \frac{1}{\Gamma(1)} (I_1 + I_2) = 0{,}785\ 381\ 308 \ .$$

La réponse exacte est

$$\beta(1) = 0{,}785\ 381\ 634 \ .$$

L'erreur est donc inférieure à 3.10^{-6} . On améliorerait les choses en utilisant une formule d'intégration de degré plus élevé. Pour obtenir la même précision par sommation directe, il eût fallu environ 170 000 termes.

8. FORMULE D'EULER-MAC LAURIN

Il s'agit à l'origine d'une formule d'intégration numérique, mais elle peut rendre bien des services dans la sommation des séries. On a, comme $B_n'(x) = n\, B_{n-1}(x)$,

$$\int_0^1 f(x)\,dx = \int_0^1 f(x)\, B_0(x)\,dx = \left[\frac{B_1(x)}{1} f(x) \right]_0^1 - \int_0^1 B_1(x)\, f'(x)\,dx$$

$$= \tfrac{1}{2}\left(f(0) + f(1) \right) - \frac{1}{2!} \int_0^1 B_2'(x)\, f'(x)\,dx$$

$$= \tfrac{1}{2}(f(0) + f(1)) - \frac{1}{2!} B_2 (f'(1) - f'(0)) + \frac{1}{3!} \int_0^1 B_3'(x)\, f''(x)\,dx \ .$$

De proche en proche, on obtient, en s'arrêtant après 2n intégrations par parties, et en se souvenant que les B_k impairs, $k > 1$, sont nuls,

$$\int_0^1 f(x)\,dx = \tfrac{1}{2}(f(0) + f(1)) - \sum_{k=1}^{n} \frac{B_{2k}}{(2k)!} \left(f^{(2k-1)}(1) - f^{(2k-1)}(0)\right)$$

$$+ R \ ,$$

avec

$$R = \frac{1}{(2n+1)!} \int_0^1 B_{2n+1}'(x)\, f^{(2n)}(x) = - \frac{1}{(2n+2)!} \int_0^1 B_{2n+2}'(x)\, f^{(2n+1)}(x)\,dx \ .$$

En choisissant pour $B_{2n+2}'(x)$, la primitive $B_{2n+2}(x) - B_{2n+2}$,

on obtient encore

$$R = \frac{1}{(2n+2)!} \int_0^1 (B_{2n+2}(x) - B_{2n+2})\, f^{(2n+2)}(x)\, dx \ .$$

Montrons que $B_{2n+2}(x) - B_{2n+2}$ est <u>de signe constant</u>. Si ce polynôme s'annulait dans $]0, 1[$, il s'annulerait donc en trois points de $[0, 1]$. Alors, par le lemme de Rolle, sa dérivée (à un facteur près) B_{2n+1} s'annulerait en deux points au moins de $]0, 1[$, ainsi qu'en 0 et 1, soit en 4 points de $[0, 1]$. $B_{2n}(x)$ s'annulerait en 3 points de $]0, 1[$; $B_{2n-1}(x)$, en deux points de $]0, 1[$, plus 0 et 1, soit 4 points de $[0, 1]$. Les $B_k(x)$, k pair, s'annuleraient donc en 3 points au moins de $[0, 1]$, et les $B_k(x)$, k impair, en 4 points. Descendant ainsi jusqu'à $B_1(x)$, ce dernier devrait s'annuler en deux points de $]0, 1[$, ce qui entraînerait $B_1(x) = 0$, puisque c'est une fonction affine. On arriverait donc à une contradiction, car $B_1(x)$ n'est pas identiquement nul. Donc, $B_{2n+2}(x) - B_{2n+2}$ a son signe constant. Ceci permet d'écrire, par le théorème de la moyenne,

$$R = \frac{1}{(2n+2)!} f^{(2n+2)}(\xi) \int_0^1 (B_{2n+2}(x) - B_{2n+2})\, dx \quad , \quad \xi \in]0, 1[\ ,$$

$$= - \frac{B_{2n+2}}{(2n+2)!} f^{(2n+2)}(\xi) \quad ,$$

car

$$\int_0^1 B_{2n+2}(x)\, dx = \frac{1}{2n+3} \int_0^1 B'_{2n+3}(x)\, dx = \frac{1}{2n+3} (B_{2n+3} - B_{2n+3}) = 0 \ .$$

Appliquons ce résultat à N intervalles successifs: il vient

$$\int_0^N f(x)\, dx + \tfrac{1}{2}(f(0) + f(N)) = \sum_{k=0}^{N} f(k) -$$

$$- \sum_{r=1}^{n} \frac{B_{2r}}{(2r)!} (f^{(2r-1)}(N) - f^{(2r-1)}(0)) + R \ ,$$

avec

$$R = \frac{1}{(2n+2)!} \int_0^1 [\, B_{2n+2}(x) - B_{2n+2} \,] \sum_{k=0}^{N-1} f^{(2n+2)}(x+k)\, dx$$

$$= - \frac{B_{2n+2}}{(2n+2)!} \sum_{k=0}^{N-1} f^{(2n+2)}(k + \xi) \quad , \quad \xi \in]0, 1[\ .$$

C'est la <u>formule d'EULER - MAC LAURIN</u>. Nous l'utiliserons sous la forme inverse:

$$\sum_{k=0}^{N} f(k) = \int_0^N f(x)\,dx + \tfrac{1}{2}(f(0) + f(N)) + \sum_{r=1}^{n} \frac{B_{2r}}{(2r)!}\left(f^{(2r-1)}(0) - f^{(2r-1)}(N)\right) - \frac{B_{2n+2}}{(2n+2)!} \sum_{k=0}^{N-1} f^{(2n+2)}(k+\xi) \quad , \quad \xi \in]0,1[$$

8.2 - Soit f une fonction intégrable avec ses dérivées jusqu'à l'ordre (2n+2), sur $[0,\infty[$. On a

$$\sum_{k=0}^{\infty} f(k) = \int_0^{\infty} f(x)\,dx + \tfrac{1}{2} f(0) - \sum_{r=1}^{n} \frac{B_{2r}}{(2r)!} f^{(2r-1)}(0) - R$$

avec

$$R = \frac{1}{(2n+2)!} \int_0^1 \left[B_{2n+2}(x) - B_{2n+2} \right] \sum_{k=0}^{\infty} f^{(2n+2)}(x+k)\,dx \ .$$

Une majoration simple de ce reste s'appuie sur le fait que

$$|B_{2n+2}(x)| < |B_{2n+2}| \quad , \quad 0 < x < 1 \ ,$$

inégalité qui découle directement du développement en série de Fourier de $B_{2n+2}(x)$. Il en résulte

$$|R| \leq \frac{2}{(2n+2)!} |B_{2n+2}| \int_0^{\infty} |f^{(2n+2)}(x)|\,dx \ .$$

Par ailleurs, de l'autre forme du reste,

$$-R = \frac{B_{2n+2}}{(2n+2)!} \sum_{k=0}^{\infty} f^{(2n+2)}(k+\xi) \ , \qquad \xi \in]0,1[\ ,$$

on déduit, pour autant que $f^{(2n+2)}$ soit de signe constant et de module décroissant pour x croissant,

$$|R| \geq \left| \frac{B_{2n+2}}{(2n+2)!} \int_1^{\infty} f^{(2n+2)}(x)\,dx \right| \ .$$

8.3 - <u>Exemple</u> - Soit à calculer $\sum_{k=1}^{\infty} \frac{1}{k^2}$. On commence par sommer directement un certain nombre de termes, par exemple, 9. On a alors

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \sum_{k=1}^{9} \frac{1}{k^2} + \sum_{k=0}^{\infty} \frac{1}{(k+10)^2} \quad ,$$

et la dernière série correspond à la fonction

$$f(x) = (x + 10)^{-2} \quad ,$$

dont les dérivées valent

$$f^{(m)}(x) = (-2)\ldots(-m-1)(x + 10)^{-m-2} = (-1)^m (m + 1)! \; (x + 10)^{-m-2} \; .$$

On a

$$\int_0^{\infty} \frac{dx}{(x+10)^2} = \frac{1}{10} \quad ,$$

d'où

$$\sum_{k=0}^{\infty} \frac{1}{(k+10)^2} = \frac{1}{10} + \frac{1}{200} - \sum_{r=1}^{n} \frac{B_{2r}}{(2r)!} \cdot \frac{(-1)^{2r-1} \cdot (2r)!}{10^{2r+1}} + R$$

$$= \frac{1}{10} + \frac{1}{200} + \sum_{r=1}^{n} \frac{B_{2r}}{10^{2r+1}} + R \; .$$

En s'arrêtant à n = 5, on trouve

$$\sum_{k=0}^{\infty} \frac{1}{(k+10)^2} = \frac{1}{10} + \frac{1}{200} + \frac{10^{-3}}{6} - \frac{10^{-5}}{30} + \frac{10^{-7}}{42} - \frac{10^{-9}}{30} + R \quad ,$$

avec un reste inférieur à

$$\left| 2 \cdot \frac{5}{66} \cdot \int_{10}^{\infty} x^{-12} \, dx \right| = (5/33) \cdot 10^{-11} = 1{,}52 \cdot 10^{-12}$$

On obtient ainsi, avec 10 chiffres significatifs,

$$\sum_{k=1}^{9} k^{-2} = 1{,}539\ 767\ 731$$

$$\sum_{k=10}^{18} k^{-2} = 0{,}105\ 166\ 336$$

$$\Sigma = 1{,}644\ 934\ 067 + \varepsilon \quad ,$$

avec

$$|\varepsilon| \leq 1{,}52 \cdot 10^{-12} \quad ,$$

valeur inférieure à l'erreur de troncature.

8.4 - La tentation est grande de pousser le développement

contenant les nombres de Bernoulli aussi loin que possible. Mais ce développement converge-t-il pour n tendant vers l'infini? Reprenons l'exemple ci-dessus. On a

$$|R| \geqslant \left| \frac{B_{2n+2}}{(2n+2)!} \int_1^{\infty} f^{(2n+2)}(x)\, dx \right| \geqslant \frac{1}{1 - 2^{-n-1}} \frac{2}{(2\pi)^{2n+2}} |f^{(2n+1)}(1)|$$

avec

$$| f^{(2n+1)}(1)| = (2n+2)!\ 11^{-2n-3}$$

Il n'y a donc pas convergence. Cependant, il peut se faire que le reste devienne très petit, comme c'était le cas dans l'exemple précédent.

9. METHODE DE STIRLING

Considérons les séries

$$S^{(m)} = \sum_{n=1}^{\infty} \frac{1}{n(n+1).....(n+m)} = \sum_{n=1}^{\infty} a_n^{(m)} .$$

On détermine aisément leur somme en notant que

$$a_n^{(m-1)} - a_{n+1}^{(m-1)} = \frac{1}{n(n+1)...(n+m-1)} - \frac{1}{(n+1)...(n+m)}$$

$$= \frac{m}{n(n+1)...(n+m)} = m\, a_n^{(m)} .$$

Il en découle

$$\sum_{n=1}^{N} a_n^{(m)} = \frac{1}{m} \left| a_o^{(m-1)} - a_1^{(m-1)} + a_1^{(m-1)} - a_2^{(m-1)} + \ldots - a_{N+1}^{(m-1)} \right|$$

$$= \frac{1}{m} \left| a_o^{(m-1)} - a_{N+1}^{(m-1)} \right|$$

et, à la limite,

$$S^{(m)} = \lim_{N \to \infty} \sum_{n=1}^{N} a_n^{(m)} = \frac{1}{m} \cdot \frac{1}{1...m} = \frac{1}{m\, m!}$$

A partir de ces séries simples, on peut déduire la somme des séries de la forme

$$\sum_{n=1}^{\infty} a_n \quad , \quad a_n = \frac{\alpha_o n^p + \alpha_1 n^{p-1} + \ldots + \alpha_p}{\beta_o n^q + \beta_1 n^{q-1} + \ldots + \beta_q}$$

p, q entiers

avec $p \leqslant q-2$ (pour assurer la sommabilité) et α_o, $\beta_o > 0$.

A cette fin, on écrit

$$a_n = \frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \cdots + \frac{A_m}{n(n+1)\cdots(n+m)} + b_n^{(m)} \quad ,$$

de manière à garantir que

$$b_n^{(m)} = 0\left(\frac{1}{n^{2+m}}\right) \quad ,$$

ce qui aura lieu si

$$A_1 = \lim_{n \to \infty} n(n+1)\, a_n \quad ,$$

$$A_2 = \lim_{n \to \infty} n(n+1)(n+2)\left[a_n - \frac{A_1}{n(n+1)}\right]$$

$$\cdots\cdots\cdots\cdots$$

$$A_m = \lim_{n \to \infty} n(n+1)\cdots(n+m)\left[a_n - \frac{A_1}{n(n+1)} - \cdots - \frac{A_{m-1}}{n(n+1)\cdots(n+m-1)}\right]$$

On obtient ainsi

$$\boxed{\sum_{n=1}^{\infty} a_n = \frac{A_1}{1.1!} + \frac{A_2}{2.2!} + \cdots + \frac{A_m}{m.m!} + \sum_{n=0}^{\infty} b_n^{(m)}} \quad ,$$

où la série des $b_n^{(m)}$ converge nettement plus vite que la série originelle.

On peut aussi commencer par sommer la série de départ jusqu'au terme numéro p et écrire

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{p} a_n + A_1 \sum_{n=p+1}^{\infty} \frac{1}{n(n+1)} + \cdots + A_m \sum_{n=p+1}^{\infty} \frac{1}{n(n+1)\cdots(n+m)} + \sum_{n=p+1}^{\infty} b_n^{(m)}$$

On vérifie aisément que

$$\sum_{n=p+1}^{\infty} \frac{1}{n(n+1)\cdots(n+m)} = \frac{1}{m}\,\frac{1}{p(p+1)\cdots(p+m)} \quad ;$$

de plus,

$$\sum_{n=p+1}^{\infty} b_n^{(m)} = 0\left(\frac{1}{p^{m+1}}\right)$$

tend vers zéro lorsque m tend vers l'infini. Il en découle le

développement de STIRLING

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{p} a_n + \frac{A_1}{1(p+1)} + \frac{A_2}{2(p+1)(p+2)} + \ldots + \frac{A_m}{m(p+1)\ldots(p+m)} + \ldots$$

Exemple - Soit à calculer $\sum_{n=1}^{\infty} \frac{1}{n^3}$. On écrit

$$\frac{1}{n^3} = \frac{A_2}{n(n+1)(n+2)} + \frac{A_3}{n(n+1)(n+2)(n+3)} + b_n^{(3)} .$$

Il vient

$$A_2 = \lim \frac{n(n+1)(n+2)}{n^3} = 1$$

$$A_3 = \lim \; n(n+1)(n+2)(n+3) \left[\frac{1}{n^3} - \frac{1}{n(n+1)(n+2)} \right]$$

$$= \lim \; (n+1)(n+2)(n+3) \; \frac{n^2 + 3n + 2 - n^2}{n^2(n+1)(n+2)} = 3$$

et

$$b_n^{(3)} = \frac{1}{n^3} - \frac{1}{n(n+1)(n+2)} - \frac{3}{n(n+1)(n+2)(n+3)}$$

$$= \frac{11n + 6}{n^3(n+1)(n+2)(n+3)} ,$$

d'où

$$\sum_{n=1}^{\infty} \frac{1}{n^3} = \frac{1}{2.2!} + \frac{3}{3.3!} + \sum_{n=1}^{\infty} \frac{11n+6}{n^3(n+1)(n+2)(n+3)} .$$

Supposons que l'on désire une erreur inférieure à $\frac{1}{2}.10^{-6}$. Le reste de la série du second membre vérifie la majoration

$$R_N \leqslant 11 \sum_{n=N+1}^{\infty} \frac{n+1}{n^3(n+1)(n+2)(n+3)} \leqslant 11 \sum_{n=N+1} \frac{1}{n^5} < \frac{11}{4} . \frac{1}{N^4} .$$

Il faudra donc assurer

$$\frac{1}{N^4} \leqslant \frac{2}{11} . 10^{-6} ,$$

soit $N \geqslant 48{,}42$, ce qui conduit à prendre N = 49 termes. On trouve

$$\Sigma = \frac{1}{4} + \frac{1}{6} + 0{,}785\ 389\ 82 = 1{,}202\ 056\ 48 .$$

La vraie valeur est 1,202 056 903... .

10. TRANSFORMATION DE KUMMER

C'est une généralisation de la méthode de STIRLING. Soit à calculer une série sommable $\sum_{n=1}^{\infty} a_n$, et soit $\sum_{n=1}^{\infty} b_n$ une autre série sommable, de somme connue et telle que

$$\lim_{n \to \infty} \frac{a_n}{b_n} = q \begin{cases} \neq 0 \\ \neq \infty \end{cases} .$$

Alors, on a

$$\sum_{n=1}^{\infty} a_n = q \sum_{n=1}^{\infty} b_n + \sum_{n=1}^{\infty} (a_n - q\, b_n) ,$$

avec

$$\frac{a_n - q\, b_n}{a_n} \longrightarrow 0$$

pour $n \to \infty$, ce qui garantit une accélération de la convergence.

Exemple - Proposons-nous de calculer la constante d'EULER

$$\gamma = 1 + \sum_{k=2}^{\infty} \left[\frac{1}{k} + \ln \left(1 - \frac{1}{k}\right) \right] .$$

On a, par la série de TAYLOR,

$$\ln \left(1 - \frac{1}{k}\right) = - \frac{1}{k} - \frac{1}{2k^2} - \frac{1}{3k^3} - \frac{1}{4k^4} - \cdots ,$$

ce qui permet d'écrire

$$= 1 - \tfrac{1}{2} \sum_{k=2}^{\infty} \frac{1}{k^2} - \frac{1}{3} \sum_{k=2}^{\infty} \frac{1}{k^3} - \frac{1}{4} \sum_{k=2}^{\infty} \frac{1}{k^4} - \tfrac{1}{5} \sum_{k=2}^{\infty} \frac{1}{k^5}$$

$$+ \sum_{k=2}^{\infty} \left(\frac{1}{k} + \ln \frac{k-1}{k} + \frac{1}{2k^2} + \frac{1}{3k^3} + \frac{1}{4k^4} + \frac{1}{5k^5} \right) .$$

On a, avec 10 chiffres significatifs,

$$\sum_{k=1}^{\infty} k^{-2} = 1{,}644\ 934\ 067$$

$$\sum_{k=1}^{\infty} k^{-3} = 1{,}202\ 056\ 903$$

$$\sum_{k=1}^{\infty} k^{-4} = 1{,}082\ 323\ 234$$

$$\sum_{k=1}^{\infty} k^{-5} = 1{,}036\ 927\ 755 ,$$

valeurs que l'on peut calculer ou trouver dans des tables. Quant à la dernière série, elle se stabilise au neuvième chiffre après la virgule après 26 termes. On trouve

$$\Sigma^* = \sum_{k=2}^{\infty} \left(\frac{1}{k} + \ln \frac{k-1}{k} + \frac{1}{2k^2} + \frac{1}{3k^3} + \frac{1}{4k^4} + \frac{1}{5k^5} \right) = - 0{,}004\ 998\ 639 \ .$$

On a donc

$$\begin{aligned}
\Sigma^* &= -0{,}004\ 998\ 639 \\
-\frac{1}{5} \sum_{k=2}^{\infty} k^{-5} &= -0{,}007\ 385\ 551 \\
-\frac{1}{4} \sum_{k=2}^{\infty} k^{-4} &= -0{,}020\ 580\ 809 \\
-\frac{1}{3} \sum_{k=2}^{\infty} k^{-3} &= -0{,}067\ 352\ 301 \\
-\frac{1}{2} \sum_{k=2}^{\infty} k^{-2} &= -0{,}322\ 467\ 034 \\
1 &= 1{,}000\ 000\ 000 \\
\hline
\gamma &= 0{,}577\ 215\ 668
\end{aligned}$$

La valeur exacte est 0,577 215 664 90...

11. TRANSFORMATION D'EULER - ABEL

11.1-Cette transformation s'applique aux séries de puissances de la forme

$$f(x) = \sum_{n=0}^{\infty} a_n x^n \ .$$

Il est peut-être utile de rappeler que le rayon de convergence ρ de ces séries admet l'expression générale

$$\boxed{\frac{1}{\rho} = \limsup_{n \to \infty} \sqrt[n]{|a_n|}}$$

Considérons en effet la valeur de ρ définie ci-dessus. Commençons par montrer que si $|x| > \rho$, la série diverge. On peut trouver deux nombres λ_1 et λ_2 tels que

$$|x| > \lambda_1 > \lambda_2 > \rho \quad ;$$

Quel que soit $N > 0$, on peut rouver $n \geqslant N$ tel que

$$|a_n|^{1/n} \geqslant \frac{1}{\lambda_2} \quad ,$$

ce qui implique

$$|a_n x^n| \geqslant \left(\frac{\lambda_1}{\lambda_2}\right)^n > 1 \quad ,$$

et la série diverge, car son terme général ne tend pas vers zéro.

Montrons à présent que pour $|x| < \rho$, la série converge. On peut trouver deux nombres λ_1 et λ_2 tels que

$$|x| < \lambda_1 < \lambda_2 < \rho \quad .$$

A partir d'une certaine valeur $N(\lambda_2)$, on a, si $n \geqslant N(\lambda_2)$,

$$\sup_{n \geqslant N} \left(|a_n|^{1/n} \right) \leqslant \frac{1}{\lambda_2} \quad .$$

Alors,

$$|a_n| \, |x|^n \leqslant \left| \frac{\lambda_1}{\lambda_2} \right|^n \quad ,$$

terme général d'une série convergente.

11.2-<u>Nous supposerons dans la suite que le rayon de convergence vaut 1</u>. Ce n'est pas une réelle restriction, car dans les autres cas, la série

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0} \left(\frac{a_n}{\rho^n} \right) y^n$$

avec

$$y = \rho x \quad ,$$

et la série du second membre a un rayon de convergence égal à 1.

11.3 - On a évidemment

$$f(x) = a_o + x \, \phi(x) \quad ,$$

avec

$$\phi(x) = \sum_{n=1}^{\infty} a_n x^{n-1} = \sum_{n=0}^{\infty} a_{n+1} x^n \quad .$$

Par ailleurs,

$$(1 - x) \, \phi(x) = \sum_{n=0}^{\infty} a_{n+1} x^n - \sum_{n=0}^{\infty} a_{n+1} x^{n+1}$$

$$= \sum_{n=0}^{\infty} a_{n+1} x^n - \sum_{n=1}^{\infty} a_n x^n$$

$$= a_o + \sum_{n=0}^{\infty} (a_{n+1} - a_n) x^n = a_o + \sum_{n=0}^{\infty} \Delta a_n x^n$$

en définissant les différences non divisées

$$\Delta a_n = a_{n+1} - a_n \ .$$

Il en découle

$$f(x) = a_o + x\,\phi(x) = a_o + \frac{x}{1-x}\,a_o + \frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n\,x^n \ ,$$

soit

$$\boxed{\sum_{n=0}^{\infty} a_n\,x^n = \frac{a_o}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty}\Delta a_n\,x^n}$$

C'est la transformation d'EULER - ABEL. On peut l'appliquer plusieurs fois:

$$\sum_{n=0}^{\infty}\Delta a_n\,x^n = \frac{a_o}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty}\Delta^2 a_n\,x^n \ ,$$

en introduisant les différences secondes

$$\Delta^2 a_n = \Delta a_{n+1} - \Delta a_n \ ,$$

et ainsi de suite:

$$\sum_{n=0}^{\infty}\Delta^k a_n\,x^n = \frac{a_o}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty}\Delta^{k+1} a_n\,x^n \ ,$$

avec

$$\Delta^{k+1} a_n = \Delta^k a_{n+1} - \Delta^k a_n \ ;$$

Il vient ainsi, en posant $\Delta^0 a_n = a_n$,

$$\boxed{\sum_{n=0}^{\infty} a_n\,x^n = \sum_{k=0}^{p-1}\frac{x^k}{(1-x)^{k+1}}\,\Delta^k a_o + \left(\frac{x}{1-x}\right)^p\sum_{n=0}^{\infty}\Delta^p a_n\,x^n}$$

Si les différences divisées $\Delta^p a_n$ tendent plus vite vers zéro que les a_n, la transformation apporte une accélération de la convergence. Cette circonstance est assez fréquente. Soit par exemple la série

$$S = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n} = \sum_{n=0}^{\infty}\frac{(-1)^n}{n+1} \ ,$$

qui est de la forme

$$\sum_{n=0}^{\infty}\frac{x^n}{n+1} \ ,$$

avec $x = -1$. Il est clair que

$$\Delta a_n = \frac{1}{n+2} - \frac{1}{n+1} = -\frac{1}{(n+1)(n+2)} \quad ,$$

$$\Delta^2 a_n = -\frac{1}{(n+2)(n+3)} + \frac{1}{(n+1)(n+2)} = \frac{2}{(n+1)(n+2)(n+3)} \quad ;$$

on a en général

$$\Delta^k a_n = \frac{(-1)^k\, k!}{(n+1)(n+2)\ldots(n+k+1)} \quad ,$$

car cette formule est vraie pour k = 1 et, si elle est vraie pour tout entier inférieur ou égal à (k-1), on a

$$\Delta^k a_n = \frac{(-1)^{k-1}(k-1)!}{(n+2)\ldots(n+k+1)} - \frac{(-1)^{k-1}(k-1)!}{(n+1)\ldots(n+k)}$$

$$= \frac{(-1)^{k-1}(k-1)!\,(n+1-n-k-1)}{(n+1)\ldots\ldots(n+k+1)} = \frac{(-1)^k\, k!}{(n+1)\ldots(n+k+1)} \quad .$$

On a en particulier

$$\Delta^k a_o = \frac{(-1)^k\, k!}{(k+1)!} = \frac{(-1)^k}{k+1} \quad ,$$

ce qui donne

$$S = \sum_{k=0}^{p-1} \frac{1}{(k+1)\,2^{k+1}} + \frac{p!}{2^p} \sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)\ldots(n+p+1)} \quad .$$

La dernière série est alternée, donc vérifie

$$\frac{p!}{2^p} \sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)\ldots(n+p+1)} \leqslant \frac{1}{p.2^p} \quad .$$

En poussant jusqu'à des valeurs suffisantes de p, cette série devient négligeable. On a donc

$$S = \sum_{k=0}^{\infty} \frac{1}{(k+1)\,2^{k+1}} = 0,693\ 147\ 180$$

valeur stabilisée après 26 termes. Comme

$$\ln(1-x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \quad ,$$

on a en fait

$$S = \ln 2 = 0,693\ 147\ 181 \ .$$

Pour obtenir un résultat à 10^{-9} près par sommation directe de la série originelle, il aurait fallu 10^9 termes.

11.4 - Une application intéressante de la transformation d'EULER-ABEL réside dans les séries du type

$$S(x) = \sum_{n=0}^{\infty} P_m(n)\, x^n \quad ,$$

où P_m est un polynôme de degré m. En effet, on a automatiquement

$$\Delta^{m+1} P_m = 0 \quad ,$$

si bien que

$$S(x) = \sum_{k=0}^{m} \frac{x^k}{(1-x)^{k+1}}\, \Delta^k P_m(0) \quad .$$

Exemple - Soit à calculer la somme de la série

$$S(x) = \sum_{n=0}^{\infty} (n^3 + n^2 + n + 1)\, x^n$$

On construit le tableau suivant:

n	P(n)	$\Delta P(n)$	$\Delta^2 P(n)$	$\Delta^3 P(n)$	$\Delta^4 P(n)$
0	1				
		3			
1	4		8		
		11		6	
2	15		14		0
		25		6	
3	40		20		
		45			
4	85				

et on obtient

$$S(x) = \frac{1}{1-x} + \frac{3x}{(1-x)^2} + \frac{8x^2}{(1-x)^3} + \frac{6x^3}{(1-x^4)} \qquad \text{pour } |x| < 1 \quad .$$

12. METHODE D'INTEGRATION DU DEVELOPPEMENT DE TAYLOR

12.1 - Cette méthode s'applique au calcul des séries du type

$$\sum_{n=1}^{\infty} f(n) \quad ,$$

avec f analytique. Partant du développement

$$f(x) = f(x_o) + \sum_{p=1}^{\infty} \frac{f^{(p)}(x_o)}{p!} (x - x_o)^p \quad ,$$

on obtient aisément

$$\int_{x_o-\frac{1}{2}}^{x_o+1/2} f(x)\,dx = f(x_o) + 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)!\ 2^{p+1}}\ f^{(p)}(x_o)$$

soit

$$f(x_o) = \int_{x_o-\frac{1}{2}}^{x_o+\frac{1}{2}} f(x)\,dx - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)!\ 2^{p+1}}\ f^{(p)}(x_o)\ .$$

On en déduit

$$\sum_{n=1}^{N} f(n) = \int_{\frac{1}{2}}^{N+\frac{1}{2}} f(x)\,dx - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)!\ 2^{p+1}} \sum_{n=1}^{N} f^{(p)}(n) \qquad (a)$$

et, en passant à la limite,

$$\sum_{n=1}^{\infty} f(n) = \int_{\frac{1}{2}}^{\infty} f(x)\,dx - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)!\ 2^{p+1}} \sum_{n=1}^{\infty} f^{(p)}(n) \qquad (b)$$

La soustraction de ces deux résultats donne une expression du reste:

$$\boxed{\sum_{n=N+1}^{\infty} f(n) = \int_{N+\frac{1}{2}}^{\infty} f(x)\,dx - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)!\ 2^{p+1}} \sum_{n=N+1}^{\infty} f^{(p)}(n)}$$

Cette formule est utile dans la mesure où les séries des dérivées convergent plus vite que les séries d'origine.

12.2 - <u>Sommation des séries de RIEMANN par récurrence rétrograde</u>

Proposons-nos de calculer les séries de RIEMANN

$$S_n = \sum_{k=1}^{\infty} k^{-n}$$

avec une erreur inférieure à 10^{-9}. Notant S_n^N la série limitée à $k = N$, on a

$$\Delta_n^N = S_n - S_n^N \leq \int_N^{\infty} x^{-n}\,dx = \frac{1}{n-1} \left(\frac{1}{N}\right)^{n-1} < 10^{-q}$$

pour

$$-(n-1) \log N - \log (n-1) < -q\ ,$$

soit

$$\log N > \frac{q - \log(n-1)}{n-1}\ .$$

Notant N* la solution de

$$\log N^* = \frac{q - \log(n-1)}{n-1} \quad ,$$

il faudra que N soit le plus petit entier supérieur à N*. Pour q = 9 , il vient

n	2	3	4	5	6	7	8	9	10	11
N*	10^9	22630	693,4	125,7	45,73	23,46	14,62	10,28	7,834	6,310
N	10^9	22630	694	126	46	24	15	11	8	7

n	12	13	14	15	16	17	18	19	20
N*	5,291	4,572	4,042	3,639	3,323	3,071	2,864	2,693	2,549
N	6	5	5	4	4	4	3	3	3

On constate qu'il est très aisé de calculer les séries à n élevé, mais que pour $n < 10$, le problème est tout autre.

En calculant les sommes S_n^{10} , on a, pour n = 10,

$$\Delta_{10}^{10} < \frac{1}{9} \left(\frac{1}{10} \right)^9 = 1{,}11 \cdot 10^{-10} \quad .$$

Cette erreur est inférieure à celle que nous nous sommes assignée. Dans ces conditions, on peut poser, pour $n \geqslant 10$, $S_n = S_n^{10}$. Pour $n < 10$, on notera que, si $f(x) = x^{-n}$, on a

$$f^{(p)}(x) = (-1)^p \, n(n+1)\ldots(n+p-1) \, x^{-n-p} = (-1)^p \, \frac{n+p-1)!}{(n-1)!} \, x^{-n-p}$$

et la formule établie au paragraphe précédent s'écrit

$$\Delta_n^{10} = S_n - S_n^{10} = \frac{1}{n-1} \left(\frac{1}{10 + \frac{1}{2}} \right)^{n-1} - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{(n+p-1)!}{(n-1)! \, (p+1)! \, 2^{p+1}} \, \Delta_{n+p}^{10}$$

On peut alors effectuer le calcul de Δ_9^{10} , Δ_8^{10} , ..., comme suit:

(a) - Calcul de Δ_9^{10}

$$\Sigma \approx 0$$

$$\frac{1}{8} \left(\frac{1}{10{,}5} \right)^8 = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 0{,}000\ 000\ 001$$

(b) - Calcul de Δ_8^{10}

Σ 0,000 000 000

$\frac{1}{7}\left(\frac{1}{10,5}\right)^7$ = ... 0,000 000 010

0,000 000 010

(c) - Calcul de Δ_7^{10}

$\frac{1}{6}\left(\frac{1}{10,5}\right)^6$ = ... 0,000 000 124

$-2\,\frac{8!}{6!\ 3!\ 2^3}\,\Delta_9^{10}$ = ... - 0,000 000 002

0,000 000 122

(d) - Calcul de Δ_6^{10}

$\frac{1}{5}\left(\frac{1}{10,5}\right)^5$ = ... 0,000 001 567

$-2\,\frac{7!}{5!\ 3!\ 2^3}\,\Delta_8^{10}$ = ... - 0,000 000 018

0,000 001 549

(e) - Calcul de Δ_5^{10}

$\frac{1}{4}\left(\frac{1}{10,5}\right)^4$ = ... 0,000 020 568

$-2\,\frac{6!}{4!\ 3!\ 2^3}\,\Delta_7^{10}$ = ... - 0,000 000 153

$-2\,\frac{8!}{4!\ 5!\ 2^5}\,\Delta_9^{10}$ = ... - 0,000 000 001

0,000 020 414

(f) - Calcul de Δ_4^{10}

$\frac{1}{3}\left(\frac{1}{10,5}\right)^3$ = ... 0,000 287 946

$-2\,\frac{5!}{3!\ 3!\ 2^3}\,\Delta_6^{10}$ = ... - 0,000 001 291

$-2\,\frac{7!}{3!\ 5!\ 32}\,\Delta_8^{10}$ = ... - 0,000 000 004

0,000 286 651

(g) - Calcul de Δ_3^{10}

$$\frac{1}{2}\left(\frac{1}{10,5}\right)^2 = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 0,004\ 535\ 147$$

$$-\ 2\ \frac{4!}{2!\ 3!\ 8}\ \Delta_5^{10} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad -\ 0,000\ 010\ 207$$

$$-\ 2\ \frac{6!}{2!\ 5!\ 32}\ \Delta_7^{10} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad -\ 0,000\ 000\ 023$$

$$-\ 2\ \frac{8!}{2!\ 7!\ 128}\ \Delta_9^{10} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad -\ 0,000\ 000\ 000$$

$$0,004\ 524\ 917$$

(h) - Calcul de Δ_2^{10}

$$\frac{1}{1}\left(\frac{1}{10,5}\right)^1 = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 0,095\ 238\ 095$$

$$-\ 2\ \frac{3!}{1!\ 3!\ 8}\ \Delta_4^{10} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad -\ 0,000\ 071\ 662$$

$$-\ 2\ \frac{5!}{1!\ 5!\ 32}\ \Delta_6^{10} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad -\ 0,000\ 000\ 097$$

$$-\ 2\ \frac{7!}{1!\ 7!\ 128}\ \Delta_8^{10} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad -\ 0,000\ 000\ 000$$

$$0,095\ 166\ 336$$

On obtient ainsi les résultats suivants:

n	S_n^{10}	Δ_n^{10}	S_n	$10^9.(S_n - S_n^{exact})$
2	1,549 767 731	0,095 166 336	1,644 934 067	-1
3	1,197 531 986	0,004 524 917	1,202 056 903	0
4	1,082 036 584	0,000 286 651	1,082 323 235	+1
5	1,036 907 342	0,000 020 414	1,036 927 756	+1
6	1,017 341 513	0,000 001 549	1,017 343 062	0
7	1,008 349 154	0,000 000 122	1,008 349 276	-1
8	1,004 077 346	0,000 000 010	1,004 077 356	0
9	1,002 008 392	0,000 000 001	1,002 008 393	0

n	S_n^{10}
10	1,000 994 576
11	1,000 494 188
12	1,000 246 087
13	1,000 122 713
14	1,000 061 248
15	1,000 030 589
16	1,000 015 282
17	1,000 007 637

n	S_n^{10}
18	1,000 003 818
19	1,000 001 908
20	1,000 000 954

12.3 - Calcul de la constante d'EULER

Proposons-nous de calculer la constante d'EULER

$$\gamma = \lim_{N \to \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{N} - \ln N \right) .$$

Partant de la fonction

$$f(x) = x^{-1} ,$$

on trouve

$$f^{(p)}(x) = (-1)^p \, p! \, x^{-p-1} ,$$

ce qui entraîne, par la formule (a) de la section 12.1,

$$S_1^N = \sum_{k=1}^{N} f(k) = \int_{\frac{1}{2}}^{N+\frac{1}{2}} \frac{dx}{x} - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)\, 2^{p+1}} \, S_{p+1}^N$$

$$= \ln(N + \tfrac{1}{2}) - \ln \tfrac{1}{2} - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)\, 2^{p+1}} \, S_{p+1}^N .$$

En particulier, pour N = 1, il vient

$$S_1^1 = 1 = \ln(3/2) - \ln(1/2) - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)\, 2^{p+1}} .$$

La soustraction de ces deux résultats donne

$$S_1^N = 1 + \ln(N + \tfrac{1}{2}) - \ln(3/2) - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)\, 2^{p+1}} \, (S_{p+1}^N - 1)$$

et

$$S_1^N - \ln N = 1 + \ln(1 + \tfrac{1}{2N}) - \ln(\tfrac{3}{2}) - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)\, 2^{p+1}} \, (S_{p+1}^N - 1).$$

Passant à la limite, on trouve

$$= \lim_{N \to \infty} (S_1^N - \ln N) = 1 - \ln(\tfrac{3}{2}) - 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{(p+1)\, 2^{p+1}} \, (S_{p+1} - 1) .$$

On a d'abord, en s'arrêtant à la neuvième décimale,

$$1 - \ln(3/2) = 0{,}594\ 534\ 892 .$$

Il nous reste à calculer la somme:

p	p+1	$S_{p+1} - 1$	$\frac{1}{(p+1)\ 2^{p+1}} (S_{p+1}-1)$
2	3	0,202 056 903	0,008 419 038
4	5	0,036 927 756	0,000 230 798
6	7	0,008 349 276	0,000 009 318
8	9	0,002 008 393	0,000 000 436
10	11	0,000 494 188	0,000 000 022
12	13	0,000 122 713	0,000 000 001
14	15	0,000 030 589	0,000 000 000
			0,008 659 613
			x 2
			0,017 319 225

On obtient

$$\gamma = 0{,}594\ 534\ 892 - 0{,}017\ 319\ 225 = 0{,}577\ 215\ 667\ .$$

La valeur exacte est:

$$\gamma = 0{,}577\ 215\ 664\ 9....$$

12.4 - <u>Détermination numérique de la formule de STIRLING</u>

Il s'agit d'obtenir l'expression asymptotique de N! pour N grand. L'expression

$$\ln N! = \ln 1 + ... + \ln N$$

est une somme de la forme

$$S_o^N = \sum_{n=1}^{N} f(n) \quad , \quad \text{avec} \quad f(x) = \ln x\ .$$

Comme

$$f'(x) = x^{-1} \quad , \quad f''(x) = -\,x^{-2}$$

et, en général,

$$f^{(p)}(x) = (-1)^{p-1}\ (p-1)!\ x^{-p} \quad ,$$

on obtient, par la formule (a) de la section 12.1 ,

$$S_o^N = \int_{\frac{1}{2}}^{N+\frac{1}{2}} \ln x\ dx + 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{p(p+1)\ 2^{p+1}}\ S_p^N$$

$$= (N+\tfrac{1}{2})\ \ln(N+\tfrac{1}{2}) - \tfrac{1}{2}\ \ln \tfrac{1}{2} - N + 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{p(p+1)\ 2^{p+1}}\ S_p^N\ .$$

En particulier, pour N = 1, il vient

$$S_0^1 = 0 = \frac{3}{2}\ln\frac{3}{2} - \frac{1}{2}\ln\frac{1}{2} - 1 + 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{p(p+1)\,2^{p+1}} .$$

Faisant la différence, on obtient

$$S_o^N = (N + \tfrac{1}{2})\ln(N + \tfrac{1}{2}) - \frac{3}{2}\ln\frac{3}{2} - N + 1 + 2 \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{p(p+1)2^{p+1}} (S_p^N - 1).$$

Lorsque N devient très grand, les sommes S_p^N finissent par se stabiliser pour valoir, à la limite, S_p. Pour l'expression asymptotique cherchée, il nous faut donc calculer

$$\Sigma = \sum_{\substack{p=2 \\ \text{pair}}}^{\infty} \frac{1}{p(p+1)\,2^{p}} (S_p - 1) .$$

p	$S_p - 1$	$\frac{1}{p(p+1)\,2^p}(S_p - 1)$
2	0,644 934 067	0,026 872 253
4	0,082 323 235	0,000 257 260
6	0,017 343 062	0,000 006 452
8	0,004 077 356	0,000 000 221
10	0,000 994 576	0,000 000 009
12	0,000 246 087	0,000 000 000
		Σ = 0,027 136 195

On a donc

$$N! \approx (N + \tfrac{1}{2})^{N + \frac{1}{2}} \; (3/2)^{-(3/2)} \; e^{-N} \exp(1 + \Sigma)$$

$$\approx e^{-N} N^{N + \frac{1}{2}} \; (1 + \frac{1}{2N})^{N} \; (1 + \frac{1}{2N})^{\frac{1}{2}} \; (3/2)^{-(3/2)} \exp(1 + \Sigma) .$$

Or, pour $N \to \infty$, on a

$$\lim (1 + \frac{1}{2N})^{N} = \left(\lim (1 + \frac{1}{2N})^{2N} \right)^{\frac{1}{2}} = e^{\frac{1}{2}}$$

et

$$\lim (1 + \frac{1}{2N})^{\frac{1}{2}} = 1 .$$

Par conséquent,

$$\boxed{N! \approx K\, e^{-N} N^{N} \sqrt{N}}$$

avec

$$K = \exp\left(\frac{3}{2} + \Sigma\right) \cdot (3/2)^{-(3/2)} = 2{,}506\ 628\ 274\ .$$

C'est la formule de STIRLING. Des considérations analytiques [16, 23] permettent de montrer que la vraie valeur de K est

$$K = \sqrt{2\pi} = 2{,}506\ 628\ 275.$$

TABLEAU DES 15 PREMIERS POLYNOMES DE BERNOULLI [5]

$$B_0(x) = 1$$

$$B_1(x) = -\frac{1}{2} + x$$

$$B_2(x) = \frac{1}{6} - x + x^2$$

$$B_3(x) = \frac{x}{2} - \frac{3}{2}x^2 + x^3$$

$$B_4(x) = -\frac{1}{30} + x^2 - 2x^3 + x^4$$

$$B_5(x) = -\frac{1}{6}x + \frac{5}{3}x^3 - \frac{5}{2}x^4 + x^5$$

$$B_6(x) = \frac{1}{42} - \frac{1}{2}x^2 + \frac{5}{2}x^4 - 3x^5 + x^6$$

$$B_7(x) = \frac{1}{6}x - \frac{7}{6}x^3 + \frac{7}{2}x^5 - \frac{7}{2}x^6 + x^7$$

$$B_8(x) = -\frac{1}{30} + \frac{2}{3}x^2 - \frac{7}{3}x^4 + \frac{14}{3}x^6 - 4x^7 + x^8$$

$$B_9(x) = -\frac{3}{10}x + 2x^3 - \frac{21}{5}x^5 + 6x^7 - \frac{9}{2}x^8 + x^9$$

$$B_{10}(x) = \frac{5}{66} - \frac{3}{2}x^2 + 5x^4 - 7x^6 + \frac{15}{2}x^8 - 5x^9 + x^{10}$$

$$B_{11}(x) = \frac{5}{6}x - \frac{11}{2}x^3 + 11x^5 - 11x^7 + \frac{55}{6}x^9 - \frac{11}{2}x^{10} + x^{11}$$

$$B_{12}(x) = -\frac{691}{2730} + 5x^2 - \frac{33}{2}x^4 + 22x^6 - \frac{33}{2}x^8 + 11x^{10} - 6x^{11} + x^{12}$$

$$B_{13}(x) = -\frac{691}{210}x + \frac{65}{3}x^3 - \frac{429}{10}x^5 + \frac{286}{7}x^7 - \frac{143}{6}x^9 + 13x^{11} - \frac{13}{2}x^{12} + x^{13}$$

$$B_{14}(x) = \frac{7}{6} - \frac{691}{30}x^2 + \frac{455}{6}x^4 - \frac{1001}{10}x^6 + \frac{143}{2}x^8 - \frac{1001}{30}x^{10} + \frac{91}{6}x^{12} - 7x^{13} + x^{14}$$

$$B_{15}(x) = \frac{35}{2}x - \frac{691}{6}x^3 + \frac{455}{2}x^5 - \frac{429}{2}x^7 + \frac{715}{6}x^9 - \frac{91}{2}x^{11} + \frac{35}{2}x^{13} - \frac{15}{2}x^{14} + x^{15}$$

NOMBRES DE BERNOULLI [5]

$B_0 = 1$

$B_1 = 1/2$

$B_2 = 1/6$

$B_4 = -1/30$

$B_6 = 1/42$

$B_8 = 1/30$

$B_{10} = 5/66$

$B_{12} = -691/2730$

$B_{14} = 7/6$

$B_{16} = -3617/510$

$B_{18} = 43867/798$

$B_{20} = -174\ 611/330$

TABLEAU DES 15 PREMIERS POLYNOMES D'EULER [5]

$$E_0(x) = 1$$

$$E_1(x) = -\frac{1}{2} + x$$

$$E_2(x) = -x + x^2$$

$$E_3(x) = \frac{1}{4} - \frac{3}{2}x^2 + x^3$$

$$E_4(x) = x - 2x^3 + x^4$$

$$E_5(x) = -\frac{1}{2} + \frac{5}{2}x^2 - \frac{5}{4}x^4 + x^5$$

$$E_6(x) = -3x + 5x^3 - 3x^5 + x^6$$

$$E_7(x) = \frac{17}{8} - \frac{21}{2}x^2 + \frac{35}{4}x^4 - \frac{7}{2}x^6 + x^7$$

$$E_8(x) = 17x - 28x^3 + 14x^5 - 4x^7 + x^8$$

$$E_9(x) = -\frac{31}{2} + \frac{153}{2}x^2 - 63x^4 + 21x^6 - \frac{9}{2}x^8 + x^9$$

$$E_{10}(x) = -155x + 255x^3 - 126x^5 + 30x^7 - 5x^9 + x^{10}$$

$$E_{11}(x) = \frac{691}{4} - \frac{1705}{2}x^2 + \frac{2805}{4}x^4 - 231x^6 + \frac{165}{4}x^8 - \frac{11}{2}x^{10} + x^{11}$$

$$E_{12}(x) = 2073x - 3410x^3 + 1683x^5 - 396x^7 + 55x^9 - 6x^{11} + x^{12}$$

$$E_{13}(x) = -\frac{5461}{2} + \frac{26949}{2}x^2 - \frac{22165}{2}x^4 + \frac{7293}{2}x^6 - \frac{1287}{2}x^8 + \frac{143}{2}x^{10} - \frac{13}{2}x^{12} + x^{13}$$

$$E_{14}(x) = -38227 + 62881x^3 - 31031x^5 + 7293x^7 - 1001x^9 + 91x^{11} - 7x^{13} + x^{14}$$

$$E_{15}(x) = \frac{929\,569}{16} - \frac{573\,405}{2}x^2 + \frac{943\,215}{4}x^4 - \frac{155\,155}{2}x^6 + \frac{109\,395}{8}x^8 - \frac{3003}{2}x^{10} + \frac{455}{4}x^{12} - \frac{15}{2}x^{14} + x^{15}$$

NOMBRES D'EULER [5]

$E_0 = 1$

$E_2 = -1$

$E_4 = 5$

$E_6 = -61$

$E_8 = 1385$

$E_{10} = -50\,581$

$E_{12} = 2\,702\,765$

$E_{14} = -199\,360\,981$

$E_{16} = 19\,391\,512\,145$

$E_{18} = -2\,404\,879\,675\,441$

$E_{20} = 370\,371\,188\,237\,525$

$E_{22} = -69\,348\,874\,393\,137\,901$

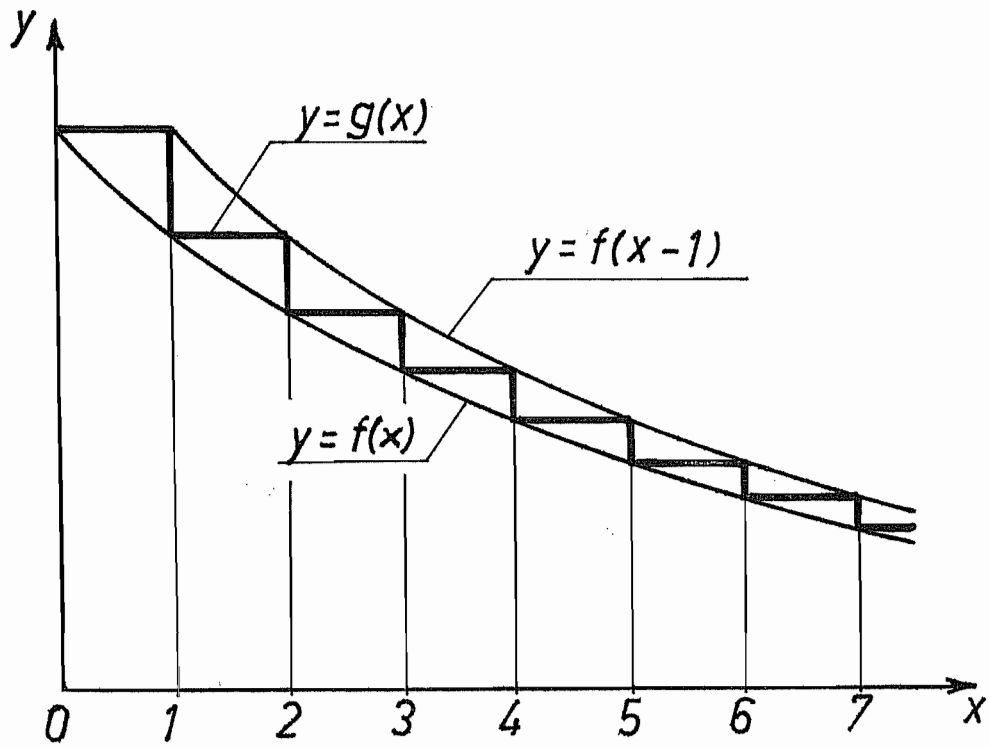


Fig. 1

BIBLIOGRAPHIE

[1] M. ZAMANSKY - Introduction à l'algèbre et l'analyse modernes
3e éd. , DUNOD, Paris, 1967

[2] F. RIESZ et B. Sz. NAGY - Leçons d'Analyse fonctionnelle
6e éd. , Gauthier-Villars, Paris, 1975

[3] A.H. STROUD - Approximate Calculation of multiple Integrals
Prentice Hall, 1971

[4] J.F. DEBONGNIE - "L'intégration dans les éléments finis"
Rapport LTAS SF-35, Université de Liège, 1976

[5] M. ABRAMOWITZ and I. STEGUN - Handbook of Mathematical Functions
Dover, New York

[6] G. HACQUES - Algorithmique numérique I
Armand Colin, Paris, 1971

[7] B. DEMIDOVITCH, I. MARON - Eléments de Calcul numérique
Mir, Moscou (1973)

[8] N. BAKHVALOV - Méthodes numériques
Mir, Moscou, 1976

[9] V. DIATCHENKO - Notions fondamentales de calcul numérique
Mir, Moscou, 1975

[10] N. VILENKINE - "Méthode des approximations successives"
in:
Quelques applications des mathématiques, par
N. VILENKINE, G. CHILOV, V. OUSPENSKI, J.LIOUBITCH, L. CHOR,
Mir, Moscou, 1975

[11] J.L. LIONS - Cours d'Analyse numérique
Cours de l'Ecole Polytechnique, Hermann, Paris, 1974

[12] V. VOIEVODINE - Principes numériques d'algèbre linéaire
Mir, Moscou, 1980

[13] B. PCHENITCHNY, Y. DANILINE - Méthodes numériques dans les problèmes d'extrémum
Mir, Moscou, 1977

[14] F.L. BAUER - "Das Verfahren der Treppen-iteration un verwandte Verfahren zur Lösung algebrische Eignewerprobleme"
ZAMP, vol VIII, pp. 214-235 , 1957

[15] F.B. HILDEBRAND Introduction to Numerical Analysis
Mc Graw Hill, 1956

[16] M. LAVRENTIEV, B. CHABAT - Méthodes de la Théorie des Fonctions d'une Variable complexe
Mir, Moscou, 1977

[17] H.G. GARNIR - Fonctions de Variables Réelles
2 tomes, Vander, Louvain

[18] H. MINEUR - Techniques du Calcul numérique
Librairie Polytechnique Béranger, Dunod, Paris, 1966

[19] A.S. HOUSEHOLDER - Principles of numerical Analysis
Mc Graw Hill, 1953

[20] H.G. GARNIR, M. DE WILDE, J. SCHMETS - Analyse fonctionnelle
Tome I, Birkhaüser, Basel und Stuttgart, 1968

[21] A. KOLMOGOROV, S. FOMINE - Eléments de la Théorie des Fonctions et de l'Analyse Fonctionnelle
Mir, Moscou, 1977

[22] A. ANGOT - Compléments de Mathématiques
5e éd., Masson, Paris, 1965

[23] R. COURANT - Differential and Integral Calculus
2 tomes, 2e éd., Blackie and Sons, London, 1937

[24] J.P. NOUGIER - Méthodes du Calcul numérique
Masson, Paris, 1983

[25] P.G. CIARLET - Introduction à l'Analyse numérique matricielle et à l'optimisation.
Masson, Paris, 1982

ISBN-13 : 978-9600313-6-2