

Octobre 1964

ORGANISATION EUROPEENNE POUR LA RECHERCHE NUCLEAIRE

TECHNIQUES DE RESOLUTION NUMERIQUE

DES SYSTEMES D'EQUATIONS :

PRINCIPALES METHODES DE MINIMISATION

# THESE

Pour l'obtention du Doctorat de Spécialité

MATHEMATIQUES (3ème cycle)

soutenue devant le jury de la Faculté des Sciences

de l'Université de NANCY

par

Michel GERARD

Thesis-1964-Gerard

Je préférerais que le Lorrain et non le mathématicien  
jette un coup d'œil sur ce texte.

Octobre 1964

Sincères amitiés

*Michel Gerard*

TECHNIQUES DE RESOLUTION NUMERIQUE

DES SYSTEMES D'EQUATIONS :

PRINCIPALES METHODES DE MINIMISATION

T H E S E

Pour l'obtention du Doctorat de Spécialité

MATHEMATIQUES (3ème cycle)

soutenue devant le jury de la Faculté des Sciences

de l'Université de NANCY

par

Michel GERARD

CERN LIBRARIES, GENEVA

CERN LIBRARIES, GENEVA



CM-P00081086

001 = 1349547

THE UNIVERSITY OF CHICAGO

LIBRARY

1000 S. MICHIGAN AVE.

CHICAGO, ILL.

1950

1000 S. MICHIGAN AVE.

CHICAGO, ILL.

1950

1000

TABLE DES MATIERES

|                                                                              | <u>Page</u> |
|------------------------------------------------------------------------------|-------------|
| <u>INTRODUCTION</u>                                                          | 1           |
| <u>EXPOSE THEORIQUE</u>                                                      |             |
| I. <u>ETUDE DIRECTE DU SYSTEME</u>                                           | 3           |
| A. Méthode de linéarisation                                                  | 3           |
| B. Méthode des approximations successives                                    | 4           |
| C. Méthode de suritération                                                   | 6           |
| D. Retour à la méthode de linéarisation                                      | 7           |
| II. <u>METHODES DE MINIMISATION</u>                                          | 9           |
| A. Fonction à une variable                                                   | 10          |
| 1) Investigations                                                            | 10          |
| 2) Interpolations par un polynôme du second degré                            | 10          |
| 3) Interpolations par un polynôme du troisième degré                         | 12          |
| B. Fonction à n variables                                                    | 18          |
| 1) Méthodes de relaxation                                                    | 19          |
| 2) Méthode du gradient                                                       | 20          |
| 3) Fonction quadratique définie positive : Méthode des directions conjuguées | 22          |
| 4) Méthode des directions conjuguées : Fonction quelconque                   | 29          |
| 5) Utilisation de n directions orthogonales                                  | 29          |
| <u>METHODES COMPAREES</u>                                                    |             |
| III. <u>TECHNIQUES UTILISEES DANS LES METHODES COMPAREES</u>                 | 33          |
| A. Méthode de linéarisation                                                  | 33          |
| B. Méthode de relaxation variable par variable                               | 37          |
| C. Méthode du gradient                                                       | 38          |
| D. Méthode des directions conjuguées                                         | 40          |
| E. Utilisation de n directions orthogonales                                  | 42          |
| 1) Interpolation par un polynôme du second degré                             | 42          |
| 2) Investigations                                                            | 44          |

|                                             | <u>Page</u> |
|---------------------------------------------|-------------|
| IV. <u>RESULTATS DES METHODES COMPAREES</u> | 45          |
| A. Résolution de systèmes d'équations       | 46          |
| 1) Premier type de fonctions                | 46          |
| 2) Deuxième type de fonctions               | 50          |
| B. Minimisation d'une fonction positive     | 55          |
| 1) Premier type de fonction                 | 55          |
| 2) Deuxième type de fonction                | 56          |
| 3) Troisième type de fonction               | 58          |
| 4) Quatrième type de fonction               | 60          |
| V. <u>CONCLUSION</u>                        | 63          |
| A. Résolution d'un système d'équations      | 63          |
| B. Minimisation d'une fonction positive     | 63          |
| 1) Ordinateur de petite capacité            | 64          |
| 2) Ordinateur de grande capacité            | 64          |
| C. Conclusion                               | 65          |
| <u>REFERENCES</u>                           | 66          |

## INTRODUCTION

Il s'agit de comparer des techniques de résolution numérique d'un système de  $n$  équations à  $n$  inconnues, qui se présenteront sous l'une des deux notations suivantes :

$$F(X) = 0 \quad (1)$$

$$X = G(X) \quad (2)$$

$X$  est le vecteur dont les composantes sont les scalaires  $x_1, x_2, \dots, x_n$ .  
 $F$  et  $G$  sont des fonctions vectorielles de composantes

$$f_1(x_1, \dots, x_n), f_2(x_1, \dots, x_n), \dots, f_n(x_1, \dots, x_n)$$

$$\text{ou } g_1(x_1, \dots, x_n), \dots, g_n(x_1, \dots, x_n) .$$

Nous supposerons que dans le domaine étudié, ces fonctions  $f_i$  ou  $g_i$  sont continues et bornées et possèdent des dérivées partielles d'ordre un et deux continues.

Une méthode commode et générale de résolution est la méthode des approximations successives; elle consiste à mettre le système (1) sous la forme (2) et à étudier la suite  $X^n$  définie par

$$X^{n+1} = G(X^n) .$$

Si à l'intérieur du domaine cette suite est convergente, sa limite  $X$  sera solution du système.

La convergence de cette suite est liée aux valeurs propres d'une matrice de terme général  $\partial g_i / \partial x_j$ . L'étude de ses valeurs propres se révèle en général très délicate, mais cette technique permet cependant la recherche de méthodes de suritération. Appliquée en particulier à un système de la forme (1), elle permet d'établir une relation de récurrence à convergence quadratique qui se trouve être identique à celle obtenue par la méthode de linéarisation. Ces méthodes seront rappelées dans la première partie de cette étude.

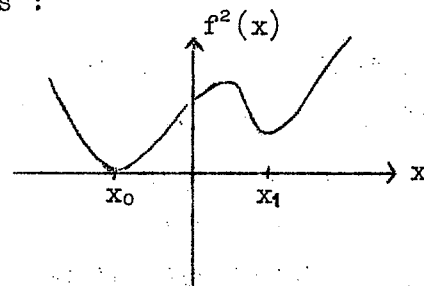
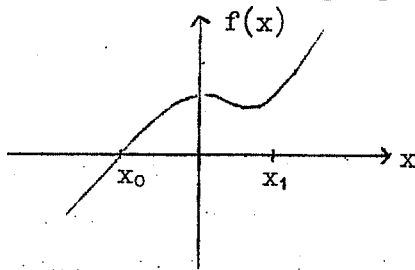
La méthode précédente a le défaut de nécessiter à chaque itération l'inversion d'une matrice d'ordre  $n$ , et la méthode devient lourde lorsque  $n$  est élevé. Il est alors tentant d'appliquer à ces systèmes assez généraux des techniques qui ont fait leur preuve dans la résolution par méthodes itératives des grands systèmes linéaires. Ces techniques consistent à associer au système (1) une expression  $\Omega(X)$ , telle que son minimum corresponde à une solution. On remplace alors la recherche de la solution de  $F(X) = 0$  par la minimisation de  $\Omega(X)$ , problème pour lequel il existe de nombreuses techniques dont la convergence est assurée.

La principale difficulté réside dans le choix de  $\Omega(X)$ . Il peut être dirigé par des considérations physiques: on sait, par exemple, que les équations définissant l'équilibre d'un système élastique correspondent à la minimisation de la somme de son énergie potentielle et de son énergie mécanique. On peut également associer la forme:

$$\Omega(X) = f_1^2(X) + f_2^2(X) + \dots + f_n^2(X) .$$

Il est évident que si nous trouvons un vecteur  $X$  tel que  $\Omega(X) = 0$ ,  $X$  est solution du système (1) et la valeur 0 est nécessairement un minimum pour  $\Omega(X)$ .

Cette méthode présente cependant le danger d'introduire des minima parasites comme le montre le cas simple d'une fonction à une variable schématisé sur les graphiques ci-dessous :



La méthode de minimisation peut conduire à la valeur  $x_1$  qui ne sera pas acceptable car  $\Omega(x_1) \neq 0$ . Pour cet exemple simple, il est facile d'éliminer cette valeur parasite; cela peut devenir délicat dans des cas plus complexes.

La recherche numérique d'un minimum d'une fonction  $\Omega(X)$  est un problème qui présente en soi-même un grand intérêt, en particulier en recherche opérationnelle, et sera l'objet de la seconde partie de cette thèse.

## CHAPITRE I

### ETUDE DIRECTE DU SYSTEME <sup>1)</sup>

Nous allons généraliser quelques méthodes itératives servant à la résolution d'une équation à une variable se présentant par analogie sous la forme :

$$\begin{aligned} f(x) &= 0 \\ x &= g(x) . \end{aligned}$$

La méthode de résolution la plus simple de  $f(x) = 0$  étant la méthode de linéarisation, nous commencerons par celle-ci.

La méthode de résolution de  $x = g(x)$  utilisant la suite  $x^{n+1} = g(x^n)$  n'est pas toujours d'un emploi facile; elle offre cependant l'avantage de pouvoir modifier la rapidité de sa convergence en étudiant la fonction  $x + \alpha x = g_1(x)$ , ce qui permet d'en tirer quelques conclusions pour la méthode de linéarisation. La généralisation de cette méthode sera développée dans la deuxième partie de ce chapitre.

#### A. METHODE DE LINEARISATION

Soit le système d'équations  $F(X) = 0$  :

$$\begin{aligned} f_1(x_1, \dots, x_n) &= 0 \\ &\dots\dots\dots \\ f_n(x_1, \dots, x_n) &= 0 . \end{aligned}$$

Prenons un vecteur  $X^0$  que nous supposerons voisin de la solution  $X$  :

$$X = X^0 + H .$$

La méthode de linéarisation consiste à développer, chacune des fonctions  $f_i$ , en série de Taylor par rapport à l'ensemble des variables et en se limitant aux premiers termes du développement :

$$f_i(x_1, \dots, x_n) = f_i(x_1^0, \dots, x_n^0) + \sum_{j=1}^n h_j \frac{\partial f_i(x_1^0, \dots, x_n^0)}{\partial x_j} .$$



On remplace ainsi le système (1) par le système approché :

$$DF(X^0) H - F(X^0) = 0 \quad (3)$$

où

$$H = \begin{bmatrix} h_1 \\ \dots \\ h_n \end{bmatrix}, \quad F(X^0) = \begin{bmatrix} f_1(X^0) \\ \dots \\ f_n(X^0) \end{bmatrix}$$

$$DF(X^0) = \begin{bmatrix} \frac{\partial f_1(X^0)}{\partial x_1}, \dots, \frac{\partial f_1(X^0)}{\partial x_n} \\ \dots \\ \frac{\partial f_n(X^0)}{\partial x_1}, \dots, \frac{\partial f_n(X^0)}{\partial x_n} \end{bmatrix}$$

Ce système approché, linéaire en  $h_i$ , admet comme solution :

$$H = - [DF(X^0)]^{-1} F(X^0)$$

d'où la solution  $X$  :

$$X = X^0 + H = X^0 - [DF(X^0)]^{-1} F(X^0).$$

En fait, cette méthode étant itérative, nous partirons d'une valeur  $X^k$  et nous chercherons la solution du système linéaire approché, ce qui donne la formulation générale pour l'itération  $k$  :

$$X^{k+1} = X^k + H^k = X^k - [DF(X^k)]^{-1} F(X^k).$$

#### B. METHODE DES APPROXIMATIONS SUCCESSIVES - CONVERGENCE

Nous cherchons, dans un domaine, une solution de  $X = G(X)$ , soit :

$$\begin{aligned} x_1 &= g_1(x_1, \dots, x_n) \\ &\dots \\ x_n &= g_n(x_1, \dots, x_n) \end{aligned}$$

Partant d'un vecteur  $X^0$ , nous formons la suite de vecteurs

$$X^{k+1} = G(X^k) .$$

Si le vecteur  $X^{k+1}$  tend vers une limite  $X$  intérieure au domaine lorsque  $k$  augmente indéfiniment, donc si chaque composante  $X_i^k$  tend vers une limite  $x_i$ ,  $X$  sera une solution du système (2).

Si les fonctions  $g_i(x_1, \dots, x_n)$  sont une fois continuellement différentiables dans le domaine, le théorème des accroissements finis nous donnera  $n$  relations :

$$g_i(x_1^k, \dots, x_n^k) - g_i(x_1^{k-1}, \dots, x_n^{k-1}) = \sum_{j=1}^n (x_j^k - x_j^{k-1}) \frac{\partial g_i(\bar{X})}{\partial x_j} \quad i = 1, 2, \dots, n ,$$

$\bar{X}$  étant un vecteur dont chaque composante  $\bar{x}_i$  est intérieure à l'intervalle  $[x_i^{k-1}, x_i^k]$ . Si, de plus, nous supposons que chaque dérivée partielle  $\partial g_i / \partial x_j$  est bornée supérieurement en valeur absolue dans le domaine par un nombre  $A_{ij}$ , nous obtenons alors  $n$  inégalités du type :

$$|g_i(X^k) - g_i(X^{k-1})| \leq \sum_{j=1}^n A_{ij} |x_j^k - x_j^{k-1}| \quad i = 1, 2, \dots, n .$$

Remplaçons l'étude de la convergence de la suite de vecteurs  $X^{k+1}$  par la convergence d'une série de vecteurs  $U^{k+1}$  en posant :

$$U^0 = X^0$$

$$U^1 = X^1 - X^0$$

$$\dots\dots\dots$$

$$U^{k+1} = X^{k+1} - X^k ,$$

ce qui donne

$$X^{k+1} = U^0 + U^1 + \dots + U^{k+1} .$$

Le vecteur  $U^{k+1}$  a pour composantes :

$$U_i^{k+1} = g_i(X^k) - g_i(X^{k-1}) = \sum_{j=1}^n U_j^k \frac{\partial g_i(\bar{X})}{\partial x_j} \quad i = 1, 2, \dots, n.$$

Si les dérivées partielles sont bornées supérieurement, nous obtenons :

$$|U_i^{k+1}| \leq \sum_{j=1}^n A_{ij} |U_j^k| \quad i = 1, 2, \dots, n.$$

La série de vecteurs  $U^{k+1}$  est majorée par la série de vecteurs  $V^{k+1}$  définie par :

$$V^0 = X^0$$

$$V_i^{k+1} = \sum_{j=1}^n A_{ij} V_j^k \quad i = 1, 2, \dots, n.$$

Une condition suffisante pour que cette série  $V^{k+1}$  converge est que les valeurs propres de la matrice A soient toutes en module inférieures à l'unité. Cependant la série  $U^{k+1}$ , majorée par la série  $V^{k+1}$ , peut très bien converger lorsque  $V^{k+1}$  diverge.

La convergence de cette méthode est d'une étude difficile; néanmoins, elle permet la recherche d'une méthode ayant une convergence plus rapide : c'est la méthode de suritération qui s'appliquera à la résolution du système (1).

### C. METHODE DE SURITERATION - APPLICATION A LA RESOLUTION DU SYSTEME (1)

La méthode de suritération utilise une matrice A rendant la convergence aussi rapide que possible. Un cas simple est celui où toutes ses valeurs propres sont voisines de zéro : la matrice A est alors une matrice presque nulle (tous ses éléments tendent vers zéro).

En introduisant une matrice inconnue auxiliaire B, nous remplaçons l'étude du système (1) par le système équivalent :

$$F(X) + BX = BX$$

que l'on peut mettre alors sous une forme analogue à celle du système (2)

$$X = X + B^{-1} F(X) .$$

Nous le résoudrons par approximations successives en déterminant, à chaque itération k, la matrice  $B^k$  qui annule la matrice A définie précédemment. Les  $A_{ij}$  étant des majorantes des  $\partial g_i / \partial x_j$ , nous les approcherons en prenant  $A_{ij} = \partial g_i(X^k) / \partial x_j$ , soit

$$X^{k+1} = X^k + [B^k]^{-1} F(X^k) \equiv G(X^k) .$$

La dérivation nous fournit

$$A = I + [B^k]^{-1} DF(X^k) ,$$

$DF(X^k)$  étant la matrice des dérivées partielles d'ordre un de  $F(X)$ . La matrice A s'annulera donc pour :

$$B^k = - DF(X^k) ,$$

ce qui donne la formulation itérative :

$$X^{k+1} = X^k - [DF(X^k)]^{-1} F(X^k) .$$

Nous retrouvons alors les formules obtenues par la méthode de linéarisation qui est donc la méthode des approximations successives ayant la convergence la plus rapide.

#### D. RETOUR A LA METHODE DE LINEARISATION

En fait, dans la méthode des approximations successives, le calcul des  $A_{ij}$  majorant les dérivées partielles  $\partial g_i / \partial x_j$  n'a pas besoin d'être effectué avec une grande précision, car c'est seulement un facteur

assurance pour la convergence de cette méthode; la seule nécessité pour cette matrice A est d'avoir toutes ses valeurs propres en module inférieures à l'unité.

Nous pourrions appliquer ces conclusions à la méthode de linéarisation qui essaye de rendre nulle toutes les valeurs propres de la matrice A. On pourra donc approcher numériquement les dérivées de F(X) en prenant :

$$\frac{\partial f_i(X^k)}{\partial x_j} = \frac{f_i(x_1^k, \dots, x_j^k + dx_j, \dots, x_n^k) - f_i(x_1^k, \dots, x_j^k, \dots, x_n^k)}{dx_j}$$

## CHAPITRE II

### METHODES DE MINIMISATION

Nous remplacerons dans certains cas l'étude directe de  $F(X) = 0$  par la minimisation d'une fonction unique. Pour des fonctions générales,

$$\Omega(X) = \sum_{i=1}^n f_i^2(X)$$

donnera une fonction positive ou nulle, dont les minima nuls correspondent à une solution du système (1).

Les méthodes de minimisation sont itératives; elles consistent à passer d'un point  $X^k$  au point  $X^{k+1}$ , de telle façon que  $\Omega(X^{k+1}) < \Omega(X^k)$ , ce qui donnera en général un procédé convergent si, dans le domaine d'étude,  $\Omega(X)$  possède au moins un minimum bien défini. Le seul inconvénient de ces méthodes est qu'elles peuvent nous conduire à un  $\Omega(X)$  non nul qui n'est pas une solution de (1).

On ramène ce problème à la minimisation d'une fonction à une variable en cherchant  $X^{k+1}$  le long d'une direction  $D^k$  que l'on se fixe à priori. Il suffit alors de minimiser la fonction

$$f(\ell) = \Omega(X^k + \ell D^k)$$

par rapport à la variable  $\ell$ . Soit  $\ell^k$  cette valeur optimum,  $X^{k+1}$  sera donné pour toutes ces méthodes par la formulation générale :

$$X^{k+1} = X^k + \ell^k D^k \quad (4)$$

Nous exposerons dans la première partie de ce chapitre les principales méthodes de minimisation d'une fonction à une variable d'un emploi facile pour le calcul automatique. La deuxième partie sera consacrée aux méthodes de minimisation d'une fonction à  $n$  variables, caractérisées par le choix de la direction  $D^k$ .

## A. METHODES DE MINIMISATION D'UNE FONCTION A UNE VARIABLE

Soit une fonction  $f(x)$  définie, continue, positive ou nulle sur un certain intervalle de travail et possédant une dérivée définie et continue sur ce même intervalle.

Si la fonction se prête mal à l'interpolation, nous rechercherons une valeur minimum de  $f(x)$  par des tâtonnements successifs. Si la fonction est régulière, nous l'interpolerons par un polynôme dont nous évaluerons le minimum. Ce procédé devant être répété un grand nombre de fois au cours des méthodes itératives de minimisation, seuls les polynômes de degré deux ou trois se prêteront à des calculs rapides.

### 1) Méthode d'investigation

Nous procéderons par itérations; partant d'un point  $x_0$ , avec un pas  $h_0$ , nous chercherons  $x_{k+1}$  tel que  $f(x_{k+1}) < f(x_k)$ .

Si  $f(x_k + h_k)$  donne satisfaction nous prendrons comme nouvelle approximation du minimum  $x_{k+1} = x_k + h_k$  et nous augmenterons le pas en espérant atteindre plus rapidement le minimum; à la prochaine itération  $h_{k+1} = h_k \cdot \alpha$ , avec  $\alpha > 1$ .

Si  $f(x_k + h_k)$  ne répond pas à la condition imposée, nous garderons  $x_k$  et prendrons un pas plus petit et de sens opposé; à la prochaine itération  $h_{k+1} = -h_k \cdot \beta$ , avec  $0 < \beta < 1$ .

### 2) Interpolation par un polynôme du second degré

Le polynôme du second degré n'admettra un minimum que si sa dérivée seconde est positive ou nulle. Ayant déterminé le minimum  $x_m$  du polynôme, il nous faudra ensuite nous assurer que  $x_m$  correspond à un minimum pour  $f(x)$ .

Bornons nous aux deux cas où l'on connaît soit les valeurs de la fonction en trois points équidistants, soit les valeurs des dérivées première et seconde en un point.

a) Trois points équidistants

Ayant choisi un pas d'interpolation  $h$  dans le sens de décroissance de  $f(x)$ , il nous faut déterminer le polynôme  $P_2(x)$ , interpolant  $f(x)$ , à l'aide des trois points  $x_0, x_1 = x_0 + h, x_2 = x_0 + 2h$ . Une méthode élégante consiste à employer les polynômes d'interpolation de Lagrange mis sous une forme matricielle<sup>2)</sup>. Notons

$$f_0 = f(x_0) \quad f_1 = f(x_1) \quad f_2 = f(x_2).$$

Après avoir fait le changement d'origine  $y = x - x_0$ , le polynôme d'interpolation en  $y$  s'écrit :

$$P_2(y) = \begin{bmatrix} 1, \frac{y}{h}, \frac{y^2}{h^2} \end{bmatrix} L_2 F_2$$

avec

$$F_2 = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{3}{2} & 2 & -\frac{1}{2} \\ \frac{1}{2} & -1 & \frac{1}{2} \end{bmatrix}$$

Le minimum  $y_m$  sera obtenu pour  $\partial P_2(y)/\partial y = P_2'(y) = 0$ , soit :

$$P_2'(y) = \begin{bmatrix} 0, \frac{1}{h}, \frac{2y}{h^2} \end{bmatrix} L_2 F_2$$

d'où

$$y_m = \frac{h}{2} \left[ \frac{3f_0 - 4f_1 + f_2}{f_0 - 2f_1 + f_2} \right],$$

ce qui donne  $x_m$ , minimum de  $P_2(x)$  si  $P_2''(x) > 0$  :

$$x_m = x_0 + \frac{h}{2} \left[ 1 - 2 \frac{f_1 - f_0}{f_0 + f_2 - 2f_1} \right].$$

Ayant ainsi obtenu quatre points auxquels sont associées quatre valeurs de  $f(x)$ , nous prendrons comme minimum de  $f(x)$ , celui des quatre points donnant à la fonction la plus petite valeur.



b) Dérivées première et seconde

Soit  $P_2(x) = ax^2 + bx + c$  le polynôme d'interpolation, les valeurs des dérivées au point  $x_0$  nous fournissent :

$$2ax_0 + b = f'(x_0)$$

$$2a = f''(x_0)$$

d'où le minimum  $x_m$  si  $f''(x_0) > 0$  :

$$x_m = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

Si les dérivées ne sont pas connues analytiquement on pourra les approcher par un procédé numérique en prenant soit

$$f'(x_0) = \frac{f(x_0 + dx_0) - f(x_0)}{dx_0},$$

ce qui donne

$$x_m = x_0 - dx_0 \frac{f(x_0 + dx_0) - f(x_0)}{f(x_0 + 2dx_0) + f(x_0) - 2f(x_0 + dx_0)}$$

soit la formulation plus exacte :

$$f'(x_0) = \frac{f(x_0 + dx_0) - f(x_0 - dx_0)}{2dx_0}$$

d'où

$$x_m = x_0 - 2dx_0 \frac{f(x_0 + dx_0) - f(x_0 - dx_0)}{f(x_0 + 2dx_0) + f(x_0 - 2dx_0) - 2f(x_0)}$$

3) Interpolation par un polynôme du troisième degré

Nous interpolerons  $f(x)$  par un polynôme  $P_3(x)$  dont nous évaluerons le minimum  $x_m$ ; nous déterminerons ensuite si  $x_m$  représente un minimum pour  $f(x)$ .

Au minimum  $x_m$ , la dérivée de  $P_3(x)$  sera nulle, ce qui nous ramène à résoudre une équation du second degré, soit

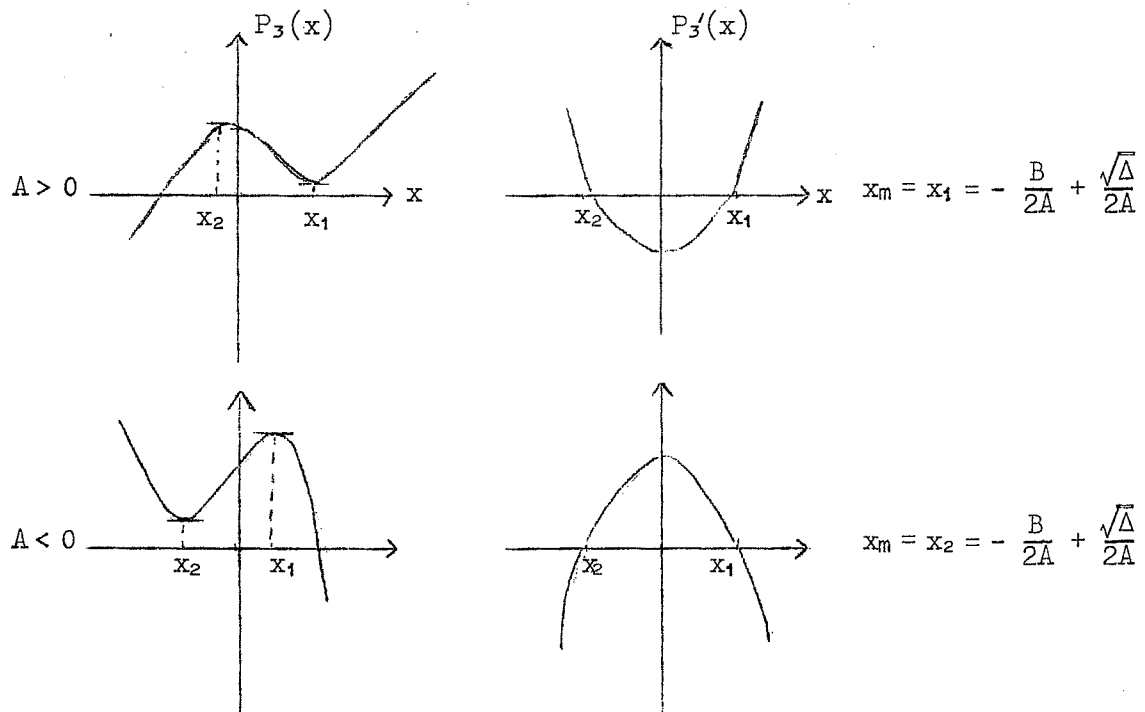
$$P_3(x) = ax^3 + \beta x^2 + \gamma x + \epsilon \quad \text{et} \quad P_3'(x) = Ax^2 + Bx + C$$

Si le polynôme représente une droite ( $\alpha = \beta = 0$ ), nous devons recommencer l'interpolation sur un intervalle différent car par définition  $f(x)$  est une fonction positive ou nulle et ne peut devenir linéaire que si  $\gamma = 0$  et  $\epsilon > 0$ .

Si  $P_3(x)$  représente une courbe du second degré ( $\alpha = 0$ ), nous pourrions soit rechercher un minimum si il existe ( $x_m = -B/A$  avec  $A > 0$ ), soit recommencer l'interpolation sur un autre intervalle.

Si  $P_3'(x)$  nous fournit deux racines imaginaires ( $B^2 - 4AC < 0$ ), la courbe d'interpolation ne conviendra pas car nous ne voulons que des valeurs réelles et nous recommencerons aussi l'interpolation.

Examinons maintenant le cas où  $P_3'(x)$  admet deux racines réelles : il nous faut déterminer laquelle de ces deux racines donne un minimum. Nous aurons deux séries de courbes suivant le signe de  $A$  :



Nous voyons que dans les deux cas  $x_m = (-B + \sqrt{\Delta})/2A$ .

Nous examinerons les deux cas où l'on connaît soit les valeurs de la fonction en quatre points équidistants, soit les valeurs de la fonction et de sa dérivée première en deux points.

a) Quatre points équidistants

Reprenons la notation matricielle pour les polynômes d'interpolation de Lagrange. Notons

$$x_1 = x_0 + h \quad x_2 = x_0 + 2h \quad x_3 = x_0 + 3h \quad f_i = f(x_i) \quad i = 0, 1, 2, 3.$$

En faisant le changement d'origine  $y = x - x_0$ , ce polynôme s'écrit :

$$P_3(y) = \left[ 1, \frac{y}{h}, \frac{y^2}{h^2}, \frac{y^3}{h^3} \right] L_3 F_3$$

avec

$$F_3 = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix}, \quad L_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -11/6 & 3 & -3/2 & 1/3 \\ 1 & -5/2 & 2 & -1/2 \\ -1/6 & 1/2 & -1/2 & 1/6 \end{bmatrix}$$

au minimum  $P_3'(y) = 0$ ,

$$P_3'(y) = \left[ 0, \frac{1}{h}, \frac{2y}{h^2}, \frac{3y^2}{h^3} \right] L_3 F_3,$$

polynôme que l'on peut mettre sous la forme

$$P_3'(y) = Ay^2 + By + C = 0$$

où

$$A = \frac{1}{2h^3} \left[ -f_0 + 3f_1 - 3f_2 + f_3 \right]$$

$$B = \frac{1}{h^2} \left[ 2f_0 - 5f_1 + 4f_2 - f_3 \right]$$

$$C = \frac{1}{6h} \left[ -11f_0 + 18f_1 - 9f_2 + 2f_3 \right]$$

ce qui donne dans le cas de racines réelles :

$$y_m = \frac{-B + [B^2 - 4AC]^{1/2}}{2A},$$

d'où le minimum  $x_m$  de  $P_3(x)$  :

$$x_m = x_0 - \frac{2C}{B + [B^2 - 4AC]^{1/2}}.$$

Il restera à prendre comme minimum de  $f(x)$  celui des cinq points donnant à la fonction la plus petite valeur.

b) Valeurs de la fonction et de sa dérivée première en deux points

Soit une fonction et sa dérivée continues sur un intervalle où nous connaissons les valeurs de  $f(x)$  et  $f'(x)$  en  $q$  points :  $x_1, x_2, \dots, x_q$ . Nous allons construire, à l'aide de ces  $2q$  paramètres, un polynôme de degré  $(2q-1)$  :  $P_{2q-1}(x)$  répondant à

$$P_{2q-1}(x_i) = f(x_i) \quad \text{et} \quad P'_{2q-1}(x_i) = f'(x_i) \quad i = 1, 2, \dots, q. \quad (5)$$

Supposons que si ce polynôme existe, il peut s'écrire sous la forme :

$$P_{2q-1}(x) = \sum_{i=1}^q [b_i(x) \cdot f(x_i) + \bar{b}_i(x) \cdot f'(x_i)]$$

où  $b_i(x)$  et  $\bar{b}_i(x)$  sont deux polynômes de degré inférieur ou égal à  $(2q-1)$ . Il est clair que la condition (5) sera réalisée si

$$b_i(x_j) = \delta_{ij}, \quad \bar{b}_i(x_j) = 0$$

$$\text{et} \quad b'_i(x_j) = 0, \quad \bar{b}'_i(x_j) = \delta_{ij}.$$

Les polynômes de Lagrange de degré  $(q-1)$  :

$$l_i(x) = \frac{(x-x_1) \dots (x-x_{i-1})(x-x_{i+1}) \dots (x-x_q)}{(x_i-x_1) \dots (x_i-x_{i-1})(x_i-x_{i+1}) \dots (x_i-x_q)} \quad \text{vérifient} \quad l_i(x_j) = \delta_{ij}.$$

Par conséquent, les polynômes de degré  $(2q-2)$  :  $c_i(x) = [\ell_i(x)]^2$  répondent à

$$c_i(x_j) = \delta_{ij} \quad \text{et} \quad c'_i(x_j) = 2\ell'_i(x_j) \cdot \ell_i(x_j) = 2\ell'_i(x_j) \cdot \delta_{ij}.$$

A l'aide de 2 fonctions  $r_i(x)$  et  $s_i(x)$  linéaires en  $x$ , nous allons former

$$\begin{aligned} b_i(x) &= r_i(x) \cdot c_i(x) \\ \bar{b}_i(x) &= s_i(x) \cdot c_i(x); \end{aligned}$$

comme  $c_i(x_j) = \delta_{ij}$  et  $c'_i(x_j) = 0$  si  $i \neq j$ , il nous reste à calculer  $r_i(x)$  et  $s_i(x)$ , de telle façon que :

$$\begin{aligned} b_i(x_i) &= 1, \quad \bar{b}_i(x_i) = 0 \\ \text{et} \quad b'_i(x_i) &= 0, \quad \bar{b}'_i(x_i) = 1 \end{aligned}$$

ce qui donne

$$\begin{aligned} r_i(x_i) \cdot c_i(x_i) &= 1, & s_i(x_i) \cdot c_i(x_i) &= 0, & \text{et} \\ r'_i(x_i) \cdot c_i(x_i) + r_i(x_i) \cdot c'_i(x_i) &= 0, & s'_i(x_i) \cdot c_i(x_i) + s_i(x_i) \cdot c'_i(x_i) &= 0. \end{aligned}$$

Les deux premières conditions fournissent, en tenant compte des propriétés de  $c_i(x)$  :

$$r_i(x_i) = 1, \quad s_i(x_i) = 0.$$

En utilisant ces résultats pour les deux dernières conditions, nous obtenons :

$$r'_i(x_i) + 2\ell'_i(x_i) = 0, \quad s'_i(x_i) = 1.$$

On en tire immédiatement :

$$\begin{aligned} r_i(x) &= 1 - 2(x - x_i) \cdot \ell'_i(x_i) \\ s_i(x) &= (x - x_i), \end{aligned}$$

d'où les deux polynômes répondant à la condition (5) :

$$b_i(x) = [\ell_i(x)]^2 \cdot [1 - 2(x - x_i) \cdot \ell'_i(x_i)]$$

$$\bar{b}_i(x) = [\ell_i(x)]^2 [x - x_i] .$$

Dans le cas où  $q=2$ , posons

$$x_2 = x_1 + h; f_1 = f(x_1), f_2 = f(x_2); f'_1 = f'(x_1), f'_2 = f'(x_2) .$$

En faisant le changement d'origine  $y = x - x_1$ , nous calculons

$$\ell_1(x) = \frac{x - x_2}{-h} = -\frac{y - h}{h} \quad \ell'_1(x_1) = -\frac{1}{h}$$

$$\ell_2(x) = \frac{x - x_1}{h} = \frac{y}{h} \quad \ell'_2(x_2) = \frac{1}{h}$$

ce qui permet de former :

$$P_3(y) = \frac{(y-h)^2}{h^2} \left[ \left( 1 + \frac{2y}{h} \right) f_1 + y f'_1 \right] + \frac{y^2}{h^2} \left[ \left( 1 - 2 \frac{y-h}{h} \right) f_2 + (y-h) f'_2 \right] .$$

La dérivation de  $P_3(y)$  donne :

$$P'_3(y) = A \frac{y^2}{h^2} - 2B \frac{y}{h} + C$$

où

$$A = 6 \frac{f_1 - f_2}{h} + 3f'_1 + 3f'_2$$

$$B = 3 \frac{f_1 - f_2}{h} + 2f'_1 + f'_2$$

$$C = f'_1 .$$

En posant

$$E = 3 \frac{f_1 - f_2}{h} + f'_1 + f'_2 ,$$

ceci conduit à

$$A = 2E + f'_1 + f'_2$$

$$B = E + f'_1$$

$$C = f'_1 .$$

Dans le cas de racines réelles, la racine donnant un minimum à  $P_3'(y)$  est :

$$y_m = h \frac{B + [B^2 - AC]^{1/2}}{A} = h \frac{C}{B - [B^2 - AC]^{1/2}}.$$

Notons par  $\Delta = [B^2 - AC]^{1/2} = [E^2 - f_1' \cdot f_2']^{1/2}$ , on peut simplifier l'écriture de  $y_m$  :

$$y_m = h \frac{f_1'}{E + f_1' - \Delta} = h \left[ 1 - \frac{(E - \Delta)(E + \Delta - f_1')}{E^2 - (f_1' - \Delta)^2} \right],$$

que l'on réduit finalement à

$$y_m = h \left[ 1 - \frac{f_2' + \Delta - E}{f_2' - f_1' + 2\Delta} \right]$$

d'où la forme condensée pour le minimum  $x_m$  de  $P_3(x)$

$$x_m = x_1 + h[1 - D]$$

avec

$$D = \frac{f_2' + \Delta - E}{f_2' - f_1' + 2\Delta}$$

$$\Delta = [E^2 - f_1' f_2']^{1/2}$$

$$E = 3 \frac{f_1 - f_2}{h} + f_1' + f_2',$$

ce qui donnera le maximum de précision au calcul de  $x_m$ .

Ce minimum ainsi obtenu sera jugé minimum de  $f(x)$  si  $f(x_m)$  est inférieur à  $f_1$  et  $f_2$ , sinon nous recommencerons l'interpolation.

#### B. MINIMUM D'UNE FONCTION DE N VARIABLES<sup>3)</sup>

Nous avons vu que l'on ramène ce problème à la minimisation d'une fonction à une variable en définissant à chaque itération une direction  $D^k$ .

Afin de diminuer la difficulté du calcul de  $f(\ell) = \Omega(X^k + \ell D^k)$ ,  $D^k$  fut prise parallèle à l'un des axes  $Ox_1, \dots, Ox_n$ , ce qui conduisit aux méthodes dites de relaxation qui ne travaillent que sur une seule composante de  $X$  à la fois.

On pensa cependant que ces méthodes allaient nécessiter un nombre d'itérations et de calcul supérieur aux méthodes faisant varier toutes les composantes de  $X$  à la fois et l'on utilisa la direction opposée à celle du gradient<sup>4)</sup> :

$$D^k = - \text{grad}[\Omega(X^k)] .$$

Cette méthode fut appelée méthode du gradient ou de la plus grande pente (Method of Steepest Descent).

La rapidité de convergence de cette dernière méthode étant imprévisible on essaya alors d'approcher itérativement la matrice  $G$  des dérivées partielles d'ordre deux de  $\Omega(X)$ . Dans le cas d'une fonction quadratique définie positive  $Q(X)$ , on utilisait des directions telles que  $(D^i, G D^k) = 0$  d'où son nom de méthode des directions conjuguées; cette manière permettait de trouver le minimum de  $Q(X)$  en  $n$  itérations. L'extension de cette méthode à des fonctions quelconques se révèle intéressante car près d'un minimum  $\Omega(X)$  sera assimilable à une fonction quadratique définie positive (développement de la fonction  $\Omega(X)$  en série de Taylor limité au deuxième ordre).

Si aucune des méthodes citées ne donne de résultats, on peut toujours essayer de rechercher le minimum de  $\Omega(X)$  à l'aide d'un système de  $n$  directions orthogonales et changer ce système afin d'accélérer la convergence. Cette méthode très générale, utilisant un grand nombre de mémoires, nécessitera un volume important de calculs.

#### 1) Méthodes de relaxation

On peut faire varier chacune des composantes de  $X$  ou alors une seule composante diminuant la fonction assez rapidement.

##### a) Relaxation variable par variable

On utilise successivement  $Ox_1, \dots, Ox_n$ , puis on recommence.



b) Méthode de Southwell

Consiste à faire varier à chaque itération  $k$  la  $j$ ème composante de  $X$ ; soit  $g^k = \text{grad}[\Omega(X^k)]$ , l'indice  $j$  est défini par :

$$|g_j^k| = \text{MAXIMUM} \{ |g_1^k|, \dots, |g_n^k| \} .$$

2) Méthode du gradient

a) Principe

Soit  $D_1^k, \dots, D_n^k$  les composantes de la direction  $D^k$  liées par la relation :

$$\sum_{i=1}^n (D_i^k)^2 = 1$$

et

$$X^{k+1} = X^k + \ell^k D^k$$

le nouveau minimum après la  $k$ ème itération. Nous allons déterminer la direction  $D^k$  donnant à

$$\Delta\Omega^k = \Omega(X^k) - \Omega(X^{k+1})$$

la plus grande valeur possible. Si nous développons  $\Omega(X^{k+1})$  en série de Taylor limitée aux premiers termes, nous obtenons

$$\Delta\Omega^k = - \ell^k \sum_{i=1}^n D_i^k \frac{\partial \Omega(X^k)}{\partial x_i} = - \ell^k (D^k, g^k) .$$

Pour que  $\Delta\Omega^k$  soit maximum en considérant  $\ell^k$  comme une constante, il suffira de maximiser

$$\sum_{i=1}^n D_i^k \frac{\partial \Omega(X^k)}{\partial x_i} .$$

Les variables  $D_i^k$  étant des variables liées, nous introduisons un multiplicateur de Lagrange que nous notons  $\mu$ , et nous cherchons le maximum de la fonction

$$\psi(D_1^k, \dots, D_n^k) = \sum_{i=1}^n D_i^k \frac{\partial \Omega(X^k)}{\partial x_i} + \mu \sum_{i=1}^n (D_i^k)^2$$

qui devra alors vérifier les  $n$  conditions

$$\frac{\partial \psi}{\partial D_i^k} = 0 \quad i = 1, 2, \dots, n,$$

ce qui donne les  $(n+1)$  équations permettant de définir  $D_1^k, \dots, D_n^k$  et  $\mu$  :

$$\frac{\partial \Omega}{\partial x_1} + \mu D_1^k = 0$$

.....

$$\frac{\partial \Omega}{\partial x_n} + \mu D_n^k = 0$$

$$(D_1^k)^2 + \dots + (D_n^k)^2 = 1$$

Nous tirons immédiatement  $D_i^k = -\frac{1}{\mu} \frac{\partial \Omega}{\partial x_i}$ .

La dernière relation fournissant

$$\mu = \left[ \sum_{i=1}^n (D_i^k)^2 \right]^{1/2},$$

$D^k$  se trouve donc être la direction opposée du gradient.

b) Convergence

La constante  $\mu$  qui représente la longueur de  $g^k = \text{grad}[\Omega(X^k)]$  pourra être englobée dans la constante  $\ell^k$ , ce qui donne :

$$D^k = -g^k$$

d'où

$$\Delta\Omega^k = +\ell^k(g^k, g^k).$$

En trouvant un nombre  $\ell^k$  positif  $\Omega(X^k) > \Omega(X^{k+1})$ . Si, dans le domaine d'étude  $\Omega(X)$  possède au moins un minimum bien défini nous aurons en général une méthode convergente.

3) Méthode des directions conjuguées. Fonction quadratique définie positive

a) Définitions - Principe de la méthode

Une fonction quadratique définie positive a pour équation

$$Q(X) = Q_0 + (A, X) + (X, BX)$$

où B est une matrice définie positive, c'est-à-dire que pour tout vecteur V non nul

$$(V, BV) > 0.$$

Deux directions U et V sont dites conjuguées ou orthogonales par rapport à la matrice B si elles vérifient

$$(U, BV) = (BU, V) = 0 ;$$

c'est une extension de la notion d'orthogonalité de deux vecteurs :

$$(U, V) = 0.$$

$\Omega(X)$  étant une fonction quadratique définie positive que nous noterons Q(X), sa matrice G des dérivées partielles d'ordre deux est une matrice définie positive de terme général  $G_{ij} = \partial^2 Q(X) / \partial x_i \partial x_j$ ; soit g le gradient de Q(X) d'élément  $g_i = \partial Q(X) / \partial x_i$ , nous pouvons donner une représentation un peu différente de Q(X) et de son gradient à l'aide de G :

$$Q(X) = Q_0 + \sum_{i=1}^n C_i x_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n x_i x_j \frac{\partial^2 Q(X)}{\partial x_i \partial x_j}$$

d'où

$$g_i = C_i + \sum_{j=1}^n x_j \frac{\partial^2 Q(X)}{\partial x_i \partial x_j}$$

que nous mettrons sous la notation abrégée :

$$Q(X) = Q_0 + (C, X) + \frac{1}{2} (X, GX) \quad (6)$$

$$g = C + GX \quad (7)$$

Soit deux valeurs  $X^k$  et  $X^{k+1}$  auxquelles sont associés  $g^k$  et  $g^{k+1}$ , convenons de noter :

$$DX^k = X^{k+1} - X^k$$

$$Dg^k = g^{k+1} - g^k$$

L'application de l'équation (7) donnera

$$Dg^k = GDX^k \quad (8)$$

soit

$$DX^k = G^{-1} Dg^k$$

Connaissant la valeur de  $G^{-1}$  nous pourrions en une itération atteindre le minimum de  $Q(X)$ ; ce calcul n'étant pas toujours aisé, Davidson<sup>5)</sup> eut l'idée d'approcher, à l'aide d'une méthode itérative, la matrice  $G^{-1}$  par une matrice  $H$ . Sa méthode étant assez complexe, Fletcher et Powell<sup>6)</sup> la simplifièrent de façon à rendre son utilisation plus facile. C'est cette dernière méthode que nous exposerons dans ce paragraphe.

Partant d'une matrice  $H^0$  définie positive que nous prendrons égale à la matrice unité, nous allons à chaque itération améliorer cette approximation de  $G^{-1}$  de façon à obtenir dans le cas d'une fonction quadratique définie positive  $H^{n+1} = G^{-1}$ . Pour cela nous minimiserons

encore  $Q(X)$  à l'aide d'une direction  $D^k$  bien déterminée; ce qui nous permettra d'obtenir

$$X^{k+1} = X^k + DX^k.$$

Nous construirons alors  $H^{k+1}$  de telle façon qu'elle vérifie la relation

$$H^{k+1} G DX^k = DX^k;$$

le dernier point trouvé  $X^{n+1}$  sera alors le minimum de  $Q(X)$ .

a) Construction de la matrice  $H^{k+1}$

Soit

$$X^{k+1} = X^k + DX^k$$

le nouveau minimum relatif. Nous voulons construire  $H^{k+1}$ , vérifiant

$$H^{k+1} G DX^k = H^{k+1} Dg^k = DX^k.$$

Nous désirons d'une part vérifier cette relation et d'autre part améliorer la matrice  $H^k$ , ce qui donne  $H^{k+1}$  sous la forme d'une somme de trois matrices :

$$H^{k+1} = R^k + H^k + S^k$$

$R^k$  vérifiera la relation  $R^k Dg^k = DX^k$  et  $S^k$  répondra à  $H^k Dg^k + S^k Dg^k = 0$ . En prenant deux vecteurs  $Y$  et  $Z$  quelconques, la forme la plus simple à leur donner serait

$$R^k = \frac{DX^k Y^T}{(Dg^k, Y)} \quad \text{et} \quad S^k = - \frac{H^k Dg^k Z^T}{(Dg^k, Z)}.$$

Comme  $H^{k+1}$  doit être symétrique, il vient :

$$R^k = \frac{[DX^k][DX^k]^T}{(Dg^k, DX^k)}, \quad S^k = - \frac{[H^k Dg^k][H^k Dg^k]^T}{(Dg^k, H^k Dg^k)}$$

ce qui donne

$$H^{k+1} = H^k + \frac{[DX^k][DX^k]^T}{(DX^k, DX^k)} - \frac{[H^k Dg^k][H^k Dg^k]^T}{(Dg^k, H^k Dg^k)}$$

matrice qui vérifie bien

$$H^{k+1} G DX^k = H^{k+1} Dg^k = DX^k.$$

b) Recherche d'un nouveau minimum relatif

Nous recherchons  $X^{k+1}$  tel que  $Q(X^{k+1}) < Q(X^k)$  à l'aide d'une direction  $D^k$ . Connaissant  $G^{-1}$  nous pourrions en une itération passer d'un point  $X^0$  au minimum  $X_m$  en utilisant la relation (8)

$$[g^0 - 0] = G[X^0 - X_m]$$

d'où

$$X_m = X_0 - G^{-1} g^0$$

nous prendrons donc

$$D^k = - H^k g^k.$$

Nous déterminerons alors un scalaire  $\ell^k$  définissant

$$X^{k+1} = X^k + DX^k = X^k - \ell^k H^k g^k$$

nous pouvons noter qu'en ce nouveau minimum

$$(D^k, g^{k+1}) = 0 = \frac{1}{\ell^k} (DX^k, g^{k+1}).$$

Développons  $Q(X^{k+1})$  en série de Taylor limitée aux premiers termes :

$$Q(X^{k+1}) - Q(X^k) = \sum_{i=1}^n DX_i^k \frac{\partial Q(X^k)}{\partial x_i} = (DX^k, g^k) = - \ell^k (H^k g^k, g^k).$$

Pour que cette méthode converge si  $Q(X)$  possède un minimum bien défini dans le domaine d'étude, il faudra que  $Q(X^{k+1}) \leq Q(X^k)$ . Si nous choisissons  $\ell^k$  positif, il suffira que  $(H^k g^k, g^k) \geq 0$ , donc que  $H^k$  soit une matrice définie positive.

$H^0$  étant définie positive, nous allons raisonner par induction et démontrer que si  $H^k$  est définie positive,  $H^{k+1}$  l'est aussi.  $H^k$  étant définie positive il existe deux vecteurs  $p$  et  $q$  tels que

$$p = [H^k]^{1/2} X^k \quad \text{et} \quad q = [H^k]^{1/2} Dg^k$$

$$\begin{aligned} (H^{k+1} X^k, X^k) &= (H^k X^k, X^k) + \frac{(DX^k, X^k)(DX^k, X^k)}{(DX^k, Dg^k)} - \frac{(H^k Dg^k, X^k)(H^k Dg^k, X^k)}{(H^k Dg^k, Dg^k)} \\ &= (p, p) + \frac{(X^k, DX^k)^2}{(DX^k, Dg^k)} - \frac{(p, q)(q, p)}{(q, q)} \\ &= \frac{(q, q)(p, p) - (q, p)^2}{(q, q)} + \frac{(X^k, DX^k)^2}{(DX^k, Dg^k)} . \end{aligned}$$

En tenant compte de l'inégalité de Schwartz :  $(p, q) \leq (p, p) + (q, q)$ , nous obtenons

$$(H^{k+1} X^k, X^k) \geq \frac{(X^k, DX^k)^2}{(DX^k, Dg^k)}$$

$H^{k+1}$  sera donc définie positive si  $(DX^k, Dg^k) > 0$

$$\begin{aligned} (DX^k, Dg^k) &= (DX^k, g^{k+1} - g^k) = - (DX^k, g^k) \quad \text{car} \quad (DX^k, g^{k+1}) = 0 \\ &= \ell^k (H^k g^k, g^k) . \end{aligned}$$

Comme  $\ell^k$  est positif et que  $H^k$  est définie positive,  $H^{k+1}$  est définie positive et  $Q(X^{k+1}) \leq Q(X^k)$ .

### c) Atteinte du minimum en n itérations

Nous avons vu que le nouveau minimum relatif de  $Q(X)$  est donné par :

$$X^{k+1} = X^k + DX^k = X^k - \ell^k H^k g^k ,$$

Nous allons montrer que  $DX^0, \dots, DX^n$  sont des vecteurs propres linéairement indépendants de la matrice  $H^{n+1}$ , associés aux valeurs propres unitaires et que  $g^{n+1}$  doit être orthogonal à  $DX^0, \dots, DX^n$ .

Nous avons déjà établi que pour tout  $i$

$$\begin{aligned} DX^i &= -\ell^i H^i g^i \\ Dg^i &= G DX^i \\ H^{i+1} Dg^i &= DX^i \\ (DX^i, g^{i+1}) &= 0 \end{aligned}$$

Nous allons démontrer que les relations

$$\begin{aligned} H^k G DX^i &= DX^i \\ (G DX^i, DX^k) &= (DX^i, G DX^k) = 0 \end{aligned}$$

sont vérifiées pour tout  $i$  tel que  $0 \leq i < k$ . Elles sont vérifiées pour  $k=1$  car par définition

$$\begin{aligned} H^1 G DX^0 &= DX^0 \\ (DX^0, g^1) &= 0 \\ &= (H^1 Dg^0, g^1) = (Dg^0, H^1 g^1) = -\frac{1}{\ell^1} (Dg^0, DX^1) \end{aligned}$$

$$\text{or } (Dg^i, DX^j) = (G DX^i, DX^j) = (DX^i, G DX^j) = (DX^i, Dg^j)$$

$$\text{donc } (DX^0, G DX^1) = 0.$$

Montrons qu'elles sont vérifiées pour  $k=2$ .

$$H^2 Dg^1 = DX^1 \text{ par définition.}$$

$$\begin{aligned} H^2 Dg^0 &= H^1 Dg^0 + \frac{DX^1 (DX^1, Dg^0)}{(DX^1, Dg^1)} - \frac{H^1 Dg^1 (H^1 Dg^1, Dg^0)}{(H^1 Dg^1, Dg^1)} \\ &= H^1 Dg^0 + (Dg^0, DX^1) \left\{ \frac{DX^1}{(DX^1, Dg^1)} - \frac{H^1 Dg^1}{(H^1 Dg^1, Dg^1)} \right\} \\ &= H^1 Dg^0 = DX^0 \end{aligned}$$



$$(DX^1, g^2) = 0 \text{ par définition.}$$

$$= (H^2 Dg^1, g^2) = -\frac{1}{\ell^2} (Dg^1, DX^2) = -\frac{1}{\ell^2} (DX^1, G DX^2)$$

$$(DX^0, g^2) = (DX^0, g^1) + (DX^0, Dg^1) = 0$$

$$= (H^2 Dg^0, g^2) = (Dg^0, H^2 g^2) = -\frac{1}{\ell^2} (DX^0, G DX^2) .$$

Montrons maintenant que si elles sont vérifiées pour  $k$ , elles le seront pour  $k+1$  :

$$H^{k+1} Dg^k = DX^k \text{ par définition.}$$

En prenant maintenant  $0 \leq j < k$

$$\begin{aligned} H^{k+1} Dg^j &= H^k Dg^j + \frac{DX^k(DX^k, Dg^j)}{(DX^k, Dg^k)} - \frac{H^k Dg^k(H^k Dg^k, Dg^j)}{(H^k Dg^k, Dg^k)} \\ &= H^k Dg^j + (DX^k, Dg^j) \left\{ \frac{DX^k}{(DX^k, Dg^k)} - \frac{H^k Dg^k}{(H^k Dg^k, Dg^k)} \right\} \\ &= H^k Dg^j = DX^j . \end{aligned}$$

Nous retrouvons comme précédemment

$$(DX^j, g^{k+1}) = -\frac{1}{\ell^{k+1}} (DX^j, G DX^{k+1})$$

$$(DX^k, g^{k+1}) = 0 \text{ par définition.}$$

En prenant maintenant  $0 \leq j < k$

$$g^{k+1} = g^k + Dg^k = \dots = g^{j+1} + \sum_{m=j+1}^k Dg^m$$

d'où

$$(DX^j, g^{k+1}) = (DX^j, g^{j+1}) + \sum_{m=j+1}^k (DX^j, Dg^m) = 0 .$$

Ces relations nous donnent pour  $k = n+1$

$$H^{n+1} G DX^i = DX^i \quad 0 \leq i < n+1$$

$$(DX^i, g^{n+1}) = 0.$$

La première relation montre que  $H^{n+1} G$  est une matrice unitaire. La deuxième impose à  $g^{n+1}$  d'être nul car il doit être orthogonal aux  $n$  vecteurs linéairement indépendants  $DX^0, \dots, DX^n$ .

#### 4) Méthode des directions conjuguées - Fonction quelconque

Dans le cas d'une fonction quelconque  $\Omega(X)$ , le développement en série de Taylor limité aux premiers termes nous donne :

$$\Omega(X^{k+1}) - \Omega(X^k) = \sum_{i=1}^n DX_i^k \frac{\partial \Omega(X^k)}{\partial x_i} = (DX^k, g^k)$$

soit en reprenant les mêmes notations que précédemment :

$$\Delta \Omega = - \ell^k (H^k g^k, g^k)$$

$\Omega(X^{k+1}) < \Omega(X^k)$  si  $\ell^k$  est positif car  $H^k$  est définie positive comme nous l'avons démontré.

Si dans le domaine  $\Omega(X)$  possède un minimum bien défini, la méthode restera convergente; de plus, près d'un minimum,  $\Omega(X)$  sera assimilable à une fonction quadratique définie positive d'où l'intérêt de cette méthode.

#### 5) Utilisation de $n$ directions orthogonales

Etant donné un système de  $n$  directions orthogonales :  $D_1^k, \dots, D_n^k$ , nous allons chercher tour à tour le long de chaque direction un nouveau minimum relatif de  $\Omega(X)$  et nous recommencerons ce cycle jusqu'à ce que l'on change de système, ce qui doit permettre d'accélérer la décroissance de la fonction. Contrairement à la méthode de relaxation variable par variable nous agissons sur toutes les composantes de  $X$  car les directions ne seront pas, en général, parallèles aux axes  $Ox_1, \dots, Ox_n$ .

a) Construction d'un système de vecteurs orthonormés

Le même système de directions se représente par un système de vecteurs orthonormés, soit :

$$D_1^k, \dots, D_n^k \text{ tels que } (D_i^k, D_j^k) = \delta_{ij}.$$

Pour déterminer un système  $D_i^{k+1}$ , la méthode la plus générale consiste à choisir un système de  $n$  vecteurs linéairement indépendants  $V_i^k$  :

$$V_1^k, \dots, V_n^k$$

et de former

$$D_1^{k+1} = V_1^k$$

$$D_2^{k+1} = V_2^k - a_{21} D_1^{k+1}$$

.....

$$D_n^{k+1} = V_n^k - \sum_{i=1}^{n-1} (a_{ni} D_i^{k+1})$$

en calculant les  $a_{ij}$  de telle façon que  $(D_i^{k+1}, D_j^{k+1}) = 0$  si  $i \neq j$ . Nous avons immédiatement pour  $i=2$  et  $j=1$  :

$$a_{ij} = \frac{(V_i^k, D_j^{k+1})}{(D_j^{k+1}, D_j^{k+1})} \text{ avec } i > j.$$

Nous simplifierons l'écriture des  $a_{ij}$  si nous normons les vecteurs  $D_i^k$ ; en effet si  $(D_i^{k+1}, D_j^{k+1}) = \delta_{ij}$

$$a_{ij} = (V_i^k, D_j^{k+1}) \quad i > j,$$

d'où le procédé :

ORTHOGONALISATION

puis

NORMALISATION

$$B_1 = V_1^k$$

$$D_1^{k+1} = \frac{B_1}{(B_1, B_1)^{1/2}}$$

$$B_2 = V_2^k - (V_2^k, D_1^{k+1}) D_1^{k+1}$$

$$D_2^{k+1} = \frac{B_2}{(B_2, B_2)^{1/2}}$$

.....

.....

$$B_n = V_n^k - \sum_{i=1}^{n-1} (V_n^k, D_i^{k+1}) D_i^{k+1}$$

$$D_n^{k+1} = \frac{B_n}{(B_n, B_n)^{1/2}}$$

Ces nouvelles directions  $D_i^{k+1}$  vont dépendre essentiellement du choix du système de vecteurs  $V_i^k$  linéairement indépendants.

b) Choix du système de vecteurs linéairement indépendants

Pour le même système de directions, désignons par

$D_1^k, \dots, D_n^k$  les n vecteurs orthornormés représentant les directions;

$e_1^k, \dots, e_n^k$  la somme des déplacements effectués sur chaque direction;

$X^k$  le minimum relatif initial;

$X^{k+p}$  le minimum relatif final.

Le déplacement du minimum dans ce système de directions a pour expression :

$$DX^k = X^{k+p} - X^k = \sum_{i=1}^n e_i^k D_i^k = \sum_{i=1}^n DX_i^k$$

où  $DX_i^k$  représente le déplacement obtenu à l'aide de la direction  $D_i^k$ .

Nous allons tenir compte de ces déplacements pour construire le système :

$$V_1^k = \sum_{i=1}^n DX_i^k = DX^k$$

$$V_2^k = V_1^k - DX_1^k$$

.....

$$V_n^k = V_{n-1}^k - DX_{n-1}^k = DX_1^k .$$

Ce système sera linéairement indépendant si aucun des déplacements  $DX_i^k$  n'est nul donc si tous les  $e_i^k$  sont différents de zéro.

Différentes possibilités de méthodes s'offriront suivant le choix que l'on fera pour

- 1°/ La méthode de minimisation le long d'une direction.
- 2°/ Le test nous permettant de changer de système de directions.

### CHAPITRE III

#### TECHNIQUES UTILISEES DANS LES METHODES COMPAREES

L'étude directe du système  $F(X) = 0$  a été faite à l'aide de la méthode de linéarisation.

Parmi les différentes méthodes possibles de minimisation, nous nous sommes limités à utiliser chacune des directions énoncées afin de faire une comparaison et nous avons associé la technique de minimisation d'une fonction à une variable qui semblait la plus appropriée.

Toutes ces méthodes ont été mises sous forme de fonctions FORTRAN très simples à utiliser (positives si la méthode converge, négatives si elle diverge). Afin de leur conférer une grande flexibilité, nous les avons divisées en sous-programmes facilement interchangeables. Nous ne donnerons dans cette description que les détails d'ordre technique et un organigramme si besoin est.

#### A. METHODE DE LINEARISATION

Cette méthode donne avec les notations employées précédemment :

$$X^{k+1} = X^k + H^k$$

où  $H^k$  est solution du système linéaire approché

$$DF(X^k) H^k = - F(X^k) .$$

La résolution de ce système par la méthode de substitution, d'une programmation facile, nous conduira au système équivalent

$$IH^k = A^k$$

d'où nous tirerons immédiatement la solution  $h_i^k = a_i^k$   $i=1,2,\dots,n$ .

Voyons cette méthode sur un exemple simple à trois dimensions :

$$DH = -F \text{ qui s'écrit } \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = - \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} .$$

Multiplier tous les éléments d'une même ligne de D et F par un même scalaire ou ajouter terme à terme deux lignes ne change rien au système, du fait de sa linéarité en  $h_i$

$$\begin{bmatrix} \alpha d_{11} & \alpha d_{12} & \alpha d_{13} \\ d_{21} - \beta d_{11} & d_{22} - \beta d_{12} & d_{23} - \beta d_{13} \\ d_{31} - \gamma d_{11} & d_{32} - \gamma d_{12} & d_{33} - \gamma d_{13} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = - \begin{bmatrix} \alpha f_1 \\ f_2 - \beta f_1 \\ f_3 - \gamma f_1 \end{bmatrix} .$$

Si l'on prend  $\alpha = 1/d_{11}$ ,  $\beta = d_{21}/d_{11}$ ,  $\gamma = d_{31}/d_{11}$ , la première colonne s'annule à l'exception du terme diagonal qui devient unitaire. Par une telle suite de transformations des colonnes, nous réduisons le système initial au système équivalent

$$IH = A .$$

Nous voyons qu'il suffit d'étudier la matrice équivalente de dimensions  $(n, n+1)$  :

$$[D | -F] \text{ que l'on réduit à } [1 | A] ,$$

la dernière colonne fournissant la solution.

Cette méthode perd cependant toute précision si, en cours d'opération, le terme diagonal que l'on veut rendre unitaire est trop petit. Une légère amélioration consistera à prendre comme jème terme diagonal le kème élément de la colonne j qui possède la plus grande valeur absolue. Nous avons imposé  $k \geq j$ , car il suffit alors d'interchanger les lignes j et k sans rien changer à l'ordre des variables.

Nous recherchons une solution intérieure à un certain domaine  $\mathcal{D}$  que nous nous fixons; si le point  $X^{k+1}$  ainsi déterminé est extérieur à ce domaine, nous réduisons de moitié  $H^k$  jusqu'à ce que l'on obtienne satisfaction :

$$X^{k+1} = X^k + H^{k+1}/_2P .$$

Ne pouvant contrôler en cours d'opération la marche des calculs, nous additionnons une condition de sécurité arrêtant les calculs dans le cas d'une divergence. Nous associons au système la fonction

$$\delta^k = \sum_{i=1}^n f_i^2(X^k)$$

et acceptons  $X^{k+1}$  comme meilleure approximation de la solution si :

$$\delta^{k+1} \leq \delta^k .$$

Dans le cas contraire nous réduisons  $H^k$  comme précédemment ce qui doit permettre, si la méthode est convergente, de trouver une meilleure approximation dès que l'on est assez près de  $X^k$ . Nous limiterons cependant pour chaque itération le nombre total de réductions de  $H^k$ .

Le point  $X^{k+1}$  sera accepté comme solution du système (1) :

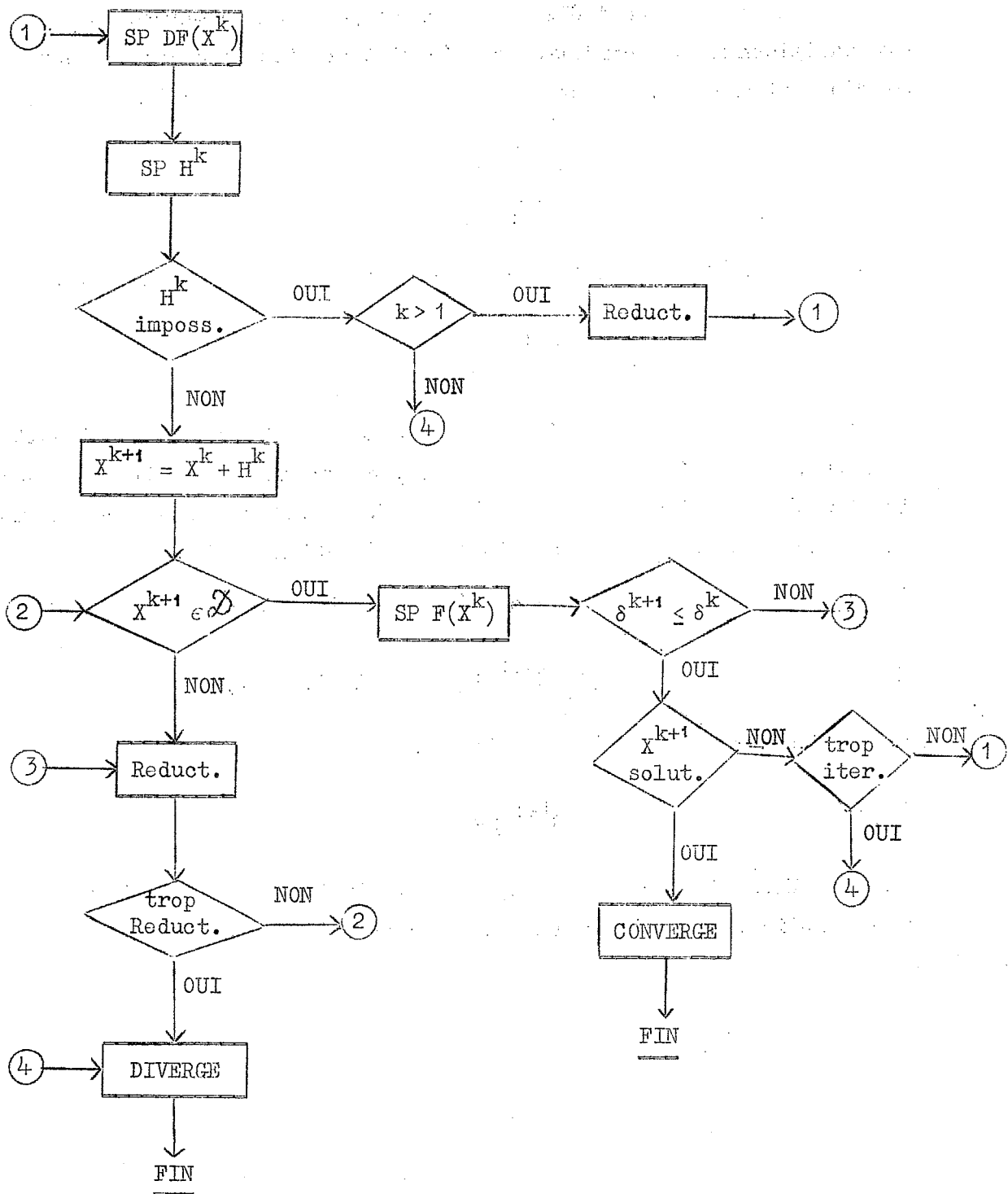
$$\text{si} \quad |f_i(X^{k+1})| \leq \epsilon_1 \quad i = 1, 2, \dots, n .$$

Nous y ajoutons une deuxième condition :

$$\text{ou si} \quad \delta^{k+1} \leq \epsilon_2 .$$

Nous limiterons le nombre total d'itérations pour l'obtention d'une solution, d'où l'organigramme de cette méthode :





## B. METHODE DE RELAXATION VARIABLE PAR VARIABLE

Nous avons, parmi les deux possibilités de méthodes de relaxation, choisi celle-ci car nous ferons la méthode du gradient. Le nouveau minimum relatif

$$X^{k+1} = X^k + \ell^k D^k$$

est recherché dans une direction  $D^k$  prise parallèle tour à tour à chacun des axes  $Ox_1, \dots, Ox_n$ , ce qui ne fait varier qu'une composante de  $X$  à la fois.

Ce procédé s'effectuant de nombreuses fois, nous avons pris une interpolation par une courbe du second degré qui est assez rapide. Les valeurs des dérivées première et seconde en un point ne tenant compte que de la variation de  $f(\ell)$  en ce point, nous avons pris trois points équidistants de

$$h = - \frac{1}{2} \left[ \frac{f(\ell)}{f'(\ell)} \right]_{\ell=0}$$

ce qui permettra de tenir compte de la variation de  $f(\ell)$  sur un certain intervalle.  $D^k$  étant parallèle à l'un des axes  $Ox_j$ , ceci nous donne en prenant une approximation numérique de la dérivée

$$h_j = - \frac{1}{2} \frac{\Omega(X^k)}{\frac{\partial \Omega(X^k)}{\partial x_j}} = - dx_j \frac{\Omega(X^k)}{\Omega(x_1^k, \dots, x_j^k + dx_j, \dots, x_n^k) - \Omega(X^k)}$$

L'interpolation perdrait toute précision si près d'un minimum nous prenions un pas d'interpolation trop grand ou si au contraire loin d'un minimum le pas était trop petit. Nous avons essayé de remédier à cet inconvénient en prenant  $h_j = dx_j$  si  $f'(0)$  devient trop petit et  $h_j = \text{MAXIMUM} \{ |dx_j|, |h_j| \}$  dans l'autre cas: nous nous ramenons alors à l'évaluation d'un minimum au moyen des dérivées première et seconde au point  $X^k$ .

Ayant ainsi quatre valeurs de  $\ell$ , nous prendrons comme minimum celui des quatre points donnant à  $\Omega$  la plus petite valeur. Si ce minimum correspond à  $X^k$  nous réduirons ou augmenterons l'intervalle d'interpolation suivant que  $h_j$  était supérieur ou égal à  $dx_j$ .

Appelons cycle la recherche de  $n$  minima sur l'ensemble des axes. Soit  $X^k$  le point initial,  $X^{k+n}$  le point final d'un tel cycle et  $H^k = X^k - X^{k+n}$  le déplacement total. Le point  $X^{k+n}$  sera jugé minimum de  $\Omega(X)$  :

si  $|x_i^k - x_i^{k+n}| \leq \epsilon_1$  MAXIMUM  $\{10^{-p}, |x_i^k|\}$   $i = 1, 2, \dots, n$

ou si  $(H^k, H^k)^{1/2} \leq \epsilon_2$ .

En prenant  $\epsilon_1 = 10^{-N_1}$  nous aurons ainsi la possibilité d'obtenir  $N_1$  chiffres significatifs pour  $X$  si tous les  $|x_i| > 10^{-p}$ . Nous avons ajouté le deuxième test qui, lui, permettra de tester la longueur du vecteur déplacement au cours d'un cycle.

### C. METHODE DU GRADIENT

Le point  $X^{k+1} = X^k + \ell^k D^k$  est obtenu à l'aide de la direction

$$D^k = - \text{grad}[\Omega(X^k)] = - g^k$$

Le calcul du gradient étant nécessaire, nous interpolerons  $f(\ell)$  par un polynôme du troisième degré à l'aide des valeurs  $f(\ell)$  et  $f'_\ell(\ell)$  en deux points, ce qui donnera de la précision pour la recherche du minimum. Le calcul de  $f'_\ell(\ell)$  se présente comme le produit scalaire :

$$f'_\ell(\ell) = (D^k, \text{grad}[\Omega(X^k + \ell D^k)])$$

Nous prendrons des précautions pour le calcul de  $\ell^k$ . Nous déterminerons tout d'abord un intervalle  $[0, \ell_M]$  dans lequel est compris  $\ell^k$ .  $D^k$  étant la direction opposée du gradient la dérivée de  $f$  sera négative, au point  $\ell = 0$  :

$$f'(0) = - (g^k, g^k)$$

Il nous faudra trouver  $\ell_M$  tel que

$$f'(\ell_M) > 0$$

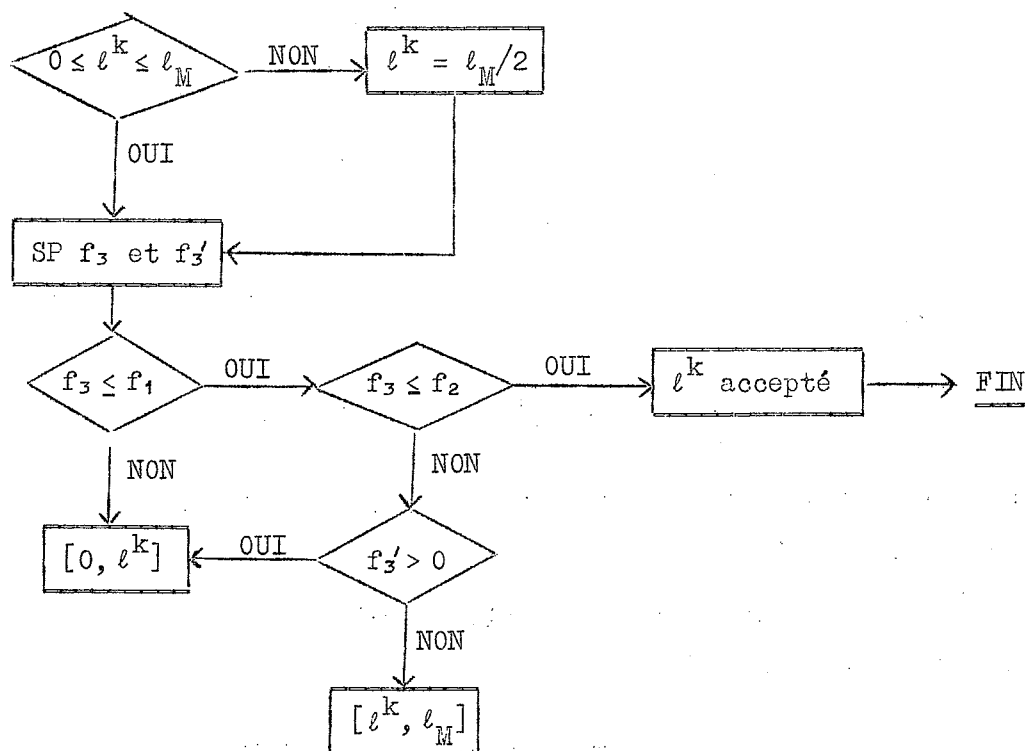
ou  $f(\ell_M) > f(0) = f_1$  si  $f'(\ell_M) < 0$ .

Nous prendrons comme valeur initiale de  $\ell_M$  la moitié de l'intervalle nécessaire à annuler  $f(\ell)$  supposée linéaire en première approximation. Nous limiterons cependant cette valeur initiale au cas où la dérivée de  $f$  serait trop petite, ce qui donne

$$\ell_M = \text{MAXIMUM} \left\{ C, -2 \frac{f(0)}{f'(0)} \right\}.$$

Si  $\ell_M$  ne répond pas aux conditions imposées nous le doublerons ou le diminuerons suivant la valeur de  $f(\ell_M)$  jusqu'à ce que l'on obtienne satisfaction. Nous aurons alors les valeurs de  $f_2$  et  $f_2'$  et pourrons faire l'interpolation.

Nous calculerons  $\ell^k$  d'où  $f_3$  et  $f_3'$ . Si  $f_3$  est inférieur à  $f_1$  et  $f_2$  le point  $X^{k+1} = X^k + \ell^k D^k$  sera accepté comme nouveau minimum relatif de  $\Omega(X)$ . Dans le cas contraire, nous recommencerons l'interpolation sur l'intervalle  $[0, \ell^k]$  ou  $[\ell^k, \ell_M]$  suivant les cas rencontrés que nous donnons dans l'organigramme suivant. Nous limiterons cependant le nombre d'interpolations pour la recherche d'une valeur de  $\ell^k$ .



Au voisinage d'une solution,  $DX^k = X^{k+1} - X^k$  doit tendre vers zéro, d'où les deux tests

$$|x_i^{k+1} - x_i^k| \leq \epsilon_1 \cdot \text{MAXIMUM} \{10^{-p}, |x_i^k|\} \quad i = 1, 2, \dots, n$$

pour obtenir le nombre de chiffres significatifs désiré,

ou 
$$(DX^k, DX^k)^{1/2} \leq \epsilon_2$$

pour tester la longueur du déplacement. Nous y ajouterons une condition supplémentaire de sécurité : au voisinage d'un minimum le gradient va lui aussi tendre vers zéro, d'où les deux autres tests :

$$|D_i^k| \leq \epsilon_1 \quad i = 1, 2, \dots, n$$

ou 
$$(D_i^k, D_i^k)^{1/2} \leq \epsilon_2$$

#### D. METHODE DES DIRECTIONS CONJUGUEES

Le point  $X^{k+1}$  est trouvé à l'aide de la direction

$$D^k = -H^k g^k$$

où  $H^k$  est une matrice calculée à chaque itération. En prenant  $H^0 = I$ ,  $D^0$  correspondra à la direction de plus grande pente.

Le calcul du gradient étant nécessaire, nous reprendrons la même méthode de minimisation que pour la méthode du gradient :

$$f'_\ell(0) = (D^k, g^k) = - (H^k g^k, g^k) .$$

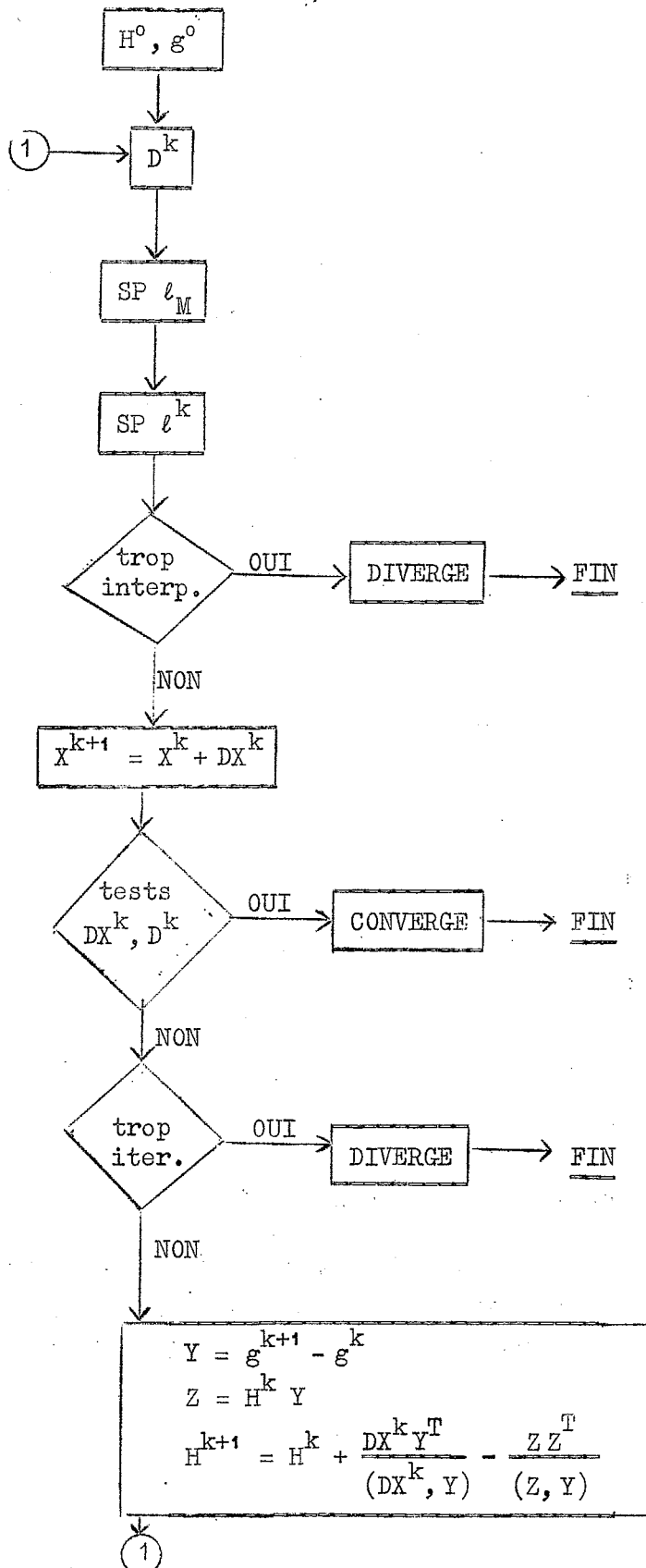
Comme  $H^k$  est définie positive  $f'_\ell(0)$  est encore négative et nous retrouvons exactement le même procédé.

Les tests pour la solution se révèlent identiques; en effet, au voisinage d'une solution

$$DX^k = X^{k+1} - X^k \text{ va tendre vers zéro}$$

et 
$$D^k = -H^k g^k \text{ aussi, car } g^k \text{ s'annulera.}$$

La matrice  $H^k$  étant symétrique, il nous suffira d'en calculer la moitié. D'où l'organigramme de cette méthode valable pour la méthode du gradient en y enlevant les calculs relatifs à  $H^k$ .



## E. UTILISATION DE N DIRECTIONS ORTHOGONALES

Nous cherchons tour à tour sur chaque direction un nouveau minimum relatif de  $\Omega(X)$  et recommençons ce cycle jusqu'à ce que l'on change de système de directions. La première méthode exposée cherche les minima en utilisant une interpolation par une courbe du second degré; la deuxième au contraire procédera par tâtonnements successifs ce qui en fait une méthode très générale car elle ne demande aucun calcul de dérivées.

Dans les deux méthodes exposées nous avons pris comme système initial de directions le système d'axes  $Ox_1, \dots, Ox_n$ .

La construction des nouvelles directions  $D_i^{j+1}$  pourra être allégée; en reprenant la notation utilisée dans le chapitre II nous construisons le système orthogonal des  $B_i$  par :

$$B_i = V_i^j - \sum_{k=1}^{i-1} (V_i^j, D_k^{j+1}) D_k^{j+1}$$

que l'on pourra mettre sous la notation abrégée

$$B_i = V_i^j - V_i^j DC_i^j,$$

avec la relation de récurrence

$$DC_i^j = \sum_{k=1}^{i-1} D_k^{j+1} [D_k^{j+1}]^T = DC_{i-1}^j + D_i^j [D_i^j]^T.$$

Cette matrice d'ordre  $(n, n)$  étant symétrique, il nous suffira d'en calculer la moitié.

### 1) Interpolation par une courbe du second degré

Nous minimiserons la fonction  $f_i(\ell) = \Omega(X + \ell D_i^j)$  par rapport à la variable  $\ell$  en faisant une interpolation de  $f_i(\ell)$  par un polynôme du second degré. Pour cela nous utiliserons trois points équidistants de :

$$h = - \frac{1}{2} \left[ \frac{f_i(0)}{\frac{\partial f_i(0)}{\partial \ell}} \right]$$

pour les mêmes raisons que dans la méthode de relaxation variable par variable. Supposons que le minimum de  $f_i(\ell)$  sur la direction  $D_i^j$  serve à déterminer le point  $X^{k+1}$ , nous prendrons alors comme pas d'interpolation

$$h_i = - \frac{1}{2} \frac{\frac{\Omega(X^k)}{\frac{\partial \Omega(X^k)}{\partial \ell}}}{\frac{\Omega(X^k + \Delta \ell D_i^j) - \Omega(X^k)}{\Delta \ell}}.$$

Nous essayerons encore de remédier aux imprécisions dues à un pas  $h_i$  trop petit ou trop grand lorsque l'on est loin ou près d'un minimum. Ayant obtenu ainsi quatre valeurs pour  $\ell$ , nous prendrons comme valeur  $\ell^k$  celle qui minimise le mieux  $f_i(\ell)$ . Si cette valeur est nulle, nous recommencerons l'interpolation sur un intervalle supérieur ou inférieur suivant que  $h_i$  était égal ou supérieur à  $\Delta \ell$ . Nous ajouterons au déplacement total  $e_i^j$  sur la direction  $D_i^j$  la valeur  $\ell^k$  et chercherons alors sur la direction  $D_{i+1}^j$  la valeur  $\ell^{k+1}$  minimisant  $f_{i+1}(\ell) = \Omega(X^{k+1} + \ell D_{i+1}^j)$  qui déterminera  $X^{k+2}$ , etc...

Nous avons vu que les vecteurs  $V_i^{j+1}$  étaient linéairement indépendants si tous les  $e_i^j$  étaient non nuls. Appelons cycle la recherche de  $n$  minima sur  $D_1^j, \dots, D_n^j$ . Nous changerons le système de directions  $D_i^j$  à la fin d'un cycle dès que, tous les  $e_i^j$  étant différents de zéro, nous rencontrerons l'un des cas suivants :

1°/ Un des  $\ell^k$  est nul au cours de ce cycle.

2°/ Nous avons effectué plus de 3 cycles sans changer de système de directions (cette condition a été ajoutée pour éviter de refaire la méthode de relaxation variable par variable ou une méthode analogue avec les directions  $D_i^0$  ou  $D_i^j$ ).

Soit  $X^k$  et  $X^{k+m}$  les minima initial et final d'un système  $D_i^j$ , le point  $X^{k+m}$  sera jugé minimum de  $\Omega(X)$ ,



si  $|x_i^{k+m} - x_i^k| \leq \epsilon_1$ . MAXIMUM  $\{10^{-p}, |x_i^k|\}$   $i = 1, 2, \dots, n$

ou si  $(H^j, H^j)^{1/2} \leq \epsilon_2$  avec  $H^j = X^{k+m} - X^k$ .

Nous avons repris les mêmes tests que pour les méthodes précédentes.

## 2) Méthode d'investigations<sup>7)</sup>

Nous rechercherons encore les nouveaux minima relatifs de  $\Omega$  en prenant tour à tour chaque direction  $D_i^j$  et recommencerons ce cycle jusqu'à ce que l'on change de système de directions. La minimisation toutefois sera faite à l'aide d'une méthode d'investigation. On se fixera au début du programme  $n+2$  constantes  $\alpha, \beta, h_1, \dots, h_n$  et, au début de chaque changement de système, nous donnerons aux pas initiaux sur chaque directions  $D_i^j$  ces valeurs  $h_i$ . Supposons maintenant que  $X^{k+1}$  soit déterminé à l'aide de la direction  $D_i^j$  possédant à ce moment là un pas déterminé  $h_i$ , nous regarderons si  $\Omega(X^k + h_i D_i^j)$  est inférieur à  $\Omega(X^k)$ .

Si oui, nous prendrons comme nouveau minimum relatif de  $\Omega$  :

$$X^{k+1} = X^k + \ell^k D_i^j$$

avec  $\ell^k = h_i$  et nous ajouterons au déplacement total  $e_i^j$  sur  $D_i^j$  la valeur de  $h_i$ . Nous multiplierons ensuite le pas de façon à avoir au prochain cycle un pas  $h_i \alpha$  sur la direction  $D_i^j$ .

Dans le cas contraire nous prendrons  $\ell^k = 0$  et au cours du prochain cycle le pas sur la direction  $D_i^j$  sera égal à  $-h_i \cdot \beta$ . Nous déterminerons  $X^{k+2}$  à l'aide de  $h_{i+1}$  et  $D_{i+1}^j$ , etc.

Le changement de système de directions sera opéré à la fin d'un cycle dès que, tous les  $e_i^j$  étant non nuls, nous trouverons au cours de ce cycle un des  $\ell^k$  égal à zéro.

Nous avons repris les mêmes tests que dans la méthode précédente pour déterminer si le dernier minimum relatif trouvé à l'aide d'un système  $D_i^j$  pouvait être considéré comme minimum de  $\Omega(X)$ .

## CHAPITRE IV

### RESULTATS DES METHODES COMPAREES

Toutes ces méthodes ont été programmées en langage FORTRAN II et exécutées sur l'ordinateur IBM 7090 de l'Organisation Européenne pour la Recherche Nucléaire à Genève. Chaque méthode ayant des opérations intermédiaires différentes (inversion d'une matrice pour la méthode de linéarisation, calcul du gradient et de la matrice  $H^k$  pour la méthode des directions conjuguées, détermination des nouvelles directions pour l'utilisation de  $n$  directions orthogonales), il nous a semblé plus équitable de faire, dans la plupart des cas, les comparaisons des temps d'exécution  $t$  qui seront donnés en secondes à  $3/100$  de seconde près.

Travaillant sur un calculateur très rapide nous avons limité le nombre d'itérations ou d'évaluations de  $\Omega$ . Dans les résultats que nous donnons le nombre  $k$  représentera des itérations pour les méthodes de linéarisation, du gradient, des directions conjuguées et des cycles pour les trois autres méthodes de recherche multidirectionnelle. Si l'on dépasse le maximum que l'on se fixe pour  $k$ , nous le soulignerons : soit  $\underline{k}$ . La limitation du nombre d'évaluations de  $\Omega$  était nécessaire pour arriver, dans le cas d'une convergence lente, à des temps d'exécution du même ordre de grandeur pour toutes les méthodes de minimisation que nous avons numérotées :

- I Relaxation variable par variable
- II Gradient
- III Directions conjuguées
- IV  $n$  directions orthogonales et minimisation par interpolations
- V " " " " " " investigations

Ce chapitre a été divisé en deux parties : l'une est consacrée à la résolution de systèmes d'équations, l'autre à la minimisation d'une fonction  $\Omega$  positive ou nulle.

## A. RESOLUTION DE SYSTEMES D'EQUATIONS

Nous distinguerons deux types de fonctions suivant que l'on peut ou non connaître facilement leurs dérivées analytiques.

### 1) Premier type de fonctions

Afin de pouvoir construire de grands systèmes non linéaires, nous utilisons un sous-programme choisissant des nombres au hasard dans l'intervalle  $[0,1]$ , ce qui nous permet de déterminer des matrices  $A$ ,  $B$  et  $\alpha$ ,  $\delta$ , dont les éléments sont choisis au hasard dans les intervalles  $[-100, +100]$  et  $[-\pi, +\pi]$ .

Nous avons calculé la matrice  $E$  :

$$E_i = \sum_{j=1}^n (A_{ij} \sin \alpha_j + B_{ij} \cos \alpha_j) \quad i = 1, 2, \dots, n.$$

Ayant alors les valeurs de toutes ces constantes, nous avons étudié le système de  $n$  équations :

$$F_i(x_1, \dots, x_n) = E_i - \sum_{j=1}^n (A_{ij} \sin x_j + B_{ij} \cos x_j) \quad i = 1, 2, \dots, n,$$

dont la solution  $X = \alpha$  est cherchée en partant de la valeur initiale :

$$X = \alpha + \delta/10.$$

Un premier essai a montré que seules les méthodes de linéarisation et des directions conjuguées aboutissaient rapidement à des résultats précis et nous avons testé ces méthodes pour des systèmes allant jusqu'à 50 équations.

Nous comparerons ensuite la rapidité de convergence des méthodes du gradient et des directions conjuguées qui utilisent toutes deux une recherche unidirectionnelle liée au gradient.

Un troisième tableau fournira les résultats des méthodes de minimisation, exception faite de la méthode des directions conjuguées, pour des temps de calcul sensiblement égaux.

a) Méthodes de linéarisation et des directions conjuguées

Nous utilisons les dérivées analytiques; un troisième essai a été fait en approchant numériquement les dérivées dans la première méthode afin de voir si la rapidité de convergence était modifiée.

Nous signalons toutefois que la convergence de cette méthode peut devenir plus lente si le calcul des dérivées n'est pas effectué avec certaines précautions; supposons qu'au point  $X^{k+1} = X^k + H^k$ , les dérivées approchées soient données par :

$$\frac{\partial f_i(X^{k+1})}{\partial x_j} = \frac{\Delta f_i(X^{k+1})}{\Delta x_j}.$$

Pour conserver une bonne convergence il faudra que  $\Delta x_j < h_j^k$ . En effet, prenons le cas suivant pour un exemple à une dimension :

| k | $x^k$   | $\Delta x = 10^{-2} \cdot  x^k $ |
|---|---------|----------------------------------|
| 1 | 10      | 0,1                              |
| 2 | 1,50000 | 0,015                            |
| 3 | 1,49982 | 0,015                            |
| 4 | 1,49976 | 0,015                            |

nous voyons qu'à partir de la troisième itération l'accroissement de  $x$  devient inférieur à  $\Delta x$ , ce qui enlève toute précision à la méthode.

Nous notons dans les tableaux de résultats :

- $n$  = nombre d'équations,
- $\delta$  = fonction associée au système,
- $\Omega$  = fonction à minimiser,
- $t$  = temps d'exécution.

La solution est jugée à l'aide des constantes  $\epsilon_1 = \epsilon_2 = 10^{-4}$ . La méthode de linéarisation a comme test de convergence :

- 1°/ moins de 30 réductions de  $H$  pour une itération,
- 2°/ limitation du nombre d'itérations.

Si l'on dépasse le nombre maximum de réductions de  $H$ , nous étoilons  $k$  : soit  $k^*$ .

| n  | $\delta^0 = \Omega^0$ | DERIV. APPR. |                       | DERIV. ANAL. |                       |       | DIRECT. CONJUGUEES |                       |        |
|----|-----------------------|--------------|-----------------------|--------------|-----------------------|-------|--------------------|-----------------------|--------|
|    |                       | k            | $\delta^k$            | k            | $\delta^k$            | t     | k                  | $\Omega^k$            | t      |
| 5  | 890                   | 3            | $0,71 \cdot 10^{-10}$ | 3            | $0,27 \cdot 10^{-10}$ | 0,18  | 9                  | $0,63 \cdot 10^{-10}$ | 0,42   |
| 5  | 1200                  | 6            | $0,17 \cdot 10^{-8}$  | 6            | $0,44 \cdot 10^{-8}$  | 0,28  | 13                 | $0,52 \cdot 10^{-11}$ | 0,93   |
| 10 | 12700                 | 4            | $0,29 \cdot 10^{-8}$  | 4            | $0,10 \cdot 10^{-9}$  | 0,55  | 17                 | $0,15 \cdot 10^{-6}$  | 2,22   |
| 10 | 3900                  | 6            | $0,14 \cdot 10^{-9}$  | 5            | $0,43 \cdot 10^{-7}$  | 0,65  | 16                 | $0,44 \cdot 10^{-8}$  | 2,02   |
| 20 | 27000                 | 6            | $0,14 \cdot 10^{-8}$  | 5            | $0,60 \cdot 10^{-9}$  | 2,63  | 26                 | $0,59 \cdot 10^{-6}$  | 10,03  |
| 20 | 21400                 | 15*          | 790                   | 15*          | 920                   | 15,17 | 28                 | $0,12 \cdot 10^{-7}$  | 10,65  |
| 20 | 33900                 | 5            | $0,96 \cdot 10^{-8}$  | 5            | $0,41 \cdot 10^{-9}$  | 2,58  | 29                 | $0,28 \cdot 10^{-6}$  | 11,65  |
| 20 | 24900                 | 13           | $0,50 \cdot 10^{-8}$  | 21           | $0,66 \cdot 10^{-8}$  | 13,85 | 34                 | $0,15 \cdot 10^{-2}$  | 15,95  |
| 30 | 145000                | 23*          | 1600                  | 30           | 2200                  | 69,98 | 36                 | $0,19 \cdot 10^{-3}$  | 29,17  |
| 30 | 89200                 | 8            | $0,13 \cdot 10^{-8}$  | 8            | $0,16 \cdot 10^{-8}$  | 11,02 | 44                 | $0,12 \cdot 10^{-8}$  | 40,32  |
| 30 | 67400                 | 4            | $0,94 \cdot 10^{-8}$  | 4            | $0,25 \cdot 10^{-8}$  | 5,57  | 36                 | $0,15 \cdot 10^{-6}$  | 29,73  |
| 30 | 78900                 | 4            | $0,22 \cdot 10^{-8}$  | 4            | $0,28 \cdot 10^{-8}$  | 5,57  | 38                 | $0,65 \cdot 10^{-7}$  | 30,98  |
| 50 | 240600                | 16           | $0,13 \cdot 10^{-7}$  | 13           | $0,15 \cdot 10^{-7}$  | 71,15 | 63                 | $0,33 \cdot 10^{-7}$  | 145,97 |
| 50 | 330100                | 8            | $0,15 \cdot 10^{-7}$  | 9            | $0,16 \cdot 10^{-7}$  | 46,88 | 68                 | $0,39 \cdot 10^{-7}$  | 166,18 |

Pour cet exemple général la méthode de linéarisation, plus rapide que la méthode des directions conjuguées, rencontre deux cas de divergence tandis que l'autre méthode nous fournit la solution, dans tous les cas. Le fait de prendre les dérivées exactes ou approchées ne change rien à la rapidité de convergence de la méthode de linéarisation et nous retrouvons les deux mêmes points de divergence qui sont dus à une matrice  $DF(X^k)$  devenant mal conditionnée, ce qui enlève toute précision au calcul de  $H^k$  et les tests de convergence nous renvoient à chaque itération vers  $X^k$ .

#### b) Méthodes du gradient et des directions conjuguées

Nous utilisons les dérivées analytiques et comparons ces deux méthodes avec un même nombre maximum d'itérations. Les valeurs de  $\Omega^0$  sont les mêmes que précédemment.

| GRADIENT |           |            |      | DIRECT. CONJUGUEES |                       |       |
|----------|-----------|------------|------|--------------------|-----------------------|-------|
| n        | k         | $\Omega^k$ | t    | k                  | $\Omega^k$            | t     |
| 5        | <u>10</u> | 4,44       | 0,47 | 9                  | $0,63 \cdot 10^{-10}$ | 0,46  |
| 5        | <u>10</u> | 0,107      | 0,47 | <u>10</u>          | $0,11 \cdot 10^{-5}$  | 0,58  |
| 10       | <u>20</u> | 9,68       | 1,92 | 17                 | $0,15 \cdot 10^{-6}$  | 2,25  |
| 10       | <u>20</u> | 7,97       | 1,92 | 16                 | $0,44 \cdot 10^{-8}$  | 2,03  |
| 20       | <u>30</u> | 20,01      | 8,50 | 26                 | $0,59 \cdot 10^{-6}$  | 10,05 |

La méthode des directions conjuguées présente une convergence bien plus rapide, les temps étant du même ordre.

### c) Méthodes de minimisation

Nous avons voulu comparer les méthodes à temps sensiblement égaux ce qui fait que les limitations pour chaque méthode sont différentes. Le nombre maximum d'évaluations de  $\Omega$  était nécessaire pour les méthodes utilisant n directions orthogonales; si ce nombre est dépassé nous étions k : soit  $k^*$ .

Les minima sont jugés avec  $\epsilon_1 = \epsilon_2 = 10^{-4}$  tandis que la méthode d'investigations utilise

$$\alpha = 1,5 \quad \beta = 0,5 \quad h_i = 0,1 \quad i = 1, 2, \dots, n.$$

| METHODES |            | I         |                      |        | II        |            |       | IV  |                      |        | V   |            |        |
|----------|------------|-----------|----------------------|--------|-----------|------------|-------|-----|----------------------|--------|-----|------------|--------|
| n        | $\Omega_0$ | k         | $\Omega^k$           | t      | k         | $\Omega^k$ | t     | k   | $\Omega^k$           | t      | k   | $\Omega^k$ | t      |
| 5        | 890        | <u>10</u> | 2,64                 | 1,85   | <u>50</u> | 0,036      | 1,82  | 4*  | 0,618                | 1,98   | 7*  | 0,190      | 2,08   |
| 5        | 1200       | <u>10</u> | $0,55 \cdot 10^{-3}$ | 1,78   | 25        | 0,047      | 0,92  | 4   | $0,66 \cdot 10^{-6}$ | 1,98   | 6   | 0,080      | 2,05   |
| 10       | 12700      | <u>20</u> | 24,6                 | 19,73  | 105       | 0,081      | 9,25  | 7*  | 4,14                 | 21,55  | 11* | 0,014      | 22,08  |
| 10       | 3900       | <u>20</u> | 19,84                | 18,20  | 179       | 0,12       | 15,82 | 6   | 0,029                | 18,67  | 10* | 0,145      | 22,10  |
| 20       | 27000      | <u>30</u> | 12,56                | 173,72 | 171       | 0,048      | 46,72 | 11* | 0,124                | 188,47 | 3*  | 1,99       | 189,65 |

La méthode de relaxation (I) donne les résultats les moins bons. A précision égale la méthode du gradient (II) reste la plus rapide.

## 2) Deuxième type de fonctions

Ce système de fonctions, dont les dérivées partielles sont assez difficiles à obtenir, se rencontre lorsque l'on étudie des lentilles cylindriques électrostatiques; en appliquant la transformation de Schwartz on aboutit à :

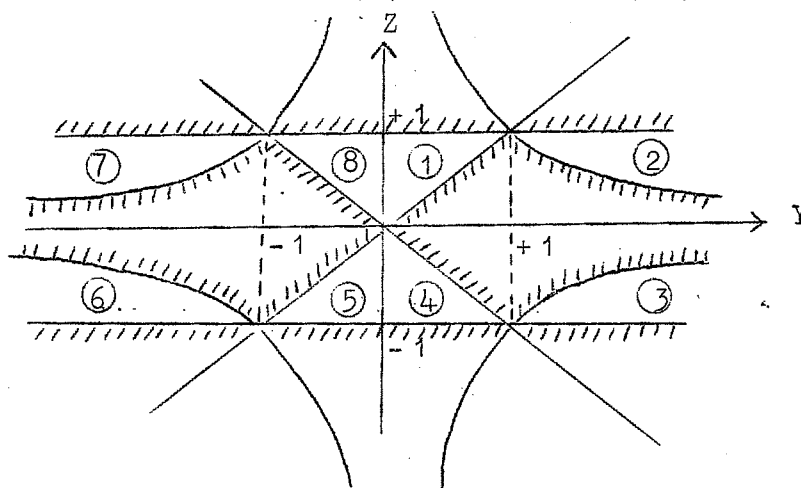
$$F(X,Y,Z) = \frac{\pi}{2} - \frac{X(1-Y^2Z^2)(Z^2-Y^2)}{Y^2Z(1+Z^2)} \quad - 5 = 0$$

$$G(X,Y,Z) = 2X - \frac{10}{\pi} \operatorname{Log} \left( \frac{1-Z}{1+Z} \right) \quad - 1 = 0$$

$$H(X,Y,Z) = X \left( Y + \frac{1}{Y} \right) - \frac{5}{\pi} \operatorname{Log} \left[ \frac{(1-YZ)(Z-Y)}{(1+YZ)(Z+Y)} \right] \quad - 2 = 0$$

Ces fonctions ne sont valables que dans les 8 domaines délimités par les courbes

$$Y = 0 \quad Z = 0 \quad |Z| = 1 \quad \text{et} \quad |YZ| = 1$$



Un premier essai à l'aide de la méthode de linéarisation nous a donné deux solutions, l'une dans (1), l'autre dans (2). Nous recherchons ces deux solutions dans leur domaine respectif. Une série de valeurs initiales permettant de balayer chaque domaine ont été fixées; nous avons numéroté ces points de départ et donné les valeurs correspondantes de  $\delta^0 = \Omega^0$ . Nous comparerons d'abord toutes les méthodes et reviendrons à la méthode des directions conjuguées.

La méthode d'investigations utilise  $\alpha = 1,5$   $\beta = 0,5$   $h_1 = 0,1$ ;  
la solution ou le minimum sont jugés avec  $\epsilon_1 = \epsilon_2 = 10^{-4}$ .

| DOM. | POINT | X | Y      | Z     | $\delta^0 = \Omega^0$ |
|------|-------|---|--------|-------|-----------------------|
| ①    | 1     | 4 | 0,6    | 0,75  | 327                   |
|      | 2     | " | 0,4    | "     | 382                   |
|      | 3     | " | 0,2    | "     | 4607                  |
|      | 4     | " | 0,1    | "     | 85548                 |
|      | 5     | 3 | 0,4    | 0,5   | 191                   |
|      | 6     | " | 0,3    | "     | 272                   |
|      | 7     | " | 0,2    | "     | 1474                  |
|      | 8     | " | 0,1    | "     | 1579                  |
|      | 9     | 2 | 0,2    | 0,25  | 169                   |
|      | 10    | " | 0,1    | "     | 3662                  |
|      | 11    | " | 0,05   | "     | 72259                 |
|      | 12    | " | 0,025  | "     | 1365360               |
|      | 13    | 1 | 0,1    | 0,125 | 142                   |
|      | 14    | " | 0,05   | "     | 3975                  |
|      | 15    | " | 0,025  | "     | 86734                 |
|      | 16    | " | 0,0125 | "     | 1494365               |
| ②    | 17    | 5 | 5      | 0,5   | 4453                  |
|      | 18    | " | 10     | "     | 90107                 |
|      | 19    | " | 15     | "     | 479673                |
|      | 20    | " | 20     | "     | 1543258               |

#### a) Résultats

Nous évaluons les dérivées en prenant  $f'(x) = [f(x+h) - f(x-h)]/2h$  et limitons le nombre d'itérations ou d'évaluations de  $\Omega$  de façon à obtenir des temps sensiblement égaux.

Le tableau ci-dessous montre que la méthode de linéarisation, bien qu'étant divergente pour cinq des vingt points, reste la plus précise et donne, dans les cas favorables, une solution en un nombre restreint d'itérations. Pour chacune des divergences  $H^k$  fournit des points



extérieurs au domaine dès les premières itérations; nous nous trouvons donc en des points où des dérivées partielles sont presque nulles, ce qui donne une mauvaise précision pour le calcul de  $H^k$ . Parmi les méthodes de minimisation, seule l'utilisation de  $n$  directions orthogonales donne de bons résultats. Nous retrouvons encore certains points divergents; on pouvait s'y attendre en voyant les résultats de la méthode de linéarisation car nous aurons aussi de faux minima si  $\partial\Omega/\partial x_i$  est nul donc  $\partial F/\partial x_i$ ,  $\partial G/\partial x_i$ ,  $\partial H/\partial x_i$ , mais le fait de changer de système de directions nous permet de réduire le nombre de faux minima. La méthode de relaxation variable par variable diminuant la fonction rapidement au cours des premières itérations a une convergence bien plus lente par la suite et ne pourra être utilisée pour la résolution de systèmes d'équations qu'en faisant un grand nombre d'itérations. Les méthodes du gradient et des directions conjuguées donnent de mauvais résultats : ceci est dû à leur méthode de recherche liée au gradient. Si le calcul approché des dérivées est trop imprécis, donc si la direction  $D^k$  devient quelconque, nous faussons les tests d'arrêt ( $DX^k$  tend vers zéro du fait d'une mauvaise direction de minimisation).

| Point | LINEAR.   |                       | I         |                      | II        |            | III       |                      | IV |                      | V  |                      |
|-------|-----------|-----------------------|-----------|----------------------|-----------|------------|-----------|----------------------|----|----------------------|----|----------------------|
|       | k         | $\delta^k$            | k         | $\Omega^k$           | k         | $\Omega^k$ | k         | $\Omega^k$           | k  | $\Omega^k$           | k  | $\Omega^k$           |
| 1     | <u>25</u> | 27,22                 | 10        | 27,70                | <u>50</u> | 19,9       | 19        | $0,37 \cdot 10^{-3}$ | 2  | 27,68                | 10 | $0,46 \cdot 10^{-3}$ |
| 2     | <u>25</u> | 27,63                 | <u>50</u> | 0,11                 | <u>50</u> | 215        | 21        | $0,28 \cdot 10^{-3}$ | 9  | $0,64 \cdot 10^{-9}$ | 24 | $0,57 \cdot 10^{-6}$ |
| 3     | 8         | $0,14 \cdot 10^{-9}$  | <u>50</u> | 0,10                 | <u>50</u> | 203        | 14        | 30,36                | 12 | $0,11 \cdot 10^{-7}$ | 24 | $0,38 \cdot 10^{-2}$ |
| 4     | 7         | $0,10 \cdot 10^{-11}$ | <u>50</u> | 0,08                 | <u>50</u> | 206        | 8         | 29,86                | 8  | $0,38 \cdot 10^{-5}$ | 21 | $0,85 \cdot 10^{-6}$ |
| 5     | <u>25</u> | 32,2                  | <u>50</u> | 29,80                | <u>50</u> | 100        | 15        | $0,24 \cdot 10^{-2}$ | 10 | 29,53                | 23 | $0,45 \cdot 10^{-5}$ |
| 6     | <u>25</u> | 28,15                 | <u>50</u> | 0,11                 | <u>50</u> | 115        | 18        | $0,31 \cdot 10^{-3}$ | 11 | $0,15 \cdot 10^{-6}$ | 15 | 0,072                |
| 7     | 9         | $0,18 \cdot 10^{-12}$ | <u>50</u> | 0,10                 | <u>50</u> | 82         | 25        | $0,61 \cdot 10^{-3}$ | 9  | $0,21 \cdot 10^{-9}$ | 20 | $0,29 \cdot 10^{-6}$ |
| 8     | 6         | $0,75 \cdot 10^{-10}$ | <u>50</u> | 0,078                | <u>50</u> | 108        | 13        | 16,72                | 8  | $0,14 \cdot 10^{-6}$ | 19 | $0,43 \cdot 10^{-8}$ |
| 9     | <u>25</u> | 27,8                  | <u>50</u> | 0,077                | <u>50</u> | 56,7       | 22        | $0,25 \cdot 10^{-3}$ | 8  | $0,54 \cdot 10^{-8}$ | 13 | $0,19 \cdot 10^{-8}$ |
| 10    | 7         | $0,30 \cdot 10^{-10}$ | <u>50</u> | 0,068                | <u>50</u> | 45         | 23        | $0,23 \cdot 10^{-3}$ | 9  | $0,69 \cdot 10^{-9}$ | 15 | $0,25 \cdot 10^{-2}$ |
| 11    | 8         | $0,10 \cdot 10^{-10}$ | <u>50</u> | 0,01                 | <u>50</u> | 41         | 10        | 10,15                | 7  | $0,31 \cdot 10^{-7}$ | 15 | $0,55 \cdot 10^{-8}$ |
| 12    | 12        | $0,13 \cdot 10^{-12}$ | 41        | 0,072                | <u>50</u> | 41,4       | 10        | 10,50                | 10 | $0,14 \cdot 10^{-3}$ | 24 | $0,57 \cdot 10^{-4}$ |
| 13    | 4         | $0,92 \cdot 10^{-8}$  | <u>50</u> | 0,023                | <u>50</u> | 19,43      | <u>50</u> | 0,017                | 13 | $0,21 \cdot 10^{-8}$ | 19 | 0,015                |
| 14    | 4         | $0,21 \cdot 10^{-8}$  | 32        | $0,11 \cdot 10^{-2}$ | <u>50</u> | 9,95       | 11        | 2,57                 | 5  | $0,24 \cdot 10^{-3}$ | 12 | $0,18 \cdot 10^{-5}$ |
| 15    | 6         | $0,86 \cdot 10^{-11}$ | 40        | 0,094                | <u>50</u> | 9,46       | 10        | 1,95                 | 8  | $0,85 \cdot 10^{-5}$ | 35 | $0,15 \cdot 10^{-5}$ |
| 16    | 8         | $0,91 \cdot 10^{-12}$ | <u>50</u> | 0,38                 | <u>50</u> | 9,73       | 11        | 1,92                 | 11 | $0,24 \cdot 10^{-8}$ | 15 | 0,001                |
| 17    | 12        | $0,12 \cdot 10^{-11}$ | <u>50</u> | 0,18                 | 24        | 861        | 4         | 33,16                | 15 | 0,16                 | 36 | 0,25                 |
| 18    | 6         | $0,89 \cdot 10^{-13}$ | <u>50</u> | 0,13                 | <u>50</u> | 2780       | 3         | 977                  | 14 | 0,13                 | 42 | $0,11 \cdot 10^{-2}$ |
| 19    | 8         | $0,94 \cdot 10^{-11}$ | <u>50</u> | 0,075                | 10        | 6192       | 34        | 3004                 | 5  | 0,37                 | 25 | $0,63 \cdot 10^{-3}$ |
| 20    | 10        | $0,82 \cdot 10^{-12}$ | <u>50</u> | 0,028                | 3         | 10830      | 3         | 10734                | 3  | 0,057                | 10 | 0,066                |

Nous allons maintenant regarder l'effet que produit l'approximation des dérivées sur la convergence de la méthode des directions conjuguées. Nous prendrons comme dérivées pour la méthode

III exactes

III bis approchées par  $\frac{f(X+h) - f(X-h)}{2h}$

III ter " "  $\frac{f(X+h) - f(X)}{h}$

| Point | k  | III                   | k  | III bis              | k  | III ter              |
|-------|----|-----------------------|----|----------------------|----|----------------------|
|       |    | $\Omega^k$            |    | $\Omega^k$           |    | $\Omega^k$           |
| 1     | 20 | $0,60 \cdot 10^{-12}$ | 19 | $0,37 \cdot 10^{-3}$ | 19 | $0,54 \cdot 10^{-2}$ |
| 2     | 11 | 3,92                  | 21 | $0,28 \cdot 10^{-3}$ | 19 | 0,036                |
| 3     | 16 | 30,31                 | 14 | 30,37                | 14 | 30,36                |
| 4     | 8  | 29,85                 | 8  | 29,85                | 8  | 29,85                |
| 5     | 17 | $0,32 \cdot 10^{-12}$ | 15 | $0,24 \cdot 10^{-2}$ | 17 | 0,014                |
| 6     | 19 | $0,27 \cdot 10^{-12}$ | 18 | $0,31 \cdot 10^{-3}$ | 16 | $0,80 \cdot 10^{-2}$ |
| 7     | 22 | $0,14 \cdot 10^{-12}$ | 25 | $0,61 \cdot 10^{-3}$ | 23 | 0,012                |
| 8     | 8  | 22,49                 | 13 | 16,72                | 15 | 10,85                |
| 9     | 23 | $0,14 \cdot 10^{-9}$  | 22 | $0,25 \cdot 10^{-3}$ | 14 | 1,05                 |
| 10    | 27 | $0,92 \cdot 10^{-12}$ | 23 | $0,22 \cdot 10^{-3}$ | 22 | $0,42 \cdot 10^{-2}$ |
| 11    | 10 | 10,17                 | 10 | 10,15                | 10 | 10,16                |
| 12    | 10 | 10,50                 | 10 | 10,50                | 10 | 10,46                |
| 13    | 8  | 2,25                  | 50 | 0,017                | 11 | 0,45                 |
| 14    | 11 | 0,39                  | 11 | 2,57                 | 13 | 2,09                 |
| 15    | 10 | 1,94                  | 10 | 1,95                 | 10 | 1,94                 |
| 16    | 11 | 1,92                  | 11 | 1,92                 | 11 | 1,90                 |
| 17    | 10 | 29,85                 | 4  | 33,16                | 8  | 30,34                |
| 18    | 14 | 28,50                 | 3  | 997                  | 3  | 1177                 |
| 19    | 12 | 27,86                 | 34 | 3004                 | 1  | $0,48 \cdot 10^6$    |
| 20    | 12 | 27,46                 | 3  | 10734                | 3  | 5821                 |

Si dans le domaine (1) nous retrouvons des résultats semblables, dans le domaine (2) les dérivées approchées donnent des résultats faux.

Même avec les dérivées exactes, la méthode des directions conjuguées est divergente en bien des points. La recherche unidirectionnelle plus sujette à des minima parasites devra être employée avec prudence pour des fonctions assez complexes, il sera préférable d'utiliser n directions orthogonales.

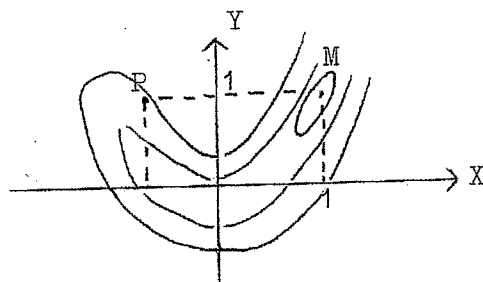
## B. MINIMISATION D'UNE FONCTION POSITIVE

Le but de cette deuxième partie est de comparer entre elles les méthodes de minimisation d'une fonction. Nous avons rassemblé quelques types de fonctions afin de faire une comparaison assez juste.

### 1) Premier type de fonction

Soit la fonction  $\Omega(X,Y) = 100 (Y - X^2)^2 + (1 - X)^2$  qui admet des minima parasites très prononcés pour  $Y = X^2$ .

Le minimum nul de cette fonction est obtenu au point  $M(1,1)$ . Représentons les surfaces équipotentielles  $\Omega(X,Y) = \text{cte}$ .



Nous allons partir du point  $P(-1,2; 1)$  où  $\Omega^0 = 24,4$  pour observer si les différentes méthodes arriveront facilement dans le premier quadrant.

Nous avons pris les dérivées analytiques et imposé un nombre maximum d'itérations. Le minimum est jugé avec  $\epsilon_1 = \epsilon_2 = 10^{-4}$ . La méthode d'investigations utilise  $\alpha = 1,5$   $\beta = 0,5$   $h_1 = h_2 = 0,1$ .

| Méthode | I         |            | II        |                      | III  |                       | IV   |                      | V    |                      |
|---------|-----------|------------|-----------|----------------------|------|-----------------------|------|----------------------|------|----------------------|
|         | k         | $\Omega^k$ | k         | $\Omega^k$           | k    | $\Omega^k$            | k    | $\Omega^k$           | k    | $\Omega^k$           |
|         | 5         | 3,997      | 5         | 3,061                | 5    | 1,289                 | 5    | 1,289                | 5    | 2,538                |
|         | 10        | 3,895      | 10        | 2,997                | 10   | 0,010                 | 10   | 0,194                | 10   | 0,377                |
|         | 20        | 3,682      | 20        | 0,025                | 15   | $0,57 \cdot 10^{-9}$  | 20   | 0,041                | 20   | $0,16 \cdot 10^{-3}$ |
|         | 30        | 0,028      | 30        | $0,28 \cdot 10^{-2}$ | 16   | $0,60 \cdot 10^{-13}$ | 28   | $0,41 \cdot 10^{-8}$ | 24   | $0,91 \cdot 10^{-7}$ |
|         | <u>50</u> | 0,019      | <u>50</u> | $0,23 \cdot 10^{-3}$ |      |                       |      |                      |      |                      |
| t       | 0,70      |            | 0,75      |                      | 0,32 |                       | 0,92 |                      | 0,50 |                      |

La méthode des directions conjuguées (III) est la plus rapide et la plus précise, les méthodes utilisant  $n$  directions orthogonales sont moins rapides, mais précises. A temps égal la méthode du gradient donne de meilleurs résultats que la méthode de relaxation qui devient très lente près de la solution.

## 2) Deuxième type de fonction

Soit la fonction de 3 variables

$$\Omega(X,Y,Z) = 100[(Z - 10\vartheta)^2 + (R - 1)^2] + Z^2$$

$$\text{où } R = (X^2 + Y^2)^{1/2} \text{ et } 2\pi\vartheta = \begin{cases} \text{Arctg}(Y/X) & \text{si } X > 0 \\ \pi + \text{Arctg}(Y/X) & \text{si } X < 0 \end{cases}$$

Cette fonction décroît le long d'une spirale d'axe OZ ayant un pas égal à 10 et un rayon égal à 1, si on la considère dans la région définie par

$$-\frac{\pi}{2} < 2\pi\vartheta < \frac{3\pi}{2}$$

$$-2,5 < Z < 7,5$$

Elle admet un minimum nul au point (1,0,0).

Nous avons encore limité le nombre d'itérations de façon à obtenir des temps comparables. La méthode d'investigations utilise les constantes  $\alpha = 1,5$   $\beta = 0,5$   $h_1 = h_2 = h_3 = 0,1$ .

Nous prenons comme point initial (-1,0,0) ce qui donne  $\Omega^0 = 2500$ . Nous donnons dans ce tableau les valeurs de la fonction à chaque itération.

| k  | I     | II    | III                   | IV                   | V                    |
|----|-------|-------|-----------------------|----------------------|----------------------|
| 0  | 2500  | 2500  | 2500                  | 2500                 | 2500                 |
| 1  | 89,88 | 521,7 | 521,7                 | 12,02                | 199                  |
| 2  | 12,30 | 53,23 | 109,6                 | 11,36                | 12,88                |
| 3  | 12,02 | 9,73  | 74,08                 | 8,15                 | 9,91                 |
| 4  | 11,67 | 6,75  | 24,19                 | 4,41                 | 9,75                 |
| 5  | 11,34 | 6,15  | 10,94                 | 2,53                 | 9,66                 |
| 6  | 11,08 | 6,05  | 9,84                  | 1,03                 | 9,49                 |
| 7  | 10,85 | 5,77  | 6,30                  | 0,38                 | 8,06                 |
| 8  | 10,54 | 5,65  | 6,09                  | 0,067                | 6,21                 |
| 9  | 10,32 | 5,58  | 1,89                  | 0,002                | 5,11                 |
| 10 | 9,97  | 5,52  | 1,75                  | $0,24 \cdot 10^{-6}$ | 3,40                 |
| 11 | 9,76  | 5,47  | 0,76                  | $0,41 \cdot 10^{-9}$ | 2,13                 |
| 12 | 9,52  | 5,42  | 0,36                  | $0,11 \cdot 10^{-9}$ | 1,46                 |
| 13 | 9,23  | 5,36  | 0,13                  |                      | 0,63                 |
| 14 | 8,97  | 5,31  | 0,05                  |                      | 0,45                 |
| 15 | 8,74  | 5,23  | 0,01                  |                      | 0,23                 |
| 16 | 8,52  | 5,17  | $0,72 \cdot 10^{-3}$  |                      | 0,16                 |
| 17 | 8,31  | 5,05  | $0,26 \cdot 10^{-4}$  |                      | 0,14                 |
| 18 | 8,10  | 4,95  | $0,59 \cdot 10^{-7}$  |                      | 0,12                 |
| 19 | 7,92  | 4,60  | $0,71 \cdot 10^{-10}$ |                      | 0,07                 |
| 20 | 7,75  | 4,28  |                       |                      | $0,22 \cdot 10^{-2}$ |
| 21 | 7,59  | 1,49  |                       |                      | $0,10 \cdot 10^{-3}$ |
| 22 | 7,44  | 1,10  |                       |                      | $0,35 \cdot 10^{-4}$ |
| 23 | 7,29  | 1,03  |                       |                      | $0,23 \cdot 10^{-4}$ |
| 24 | 7,15  | 1,01  |                       |                      | $0,21 \cdot 10^{-4}$ |
| 25 | 7,01  | 0,98  |                       |                      |                      |
| 30 | 6,39  | 0,81  |                       |                      |                      |
| 50 |       | 0,36  |                       |                      |                      |
| t  | 0,92  | 0,87  | 0,50                  | 1,10                 | 1,03                 |

La méthode des directions conjuguées qui utilise ici les dérivées exactes est la plus rapide. Les méthodes utilisant  $n$  directions orthogonales donnent aussi de bons résultats. La méthode du gradient, diminuant la fonction rapidement au cours des premières itérations, converge ensuite très lentement tandis que la méthode de relaxation donne les moins bons résultats.

### 3) Troisième type de fonction

C'est un problème rencontré pour des colonnes de distillation : il s'agit de minimiser une quantité de chaleur  $Q = Q_1 + Q_2 + Q_3$  fonction positive de deux variables  $Q = Q(\lambda, \mu)$ , déterminée dans l'intervalle  $0 < \lambda < 5$  et  $0 < \mu < 0,5$ . Le calcul des dérivées partielles de  $Q$  est impossible.

Etant donné une série de constantes

$$\begin{aligned} & a_1 \ a_2 \ a_3 \ a_4; \ a_1' \ a_2'; \ a_1'' \ a_2''; \ A_1 \ A_2 \ A_3; \\ & TR \ TF \ TB \ ; \ TR' \ TF' \ TB' \ ; \ TR'' \ TF'' \ TB''; \\ & C_1 \ C_2 \ C_3 \ C_4; \ B_1 \ B_2 \ B_3 \ B_4 \ ; \ F_1 \ F_2 \ F_3 \ F_4; \end{aligned}$$

on calcule  $P_1 \ P_2 \ P_3 \ P_4$  donnés par le système linéaire en  $\lambda$  et  $\mu$

$$[F_i] = [L(\lambda, \mu)] \cdot [P_i]$$

ce qui permet de déterminer facilement un groupe de variables, fonction de  $\lambda$  et de  $\mu$  :

$$\begin{aligned} & t_1 \ t_2; \ w_3 \ w_4; \\ & u_1 \ u_2 \ u_3; \ v_2 \ v_3 \ v_4; \ y_1 \ y_2 \ y_3; \ z_2 \ z_3 \ z_4 \\ & x_1 \ x_2 \ x_3 \ x_4 \ . \end{aligned}$$

On tire ensuite  $R_1 \ R_2 \ R_3$  des expressions :

$$\begin{aligned} \text{Log} \left( \frac{y_2 \ z_3}{y_3 \ z_2} \right) &= A_1 \left( \frac{R_1}{\sum y_i} \right)^{a_1} \left( \frac{x_1}{x_3} \right)^{a_2} \left( \frac{x_2}{x_3} \right)^{a_3} \left( \frac{x_4}{x_3} \right)^{a_4} \\ \text{Log} \left( \frac{t_1 \ u_2}{t_2 \ u_1} \right) &= A_2 \left( \frac{R_2}{\sum t_i} \right)^{a_1'} \left( \frac{y_1}{y_2} \right)^{a_2'} \\ \text{Log} \left( \frac{v_3 \ w_4}{v_4 \ w_3} \right) &= A_3 \left( \frac{R_3}{\sum v_i} \right)^{a_1''} \left( \frac{z_3}{z_4} \right)^{a_2''} \end{aligned}$$

On calcule alors :

$$\alpha_1 = \frac{\sum y_i B_i}{\sum y_i} ; \beta_1 = TR \frac{\sum y_i C_i}{\sum y_i} ; \gamma_1 = TB (\sum z_i C_i)$$

$$\alpha_2 = \frac{\sum t_i B_i}{\sum t_i} ; \beta_2 = TR' \frac{\sum t_i C_i}{\sum t_i} ; \gamma_2 = TB' (\sum u_i C_i)$$

$$\alpha_3 = \frac{\sum v_i B_i}{\sum v_i} ; \beta_3 = TR'' \frac{\sum v_i C_i}{\sum v_i} ; \gamma_3 = TB'' (\sum w_i C_i) ,$$

ce qui permet de tirer  $Q_1$   $Q_2$   $Q_3$  des relations :

$$Q_1 + R_1 \beta_1 + TF(\sum x_i C_i) = (\sum y_i + R_1)(\alpha_1 + R_1) + \gamma_1$$

$$Q_2 + R_2 \beta_2 + TF'(\sum y_i C_i) = (\sum t_i + R_2)(\alpha_2 + R_2) + \gamma_2$$

$$Q_3 + R_3 \beta_3 + TF''(\sum z_i C_i) = (\sum v_i + R_3)(\alpha_3 + R_3) + \gamma_3$$

d'où

$$\Omega = Q = Q_1 + Q_2 + Q_3 .$$

Le test pour le minimum s'effectue avec  $\epsilon_1 = \epsilon_2 = 10^{-4}$ . La méthode d'investigations utilise  $\alpha = 3$   $\beta = 0,5$   $h_1 = h_2 = 0,1$ .

Nous sommes partis du point(1;.0,1) d'où  $\Omega^0 = 143960,45$ .

| METH. | I    |            | II   |            | III  |            | IV   |            | V    |            |
|-------|------|------------|------|------------|------|------------|------|------------|------|------------|
|       | k    | $\Omega^k$ | k    | $\Omega^k$ | k    | $\Omega^k$ | k    | $\Omega^k$ | k    | $\Omega^k$ |
|       | 1    | 117898,36  | 1    | 137171,60  | 1    | 137171,60  | 1    | 116168,25  | 1    | 137789,06  |
|       | 2    | 116191,68  | 2    | 127989,58  | 2    | 126233,80  | 2    | 116168,19  | 2    | 130298,72  |
|       | 3    | 116168,25  | 3    | 127324,54  | 3    | 117356,77  |      |            | 3    | 129142,54  |
|       | 4    | 116,168,19 | 4    | 127021,48  | 4    | 116233,39  |      |            | 4    | 119928,39  |
|       |      |            | 5    | 124195,45  | 5    | 116186,81  |      |            | 5    | 116171,38  |
|       |      |            | 10   | 116325,81  | 6    | 116168,21  |      |            | 6    | 116168,33  |
|       |      |            | 15   | 116170,85  | 7    | 116168,19  |      |            | 7    | 116168,19  |
|       |      |            | 20   | 116168,46  |      |            |      |            |      |            |
| t     | 0,86 |            | 4,50 |            | 1,44 |            | 1,32 |            | 1,44 |            |



La méthode de relaxation est la plus rapide. La méthode du gradient est de loin la plus lente mais arrive quand même au minimum. Les trois autres méthodes sont équivalentes.

#### 4) Quatrième type de fonction

C'est un problème qui se pose lorsque l'on veut déterminer un filtre électrique : on connaît  $n$  points d'une courbe expérimentale représentant une impédance  $g(t)$  en fonction de la fréquence  $t$  sur l'intervalle  $[0, 1; 1, 1]$  mégahertz. Nous voulons approximer cette courbe au sens des moindres carrés par une fonction  $f(t)$  que l'on définit par

$$f(t) = |y(t)|$$

et 
$$y(t) = K \frac{(z^2 + s_1 z)(z^2 + s_2 z + P_2)}{(z^2 + s_3 z + P_3)(z^2 + s_4 z + P_4)} \quad \text{avec } z = +it.$$

Il faut déterminer les valeurs des coefficients réels et positifs  $K, s_1, s_2, s_3, s_4, P_2, P_3, P_4$  de façon à minimiser la fonction

$$\Omega(K, \dots, P_4) = \sum_{i=1}^n [g(t_i) - f(t_i)]^2.$$

Un calcul simple nous permet de mettre  $f(t)$  sous la forme :

$$f(t) = K \left( \frac{AB}{CD} \right)^{1/2}$$

$$A = 1 + \left( \frac{s_1}{t} \right)^2$$

$$B = \left( \frac{P_2}{t^2} - 1 \right)^2 + \left( \frac{s_2}{t} \right)^2$$

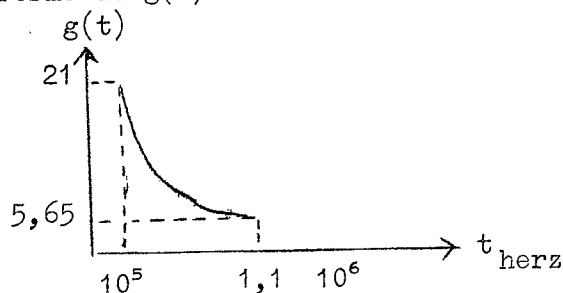
$$C = \left( \frac{P_3}{t^2} - 1 \right)^2 + \left( \frac{s_3}{t} \right)^2$$

$$D = \left( \frac{P_4}{t^2} - 1 \right)^2 + \left( \frac{s_4}{t} \right)^2$$

ce qui explique les valeurs initiales voisines de la solution qui sont d'un ordre de grandeur très différent.

$$K_1 = 1 \quad s_1 = s_2 = s_3 = s_4 = 10^7 \quad P_2 = P_3 = P_4 = 10^{13}.$$

La forme de  $g(t)$  est semblable à un arc d'hyperbole



Nous avons utilisé onze points équidistants de  $10^5$  herz, ce qui donne  $\Omega^0 = 1256$ .

Le minimum est testé avec  $\epsilon_1 = \epsilon_2 = 10^{-4}$ ; la méthode d'investigations utilise  $\alpha = 3$   $\beta = 0,5$   $h_1 = \dots = h_8 = 10^6$ .

| I   |            | II |            | III |            | IV |            | V |            |
|-----|------------|----|------------|-----|------------|----|------------|---|------------|
| k   | $\Omega^k$ | k  | $\Omega^k$ | k   | $\Omega^k$ | k  | $\Omega^k$ | k | $\Omega^k$ |
| 1   | 671,81     | 1  | 845,21     | 1   | 845,21     | 1  | 462,00     | 1 | 19,91      |
| 2   | 551,23     | 2  | 845,21     | 2   | 845,21     | 2  | 461,91     | 2 | 19,38      |
| 3   | 462,00     |    |            |     |            |    |            | 3 | 19,33      |
| 4   | 334,96     |    |            |     |            |    |            |   |            |
| 5   | 177,35     |    |            |     |            |    |            |   |            |
| 10  | 11,40      |    |            |     |            |    |            |   |            |
| 20  | 10,05      |    |            |     |            |    |            |   |            |
| 50  | 1,30       |    |            |     |            |    |            |   |            |
| 100 | 1,16       |    |            |     |            |    |            |   |            |

Les méthodes du gradient et des directions conjuguées n'ont agit que sur la première composante K laissant les autres inchangées d'où leur arrêt à la deuxième itération. Les méthodes utilisant n directions perpendiculaires changent toutes les composantes à la première itération, mais les nouvelles directions seront telles que seule la composante K changera d'où l'arrêt.

La méthode de relaxation faisant varier tour à tour  $K, s_1, \dots, P_4$  ne sera pas affectée par les ordres de grandeur très différents des variables et minimisera très bien cette fonction.



## CHAPITRE V

### CONCLUSION

Ce chapitre sera divisé en deux parties : l'une relative à la résolution des systèmes d'équations, l'autre à la minimisation d'une fonction positive.

#### A. ETUDE D'UN SYSTEME D'EQUATIONS

La méthode d'étude directe du système reste préférable même pour de grands systèmes. La méthode de linéarisation nous a donné de bons résultats dans les deux types de systèmes envisagés; néanmoins, le calcul des dérivées approchées devra être effectué avec certaines précautions ( $\Delta x_i$  inférieur à  $|x_i^k - x_i^{k-1}|$ ) de façon à conserver une convergence rapide. L'adjonction des tests sur la fonction associée  $\delta = \sum f_i^2$  et sur le nombre d'itérations en font une méthode sûre, d'un emploi facile pour des calculateurs rapides comme se fut le cas sur l'ordinateur I.B.M. 7090.

Lorsque cette méthode diverge, on pourra essayer d'autres valeurs initiales ou alors minimiser la fonction  $\Omega = \sum f_i^2$ . Seules les méthodes évoluées (directions conjuguées, utilisation de  $n$  directions orthogonales) pourront être envisagées car elles se sont révélées précises. Les méthodes utilisant  $n$  directions orthogonales, d'une programmation assez difficile et demandant un temps de calcul élevé, seront prises si la méthode des directions conjuguées se révèle inefficace. Il faudra de plus tester si le minimum trouvé est nul car dans certains cas nous aurons des minima parasites. Nous reparlons de ces méthodes dans le paragraphe suivant.

#### B. MINIMISATION D'UNE FONCTION

Le problème se présente sous deux aspects différents suivant le type de calculateur mis à la disposition de l'utilisateur.

### 1) Ordinateur de petite capacité

Un ordinateur à capacité moyenne s'accommodera mal des méthodes de minimisation évoluées qui demandent déjà un certain nombre de mémoires; si on ne travaille pas avec un langage évolué (FORTRAN, ALGOL) elles seront d'une programmation ardue. Le choix entre la méthode du gradient et la méthode de relaxation variable par variable est délicat.

La méthode du gradient dans beaucoup de cas est plus précise, à égalité de temps, que la méthode de relaxation. Par contre, elle demandera quelques précautions si on approche les dérivées numériquement ( $\Delta x_i$  inférieur à  $|x_i^k - x_i^{k-1}|$ ); elle s'avère mauvaise pour des variables d'un ordre de grandeur très différent.

La méthode de relaxation variable par variable, d'une programmation facile, nécessite le minimum d'encombrement; elle donne de bons résultats pour des variables d'ordre de grandeur très différent. Son principal défaut réside dans sa convergence lente près d'un minimum. Elle restera quand même préférable si l'on désire seulement une bonne approximation d'un minimum.

### 2) Ordinateur de grande capacité

Les méthodes évoluées sont indiquées. La méthode des directions conjuguées sera à utiliser en premier car elle s'est révélée efficace et rapide même pour de grands systèmes; nous ferons toutefois la même remarque pour l'évaluation numérique des dérivées ( $\Delta x_i$  inférieur à  $|x_i^k - x_i^{k-1}|$ ). Dans le cas de fonctions assimilables à des fonctions quadratiques définies positives (fonction près de son minimum) elle possède une très bonne convergence. La recherche unidirectionnelle peut nous conduire à de faux minima surtout si on évalue numériquement les dérivées.

Les méthodes utilisant  $n$  directions orthogonales sont précises mais plus lentes, surtout si le nombre de variables devient élevé. Leur méthode de recherche conviendra parfaitement pour des fonctions dont les

dérivées analytiques sont impossibles à calculer ou pour des fonctions ayant des minima parasites. La méthode utilisant des interpolations d'une logique un peu moins complexe sera préférable car elle ne nécessitera pas l'adjonction des constantes  $\alpha, \beta, h_1, \dots, h_n$  dépendant du type de fonction que l'on étudie, ceci demandant en général des essais préliminaires afin de leur donner une valeur convenable.

Pour des fonctions possédant des variables d'ordre de grandeur très différent ces méthodes évoluées se révèlent inopérantes et l'on aura encore recours à la méthode de relaxation variable par variable.

### C. CONCLUSION

Nous pourrions dire en guise de conclusion qu'il n'existe aucune méthode absolument sûre pour la résolution d'un système d'équations ou la minimisation d'une fonction positive. Il faudra étudier ce système ou cette fonction afin de déterminer la méthode qui semble la plus appropriée. Dans certains cas, l'élaboration d'un programme adapté aux conditions d'utilisation reste préférable à l'emploi automatique d'une méthode dite générale.

REFERENCES

- 1) J. Legras, Précis d'Analyse Numérique, DUNOD, 43 (1963).
- 2) J. Legras, " " " " " " " " , 98 (1963).
- 3) H.A. Spang, "A Review of Minimization Techniques for Non-linear Functions", SIAM Review, Vol. 4, No 4, 343 (1962).
- 4) H.D. Curry, "The Method of Steepest Descent for Non-linear Minimization Problem", QUART. APPL. MATHS., Vol. 2, 258 (1944).
- 5) W.C. Davidson, "Variable Metric Method for Minimization", Report ANL 5990 (1959).
- 6) R. Fletcher and M.J.D. Powell, "A Rapidly Convergent Descent Method for Minimization", The Computer Journal, Vol. 6, No 2, 163 (1963).
- 7) H.H. Rosenbrock, "An Automatic Method for Finding the Great or Least Value of a Function", The Computer Journal, Vol. 3, No 3, 175 (1960).

REMERCIEMENTS

Que Monsieur le Professeur J. LEGRAS, Directeur du Centre Universitaire de Calcul Automatique de NANCY, ainsi que Monsieur le Docteur H. LIPPS, Chef de la Section de Programmation à l'Organisation Européenne pour la Recherche Nucléaire, veuillent bien trouver ici l'expression de ma profonde reconnaissance pour les nombreux conseils qu'ils m'ont prodigués et les heureuses suggestions qu'ils m'ont fournies.

Je tiens également à remercier Monsieur le Docteur G. MACLEOD, Directeur de la Division des Données et Documents, de la bienveillance dont il a fait preuve.

Enfin, je remercie Messieurs les Professeurs qui ont bien voulu me faire l'honneur de composer le jury.



