ALGORITHMES POUR UN PROBLEME INVERSE DISCRET DE STURM-LIOUVILLE.

P. Morel
Université de Bordeaux I
33405 Talence / France

I - Introduction.

On considère sur $[o,\pi]$ l'opérateur différentiel de Sturm-Liouville $L(y) = -y'' + q(x)y = \lambda y$ avec les conditions aux limites $\alpha_1 y(o) + \beta_1 y'(o) = 0$, $\alpha_2 y(\pi) + \beta_2 y'(\pi) = 0$. On appelle problème inverse de valeurs propres la recherche de la fonction q connaissant le spectre de l'opérateur.

Nous voulons obtenir numériquement la fonction q ; c'est donc la version discrétisée de ce dernier problème qui nous intéresse. Après avoir introduit un pas de discrétisation, un maillage ,on obtient un problème matriciel qui légèrement généralisé s'énonce de la manière suivante. On appelle problème (P_s) la recherche d'une matrice nxn diagonale réelle $X = (x_i \delta_{ij}) \in \mathcal{M}_{nn}(\mathbb{R})$ telle que A étant une matrice nxn donnée symétrique de $\mathcal{M}_{nn}(\mathbb{R})$ le spectre: Sp (A+X) de A+X soit égal au spectre de la matrice fixée $S = (s_i \delta_{ij}) \in \mathcal{M}_{nn}(\mathbb{R})$.

On peut faire deux hypothèses ne diminuant en rien la généralité du problème traité. On peut supposer que la diagonale de la matrice A est nulle. En effet X est une solution pour A et S fixées ie. $Sp(A+X)=SpS$ si et seulement si X-Diag A est une solution pour A et S données.
D'autre part soit $t \in \mathbb{R}$ tel que pour $i=1, 2, \ldots, n$ $s_i > t > 0$
Si alors X est telle que :

$$Sp(A+X) = Sp((s_i+t)\,\delta_{ij})$$

alors :

$$X-(t\delta_{ij}) \text{ vérifie } Sp(A+X-(t\delta_{ij})) = Sp(S) ;$$

en d'autres termes on peut supposer que le spectre visé est strictement positif. Nous incluons ces deux hypothèses, non restrictives, dans la formulation du problème (P_S).

2 - Des conditions nécessaires et des conditions suffisantes.

L'étude en dimension 2x2 montre immédiatement que le problème (Ps) ne possède pas toujours de solution. K. Hadeler [2], F. Laborde [4], P. Morel [6, 7] ont donné des conditions nécessaires de plus en plus précises pour que le problème

(Ps) possède des solutions. On montre dans Morel [7] que nécessairement $S = (s_i\,\delta_{ij})$ doit vérifier :

$$\sum_{i=1}^{n} s_i^2 - \frac{1}{n}(\Sigma s_i)^2 \geqslant \sum_{i,j} a_{ij}\,a_{ji}$$

ce qui d'une manière équivalente mettant en évidence une nécessaire séparation du spectre visé s'écrit :

$$2n \sum_{i,j} a_{ij}\,a_{ji} \leqslant \sum_{i,j} (s_i - s_j)^2$$

Dans le cas où A est symétrique $\sum_{i,j} a_{ij}a_{ji} = \sum_{i,j} a_{ij}^2 = \operatorname{tr} A^2 = \|A\|_S^2$; $\|A\|_s$ désignant la norme de Schur de A. Sous cette hypothèse cela permet d'affirmer que : l'application $x \to \mu(A+(x\delta_{ij}))$, où $\mu(A+x_i\delta_{ij})$ désigne le vecteur dont les composantes sont les valeurs propres de $A+(x_i\delta_{ij})$ numérotées dans l'ordre non croissant n'est surjective que si A est nulle.

Dans [6], on obtient la condition nécessaire suivante, qui est strictement plus précise que les précédentes

$$(\overleftarrow{\mu}(S)\,|\,\mu(A)) \leqslant \sum_{i,j} a_{ij}^2 \leqslant (\mu(S)\,|\,\mu(A))$$

où $(\overleftarrow{\mu}(S)\,|\,(\mu(A))$ désigne le produit scalaire entre les vecteurs $\overleftarrow{\mu}(S)$ et $\mu(A)$ qui désignent respectivement les valeurs propres de S dans l'ordre croissant et les valeurs propres de A dans l'ordre décroissant.

Ces conditions nécessaires, ne sont suffisantes que si la dimension est inférieure ou égale à 2. Dans Morel [7] on recherche systèmatiquement une localisation de la solution ; cela pour choisir le plus correctement possible une approximation initiale lors de la mise en oeuvre d'un algorithme. Dans cet ordre d'idée citons:

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} s_i^2 - \sum_{i\,j=1}^{n} a_{ij}^2$$

Toutes les conditions suffisantes connues expriment que le spectre visé est suffisamment séparé. Plus précisément que $d(s) = \min_{i \neq j} |s_i - s_j| > f(\lambda_1(A), \ldots, \lambda_n(A))$ où f est une fonction des valeurs propres de A.

Donnons celle de Morel [7] :

$$d(s) = \min_{i \neq j} |s_i - s_j| \geqslant 2^{(1-1/p)} \left(\sum_{i=j}^{n} |\lambda_i(A)|^p\right)^{1/p}$$

Elle recouvre celle de Laborde ($p=+\infty$) et rappelle sans être pourtant identique celle

de de Oliviera (p=1) et, la première connue, celle de K. Hadeler (p=2).

Notons que l'on connait également des conditions suffisantes pour un problème analogue à (P_s) mais dans lequel on ne suppose pas que A soit symétrique cf [2, 4, 5, 6, 7].

On peut adapter une démonstration de de Oliveira [1] et de Friedland [9] pour obtenir le résultat d'existence, et en quelque sorte d'unicité suivant.

PROPOSITION 1. - Si $d(s) = \min_{i \neq j} |s_i - s_j| > \delta(A) = \max_{i \neq j} |\lambda_i(A) - \lambda_j(A)|$
alors le problème (P_s) possède

i) $n!$ solutions

ii) une et une seule solution $X = (x_i \delta_{ij})$ vérifiant
$x_1 \geqslant x_2 \geqslant \ldots\ldots \geqslant x_n > 0$

La démonstration est basée sur le théorème du point fixe de Brouwer.

3 - Un algorithme du type des approximations successives et un algorithme du type Newton.

Pour $n \geqslant 2$ toutes les conditions suffisantes assurant l'existence proviennent de l'application du théorème de Brouwer ; il est donc naturel de rechercher sous quelles conditions l'algorithme des approximations successives sera convergent. Pour montrer qu'un opérateur est une contraction il est classique d'étudier sa dérivée, ce qui entraîne à regarder la dérivabilité de $x \to \mu(A+X)$.

Si x est tel que $A+X$ n'a que des valeurs propres simples alors en ce point $\mu(A+X)$ est de classe C^∞, d'après Lancaster [13] et Kato [12]. Le fait d'imposer que $A+X$ n'ait que des valeurs propres simples est assez restrictif mais l'on peut donner des exemples où une telle situation a lieu.

Supposons que A soit symétrique, tridiagonale et que $a_{i-1,\,i} \neq 0$ $i=2, \ldots, n$ alors d'après Wilkinson [11] page 300, on sait que pour tout $x \in \mathbb{R}^n$ $A+(x_i \delta_{ij})$ n'aura que des valeurs propres simples. Dans ce cas $x \to \mu(A+(x_i \delta_{ij}))$ appartient à $C^\infty(\mathbb{R}^n)$. Notons que ce cas correspond exactement à la discrétisation de l'opérateur de Sturm-Liouville.

Supposons que $x \to \mu(A+x_i\delta_{ij})$ soit dans $C^1(\Omega)$. Notons $J(x)$ la valeur en x de la matrice jacobienne de $x \to (\mu(A+X))$; d'après Lancaster [13] on obtient :

$$J(x) = \left[\frac{\partial \mu_i(A+X)}{\partial x_j} \right]_{ij} = (u^2_{ji})_{ij}$$

où $U=(u_{ij})_{ij}$ est la matrice orthogonale qui diagonalise la matrice symétrique $A+X$, ie $A+X=U.\ \mathrm{Diag}\ \mu(A+X).\ U^T$. Il est important de remarquer que $J(x)$ est une matrice doublement stochastique.

Pour les propriétés des matrices doublement stochastiques, on pourra consulter Horn [14], Hardy-Littlewood-Polya [15].

ALG 1 : Un algorithme du type approximations successives : Alg 1

C'est Hadeler [2] qui a obtenu les résultats les plus précis sur l'algorithme des approximations successives :

Alg 1 : $$x^{n+1} = x^n + \mu(S) - \mu\,(A+(x^n_i\delta_{ij})) \qquad n \geq 0$$

Reformulons son résultat en introduisant un coefficient de relaxation ω qui assure un meilleur comportement numérique.

PROPOSITION 2. - Soit A appartenant à $\mathcal{M}_{nn}(\mathbb{R})$, symétrique à diagonale nulle.

Si $\min_{i \neq j} |s_i - s_j| \geq 4 \max_i \sqrt{\Sigma a^2_{ij}}$ alors quelque soit $\omega \in]0, 1]$ l'application $T : x \to x + \omega(\mu(S) - \mu(A+X))$ est k-lipschitzienne de constante $k \leq 13/18$ de la boule $B(s\ d(s)/12)$ dans elle même.

Laborde [4] a démontré également que sous l'hypothèse $\min_{i \neq j} |s_i - s_j| > 2\rho(A)$ $\rho(A)$ rayon spectral de A la solution était un point attractif (cf Ortega [16] page 383)pour les approximations successives.

De fait les conditions de Hadeler, aussi bien que celles de Laborde impliquent que sur la solution $\bar{x}$ le jacobien $J(\bar{x})$ est inversible. Cela donne en quelque sorte la limite de leur résultat car il est facile de construire des exemples pour lesquels une solution existe mais dont le jacobien en ce point n'est pas inversible.

Nous n'avons pas réussi à construire d'exemple pour lequel à la fois

$d(s) \geq \ell(A)$ et $J(\bar{x})$ non inversible ; mais cette conjecture semble plausible.

L'avantage majeur de cet algorithme est le fait qu'il n'utilise pas les vecteurs propres ; l'unique opération couteuse est l'extraction des valeurs propres de $A+X^n$ ce que l'on réalise par une méthode du type Q. R. avec shift.

Un autre avantage est sa tendance à conserver l'invariant important pour le problème, qu'est la trace. Appelons défaut de trace à l'itération k le nombre :

$$e_k = \sum_{i=1}^{n} x_i^k - \sum_{i=1}^{n} s_i$$

PROPOSITION 3. - Pour tout ω de $]0, 1]$ considérons l'algorithme

$$x^{n+1} = x^n + \omega(s - \mu(A+(x_i^n \delta_{ij}))) \qquad n \geq 0$$

alors si 1/ $e_0 = 0 \Rightarrow \forall k \geq 0 \quad e_k = 0$

2/ $e_0 \neq 0 \Rightarrow \lim_{k \to \infty} e_k = 0$

ALG 2 : Un algorithme du type Newton : Alg 2.

Le problème à résoudre étant essentiellement celui de la résolution d'un système non linéaire, il est naturel d'envisager l'algorithme de Newton.

PROPOSITION 4. - Supposons :

1/ qu'il existe une solution $\bar{x}$ ie $\mu(A+(\bar{x}_i \delta_{ij})) = \mu(S)$

2/ $f : x \to \mu(A+\bar{X})$ soit de classe C^1 dans un voisinage Ω de $\bar{x}$

3/ $\forall i=1, 2, \ldots . n \quad \Re e\, \rho_i > 0$ ou ρ_i valeurs propres de $J(\bar{x})$

alors $\forall \lambda > 0$ et $\forall \omega \in]0\ 1]$ l'algorithme de Newton

alg 2 : $x^{n+1} = x^n - \omega (J(x_n)+\lambda I)^{-1} (\mu(A+(x_i^n \delta_{ij})) - \mu(S))$

possède $\bar{x}$ comme point d'attraction.

Remarquons que s'il existe une solution $\bar{x}$ pour un spectre visé $Sp(s_i \delta_{ij})$ qui est bien séparé alors nécessairement $f : x \to \mu(A+X)$ est de classe C^1 dans un voisinage Ω de $\bar{x}$; la seule hypothèse restante est la 3°. $J(\bar{x}) = (u_{ji}^2)_{ij}$ est une matrice doublement stochastique, ce qui implique d'après le théorème de Gerchgorin que 1 est toujours la valeur de plus grand module d'une part, d'autre

part que toutes les autres valeurs propres sont contenues dans la réunion pour $i=1, 2, \ldots\ldots, n$ des disques centrés en u_{ii}^2 et de rayon $1-u_{ii}^2$. Tous ces disques seront contenus dans le 1/2 plan $\Re e\, z > 0$ dès que $\forall\, i=1\ 2, \ldots\ldots, n \quad u_{ii}^2 > \frac{1}{2}$. Or d'après Laborde [4] cela est réalisé si $\min |s_i - s_j| > 2\rho(A)$. D'où le corollaire

COROLLAIRE 1. - Si A est une matrice symétrique à diagonale nulle et si $\min |s_i - s_j| > 2\rho(A)$ alors

1/ il existe $\bar{x}$ solution de (Ps)

2/ $\bar{x}$ est un point attractif pour l'algorithme de Newton :

$$\text{Alg 2 : } x^{n+1} = x^n - \omega(J(x^n)+\lambda I)^{-1} -(\mu(A+(x_i^n \delta_{ij}))-\mu(S)).$$

Une autre façon d'obtenir que tout les disques de centre u_{ii}^2 et de rayon $1-u_{ii}^2$ soient dans $\Re e\, z > 0$ et d'imposer que $1-u_{ii}^2 = \sum_{\substack{j=1 \\ j\neq 1}}^{n} u_{ij}^2 < \frac{1}{2}$, puisqu'ils passent tous par le point 1. En adaptant une partie de démonstration de Hadeler [2] on obtient

COROLLAIRE 2. - Si A est symétrique à diagonale nulle et si

$$d(s) = \min_{i\neq j} |s_i - s_j| \geqslant 2\sqrt{3} \max_i \sqrt{\sum_j a_{ij}^2} \quad \text{alors}$$

1/ il existe $\bar{x}$ solution de (Ps)

2/ $\bar{x}$ est un point attractif pour l'algorithme de Newton Alg 2.

On peut résumer ces deux corollaires en disant que les conditions qui assurent la convergence des approximations successives, suffisent pour entraîner la convergence de la méthode de Newton.

L'algorithme de Newton nécessite à chaque étape la connaissance de $J(x^n)$ c'est à dire de toutes les valeurs propres et de tous les vecteurs propres de $A+(x_i^n \delta_{ij})$. C'est un accroissement de la masse des calculs pour chaque itération, de fait lors des essais numériques nous nous sommes bornés à des matrices tridiagonales symétriques et nous avons employé l'algorithme du type Q. R nommé tq 12 dans Wilkinson-Reinsch [17]. Pour contre partie nous obtenons une convergence très rapide, et le fait assez surprenant que pour des approximations initiales qui sont en normes plus éloignées de la solution, que celles nécessaires à la convergence des approximations successives, nous ayons encore convergence. Ce bon

comportement numérique est peut être dû au fait que l'algorithme conserve la trace, ou réduit le défaut de trace.

En effet, on a la :

PROPOSITION 5. - Pour $\omega \in]0\ 2[$ et $\lambda>0$ considérons l'algorithme

$$x^{n+1} = x^n - \omega(J(x^n)+\lambda I)^{-1}(\mu(A+X^n)-\mu(S))$$

alors

$$e_0 = 0 \Rightarrow \forall k\ e_k = 0$$

$$e_0 \neq 0 \Rightarrow \lim_{k\to\infty} e_k = 0$$

Dans la démonstration on utilise le fait que $J(x^n)$ est une matrice doublement stochastique.

4 - Algorithmes de minimisation.

Dès que l'on sait calculer la dérivée de $x \to \mu(A+(x_i\delta_{ij}))$ il est naturel pour approximer la solution de l'équation $\mu(A+x_i\delta_{ij}) = \mu(S)$ de songer à minimiser $f(x) = \frac{1}{2} \| \mu(A+x_i\delta_{ij})-\mu(S) \|_2^2$.

Nous ferons l'hypothèse que A est une matrice tridiagonale à diagonale nulle telle que de plus $a_{i,i-1} \neq 0$ $i=2\ 3\ldots\ldots,n$; cela pour assurer la dérivabilité de $x \to \mu(A+x_i\delta_{ij})$ en tout $x \in \mathbb{R}^n$.

La fonction $f:x \to f(x) = \frac{1}{2} \|\mu(A+x_i\delta_{ij}) - \mu(S)\|_2^2$ n'est pas convexe, mais elle possède de bonnes propriétés vis à vis d'une méthode de gradient. Par construction f est bornée inférieurement par zéro et il résulte d'un calcul facile que son gradient $\nabla f(x)$ en x vaut :

$$\nabla f(x) = J(x)^T \lfloor \mu(A+x_i\delta_{ij})-\mu(S) \rfloor$$

où :

$$J(x) = (u_{ji}^2)_{ij}$$

$U = (u_{ij})$ étant la matrice orthogonale qui diagonalise $A+(x_i\delta_{ij})$.

Notons également que : $\lim f(x) = +\infty$

$\|x\| \to +\infty$

Appelons Alg 3 l'algorithme de plus grande descente décrit par

$$\text{Alg 3} : x^{n+1} = x^n - \rho_n \nabla f(x^n) \qquad n \geq 0$$

Pour assurer la convergence de cet algorithme il reste à faire un choix convergent, au sens de Cea [18], du pas ρ_n.

Notons $\mu(x) = (\mu_1(x), \ldots \mu_n(x))$ le vecteur de $\mathbb{R}^n$ obtenu à partir du vecteur x en renumérotant ses composantes dans l'ordre non croissant. Sur $\mathbb{R}^n \times \mathbb{R}^n$ introduisons après Hardy-Littlewood-Polya la relation $x \mathcal{R} y = x \prec y$ qui est vraie si et seulement si :

$$\forall k = 1, 2, \ldots, n-1 \qquad \sum_{i=1}^{k} \mu_i(x) \leq \sum_{i=1}^{k} \mu_i(y)$$

$$\sum_{1}^{n} \mu_i(x) = \sum_{1}^{n} \mu_i(y)$$

On a alors le résultat suivant dû à Horn [14].

PROPOSITION 6. - Soit $X=(x_i \delta_{ij})$ fixée. Une condition nécessaire et suffisante pour qu'il existe une matrice réelle symétrique A à diagonale nulle telle que $Sp\,(A+X) = (s_1, s_2, \ldots, s_n)$ est que $x \prec s$.

Notons alors W_s l'ensemble des $x \in \mathbb{R}^n$ tels que $x \prec s$ Horn [14] reprenant Hardy-Littlewood-Polya montre que W_s peut encore s'écrire $W_s = \{x \in \mathbb{R}^n \mid x = Ms$ M matrice doublement stochastique$\}$ ce qui prouve que d'après un résultat de Birkoff [23] que W_s est un polyedre convexe compact dont les sommets sont les Ps ; P décrivant l'ensemble des matrices de permutation.

Considérons d'autre part l'orbite $\mathcal{O}(S)$ de $S=(s_i \delta_{ij})$, c'est à dire l'ensemble des matrices orthogonalement semblables à S.

$$\mathcal{O}(S) = \{B \in \mathcal{M}_m(\mathbb{R}) \mid B = USU^T \qquad U \text{ orthogonale}\}$$

est un ensemble compact, mais non convexe. Il est clair que si x est une solution de (Ps) alors $A + x_i \delta_{ij} \in \mathcal{O}(S)$.

Sur l'ensemble des matrices symétriques considérons le produit scalaire $(A\ ,B) = tr\,(AB)$ et la norme induite, dite norme de Schur $\|A\|_S^2 = tr\,A^2 = \sum_{i,j=1}^{n} a_{ij}^2$.

L'ensemble des matrices symétriques est alors un espace de Hilbert qui peut se décomposer en somme directe orthogonale entre les matrices diagonales et les matrices symétriques à diagonale nulle. D'après le théorème de Wiedlant-Hoffman [21] la distance de $A+x_i\delta_{ij}$ à $\mathcal{O}(S)$ est donnée par $\|\mu(A+x_i\delta_{ij})-\mu(S)\|_2$ c'est à dire que f(x) représente au facteur 1/2 près le carré de la distance de A+X à $\mathcal{O}(S)$.

Notons $A+W_s$ l'ensemble convexe compact des matrices $A+x_i\delta_{ij}$ où $x\in W_s$. $A+W_s$ est contenu dans un hyperplan passant par A parallèle à l'ensemble des matrices diagonales. Il est clair que résoudre (P_s) c'est trouver un point de l'intersection $\mathcal{O}(S)\cap(A+W_s)$, et qu'une méthode constructive sera l'obtention d'une suite minimisant la distance. Gubin-Polyak-Raik [19], Pierra [20] ont développé des algorithmes de projection successives pour trouver un point de l'intersection de plusieurs convexes ; reprenons cette idée en l'adaptant.

Soit $A^k = A+X^k = A+(x_i^k\,\delta_{ij})$ une matrice de $(A+W_s)$. D'après le théorème de Wiedlandt-Hoffman [21] la distance de A^k à $\mathcal{O}(S)$ est donnée par :

$$\text{dist}\,(A^k,\mathcal{O}(S) = \min_{B\in\mathcal{O}(S)} \|A^k-B\|_s = \|\mu(A+X^k)-\mu(S)\|_2$$

Car si $A^k=A+X^k=M^k.\ \text{Diag}\,\mu(A+X^k).\ {M^k}^T$, la matrice $B^k=M^k\ \text{Diag}\,\mu(S)\ {M^k}^T$ réalise le minimum de la distance. En d'autres termes, on sait projeter sur $\mathcal{O}(S)$. Notons que $\mathcal{O}(S)$ n'étant pas convexe il peut exister plusieures projections, mais notre façon de procéder en détermine une seule.

La détermination de la projection $A^{k+1}=A+X^{k+1}$ de B^k sur le convexe compact $(A+W_s)$ est particulièrement simple, d'après la proposition 6, il vient :

$$\underset{(A+W_s)}{\text{Proj}}\ B^k = A+\text{Diag}\,B^k$$

car diag $B^k\in W_s$.

On appellera algorithme des projections successives ou Alg 4 l'itération des deux étapes suivantes.

a) $B^k = M^k\ \text{Diag}\mu(S)\ {M^k}^T$ si $A^k\ A+X^k=M^k.\ \text{Diag}\,(A+X^k).\ {M^k}^T$

b) $A^{k+1} = A+X^{k+1} = A+\text{Diag}\,B^k$.

On a par construction la proposition suivante :

PROPOSITION 7. - Si $f(x) = \frac{1}{2}\ \|\mu(A+x_i\delta_{ij})-\mu(S)\|_2^2$ alors pour la suite

$\{x^k\}_1^\infty$ fournit par Alg 4 on a $f(x^{k+1}) \leqslant f(x^k)$

Explicitons le passage de X^n à X^{n+1} dans l'algorithme précédent.

On a :

$$X^{n+1} = X^n + \text{Diag}\,\{M^n(\text{Diag}\,\mu(S) - \text{Diag}\,\mu\,(A+X^n))\,M^{n\,T}\}.$$

Cette écriture prouve que l'algorithme Alg 4, de projections successives est l'algorithme de O. Hald [9].

Multiplions à droite chaque terme par e où $e^T=(1,1,1,\ldots\ldots,1)$; il vient :

$$X^{n+1}\,e = x^{n+1} + \text{Diag}\,\{M^n(\text{Diag}\,\mu(S) - \text{Diag}\,\mu(A+X^n)\,M^{n\,T}\}\,e$$

d'où

$$x^{n+1} = x^n + J(x^n)^T\,\{\mu(S)-\mu(A+x_i^n\delta_{ij})\} = x^n - \nabla f(x^n).$$

Il y a donc coïncidence entre l'algorithme de double projection Alg 4, l'algorithme de Hald, la méthode de plus grande descente Alg 3 avec le choix du pas $\rho_n=1$. C'est la concordance de ces trois méthodes qui va permettre de prouver la convergence.

LEMME. - (Hald [9] page 162). Pour la suite de matrices diagonales obtenues par l'algorithme de Hald [resp Alg 3, Alg 4] on a

$$\sum_{n>0} \|X^{n+1}-X^n\|_s^2 < +\infty$$

Il est connu que ce résultat n'implique pas à lui seul la convergence de la suite $\{x^n\}_1^\infty$. C'est notre interprétation comme méthode de gradient qui permet de conclure.

PROPOSITION 8. - Soit $A \in \Pi_{nn}(\mathbb{R})$ symétrique à diagonale nulle

Soit $S=(s_i\delta_{ij})$ fixée

Alors

1/ il y a coïncidence entre les trois algorithmes de Hald, des projections successives : Alg 4, de gradient à pas fixe $\rho_n=1$: Alg 3

ie. $x^{n+1}=x^n-J(x^n)^T\{\mu(A+x_i^n\delta_{ij})-\mu(S)\}$

2/ tout point adhérant $\bar{x}$ à la suite $\{x^n\}_1^\infty$ est un point stationnaire pour $f(x)=\frac{1}{2}\|\mu(A+x_i\delta_{ij})-\mu(S)\|_2^2$.

De même que l'interprétation de l'algorithme de Hald comme algorithme de descente fournit des variantes, l'interprétation géométrique suscite de même des variantes numériquement intéressantes.

De la relation

$$A+X=M\ \mathrm{Diag}\,\mu(S)\ M^T$$

on déduit en prenant le carré de la norme de Schur des deux membres que

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} s_i^2 - \sum_{i,j=1}^{n} a_{ij}^2 = r^2$$

Si une solution x au problème (Ps) existe alors, nécessairement, $x \in S(o, r) \subset B(o, r)$. posons

$$A+B(o, r)=\{M \in \mathbb{M}_{nn}(\mathbb{R}) \mid M=A+(x_i \delta_{ij}) \ , \ \sum_{i=1}^{n} x_i^2 \leqslant r^2\}.$$

Cet ensemble est contenu dans hyperplan parallèle à l'ensemble des matrices diagonales et passant par A.

La solution si elle existe appartient à $(A+B(o, r)) \cap \mathcal{O}(S)$; il est facile d'adapter l'algorithme des projections successives.
Soit :

$$A^k = A+X^k = M^k = M^k.\ \mathrm{Diag}\,\mu(A+X^k).\ M^{k\,T}$$

la matrice $C^k=M^k\ \mathrm{Diag}\,\mu(S)M^{k\,T}$ réalise le minimum de la distance entre A^k et $\mathcal{O}(S)$. La projection de C^k sur le convexe compact $(A+B(o, r))$ est facile à obtenir ; c'est $A^{k+1} = A+X^{k+1} = A+r\ \frac{\mathrm{Diag}\ C^k}{\|\mathrm{Diag}\ C^k\|_2}$. On itère ces deux projections successives .

L'intérêt dans cet algorithme est la conservation à priori de la norme de la solution.

Il est clair que la suite $\{x^n\}_1^\infty$ obtenue définie un algorithme de descente pour $f(x) = \frac{1}{2}\ \|\mu(A+x_i\delta_{ij})-\mu(S)\|_2^2$, c'est à dire $f(x^n) \geqslant f(x^{n+1})$.
De fait cet algorithme n'est rien d'autre que la minimisation de f(x) sous la contrainte $x \in B = B(o, r) = \{x \mid \|x\| \leqslant r\}$ par une méthode de gradient projeté.

On obtient alors le résultat suivant :

PROPOSITION 9. - Soit A tridiagonale symétrique à diagonale nulle telle que

$a_{i-1,i} \neq 0 \quad i=2, 3, \ldots, n$

alors si

$$K = \sup_{\xi \in \mathbb{R}^n} \|f''(\xi)\| < 2$$

où

$$f(x) = \frac{1}{2} \left\| \mu' A + x_i \delta_{ij} \right) - \mu(S) \right\|_2^2$$

l'algorithme $x^{n+1} = \mathrm{Proj}(x^n - \nabla f(x^n))$ produit une suite dont tous les points adhérants sont des points stationnaires de f.

5 - Essais numériques.

Nous avons fait de nombreuses expériences numériques, aussi nous n'en présentons ici qu'une partie. Dans les essais suivants nous aurons $\omega = 1$ Alg 1, $\omega = 1$ et $\lambda = 0$. $1/(n^\circ$de l'itération) dans Alg 2.

Nous avons également multiplié les essais pour étudier le domaine d'attraction d'une solution relativement aux divers algorithmes. L'expérience a révélé un fait assez inhabituel : l'ensemble des approximations initiales pour lesquelles l'algorithme de Newton Alg 2 converge semble plus grand que l'ensemble des approximations initiales assurant la convergence de l'algorithme des approximations successives Alg 1. Naturellement lorsque tous les deux convergent Alg 2 est bien plus rapide que Alg 1.

Pour comparer l'efficacité des divers algorithmes étudiés nous avons construit des problèmes tests à partir de la discrétisation avec un pas $h=1/(n+1)$, de problèmes de Sturm-Liouville.

$$Ly = y'' + q(x)y = \lambda y$$

$$\alpha_1 y(0) - \beta_1 y'(0) = 0$$

$$\alpha_2 y(1) + \beta_2 y'(1) = 0$$

On approxime $y''(x_i) = y''(ih)$ par $(y_{i-1} - 2y_i + y_{i+1})/h^2$, $y'(0)$ par $(y_1 - y_0)/h$ et $y'(1)$ par $(y_{n+1} - y_n)/h$. Il est alors facile de vérifier que l'approximation de l'opérateur L est la matrice A symétrique tridiagonale ayant des -1 sur les deux codiagonales et $\{C_a + (2+h^2 q_1),\ 2+h^2 q_2, \ldots, 2+h^2 q_{n-1}, (2+h^2 q_n) + C_b\}$ comme diagonale, avec $C_a = -\beta_1/(\alpha_1 h + \beta_1)$ et $C_b = -\beta_2/(\alpha_2 h + \beta_2)$. La procédure est alors la suivante. On se donne une fonction q et les constantes $\alpha_1, \alpha_2, \beta_1, \beta_2$ pour n=10 on obtient des matrices 10 x 10. On calcule alors le spectre de cette matrice par l'algorithme du type Q. R que Wilkinson-Reinsh nomme tq 12 cf (17).

Les programmes pour les calculs des jeux d'essais et les algorithmes Alg 1, Alg 2, Alg 3, Alg 4 sont écrits en Fortran et testés sur IRIS 80 de CII. Pour chaque programme la phase essentielle du calcul des valeurs propres et vecteurs propres est effectuée par le sous programme tq 12.

Nous reproduisons dans les tableaux ci-dessous deux séries d'essais ; l'une est construite à partir d'une fonction $q(x)=x(1-x)$, c'est à dire symétrique sur $]0,1[$, l'autre à partir de $q(x)=1-x$. Pour chaque série nous faisons varier les conditions aux limites.

i) $y(o) = y(1) = 0$

ii) $y'(o) = o$ et $y(1) = 0$

iii) $y(o) = y'(o)$ et $y(1) = y'(1)$

Nous donnons pour diverses itérations l'erreur relative :

$$\| \text{Spectre visé - Spectre obtenu} \|_2 \; / \; \| \text{Spectre visé} \|_2 .$$

L'approximation initiale de la diagonale est pour tous les essais le spectre visé.

Tableau 1 -$q(x) = x(1-x)$; $y(o) = y(1) = 0$

N°	Alg 1	Alg 2	Alg 3	Alg 4
0	0.260	0.260	0.260	0.260
10	$0.351\ 10^{+1}$	$0.414\ 10^{-4}$	$0.196\ 10^{-1}$	$0.174\ 10^{-1}$
20	diverge	$0.373\ 10^{-5}$	$0.106\ 10^{-1}$	$0.940\ 10^{-2}$
50		$0.153\ 10^{-5}$	$0.494\ 10^{-2}$	$0.421\ 10^{-2}$
80		$0.381\ 10^{-7}$	$0.322\ 10^{-2}$	$0.272\ 10^{-2}$

Tableau 2 $q(x) = x(1-x)$; $y'(o)=0 \quad y(1)=0$

N°	Alg 1	Alg 2	Alg 3	Alg 4
0	0.268	0.268	0.268	0.268
10	$0.972\ 10^{+1}$	0.257	$0.234\ 10^{-1}$	$0.209\ 10^{-1}$
20	diverge	$0.159\ 10^{-9}$	$0.160\ 10^{-1}$	$0.141\ 10^{-1}$
50		arrêt	$0.577\ 10^{-2}$	$0.463\ 10^{-2}$

Tableau 2 (suite)

N°	Alg 1	Alg 2	Alg 3	Alg 4
80			$0.150\ 10^{-2}$	$0.117\ 10^{-2}$

Tableau 3 $q(x) = x(1-x); y(o) = y'(o); y(1) = y'(1)$

N°	Alg 1	Alg 2	Alg 3	Alg 4
0	0.280	0.280	0.280	0.280
10	$0.112\ 10^{+3}$	0.179	$0.427\ 10^{-1}$	$0.408\ 10^{-1}$
20	diverge	0.128	$0.372\ 10^{-1}$	$0.336\ 10^{-1}$
50		$0.253\ 10^{-5}$	$0.115\ 10^{-1}$	$0.864\ 10^{-2}$
80		$0.669\ 10^{-7}$	$0.386\ 10^{-2}$	$0.351\ 10^{-2}$

Tableau 4 $q(x) = 1-x\ ;\ y(o) = y(1) = 0$

N°	Alg 1	Alg 2	Alg 3	Alg 4
0	0.259	0.259	0.259	0.259
10	$0.710\ 10^{+1}$	$0.282\ 10^{-4}$	$0.195\ 10^{-1}$	$0.174\ 10^{-1}$
20	diverge	$0.250\ 10^{-5}$	$0.106\ 10^{-1}$	$0.941\ 10^{-2}$
50		$0.584\ 10^{-4}$	$0.494\ 10^{-2}$	$0.422\ 10^{-2}$
80		$0.345\ 10^{-6}$	$0.322\ 10^{-2}$	$0.272\ 10^{-2}$

Tableau 5 $q(x) = 1-x\ ;\ y'(o) = 0\ ;\ y(1) = 0$

N°	Alg 1	Alg 2	Alg 3	Alg 4
0	0.268	0.268	0.268	0.268
10	0.976	0.577	$0.235\ 10^{-1}$	$0.408\ 10^{-1}$
20	diverge	0.161	$0.161\ 10^{-1}$	$0.136\ 10^{-1}$
50		$0.196\ 10^{-1}$	$0.573\ 10^{-2}$	$0.863\ 10^{-2}$
80		$0.145\ 10^{-9}$ en N° 60	$0.386\ 10^{-2}$	$0.118\ 10^{-2}$

Tableau **6** $q(x) = 1-x$; $y(o) = y'(o)$; $y(1)=y'(1)$

N°	Alg 1	Alg 2	Alg 3	Alg 4
0	0.279	0.279	0.279	0.279
10	$0.140\ 10^{+2}$	0.314	$0.426\ 10^{-1}$	$0.208\ 10^{-1}$
20	diverge	$0.219\ 10^{-1}$	$0.371\ 10^{-1}$	$0.140\ 10^{-1}$
50		$0.134\ 10^{-5}$	$0.114\ 10^{-1}$	$0.464\ 10^{-2}$
80		$0.157\ 10^{-6}$	$0.386\ 10^{-2}$	$0.118\ 10^{-2}$

A la vue de ces résultats deux remarques au moins s'imposent. La plus importante est que pour l'approximation initiale choisie l'algorithme Alg 1 diverge à chaque fois tandis que les autres convergent ; cela corrobore une remarque déjà faite.

Le seconde remarque consiste en l'opposition entre d'une part les deux algorithmes de minimisation Alg 3 et Alg 4 et d'autre part l'algorithme de Newton Alg 2; Alg 3 et Alg 4 donnent des résultats très similaires avec un très léger avantage pour Alg 4. Mais pour ces deux méthodes la convergence bien que très régulière est aussi très lente ; on n'arrive pas en 80 itérations à dépasser le seuil d'une erreur relative en 10^{-3}, ce qui dans notre cas assure 10 pour cent d'erreur sur la plus petite composante et 0,25 pour cent d'erreur sur la plus grande composante du spectre visé. Par contre Alg 2 donne à chaque fois une erreur relative de l'ordre de 10^{-7}, ce qui assure sept chiffres caractéristiques exacts pour toutes les composantes du spectre visé.

La régularité des algorithmes 3 et 4 et la rapidité de l'algorithme de Newton incite à étudier un algorithme pour le problème inverse des valeurs propres qui aurait ces deux excellentes propriétés.

BIBLIOGRAPHIE

[1] De OLIVEIRA G. - Note on inverse characteristic problem. Numer. Math Vol 15, (1970), 339-341.

[2] HADELER K. P. - Ein inverses Eigen wert problem. Linear algebra and its appl. Vol 1, (1968), 83-101.

[3] HADELER K. P. - Newton-Verfahren für inverse Eigenwertaufgen. Num. Math. Vol 12, (1968), 35-39.

[4] LABORDE F. - Sur un problème inverse de valeurs propres. CRAS tome 268, (1969), 153-156.

[5] CHATELIN -LABORDE F. - Thèse Mathodes numériques de calcul de valeurs propres et vecteurs propres d'un opérateur linéaire. Grenoble 1971.

[6] MOREL P. - A propos d'un problème inverse de valeurs propres. GRAS tome 277, (1973), 125-128.

[7] MOREL P. - Sur le problème inverse des valeurs propres. Numer. Math 23, (1974), 83-94.

[8] HALD O. - On discrete and numerical inverse Sturin-Liouville problems, Uppsala University, Dep. of Computer Sciences, Report 42, 1972.

[9] FRIEDLAND S. - Matrices with prescribed off diagonal elements, Israel J. of Math. Vol 11, (1972), 184-189.

[10] FRIEDLAND S. - Inverse eigenvalue problems. A paraître.

[11] WILKINSON - The algebric eigenvalue problem. Oxford University Press (1965).

[12] KATO T. - Perturbation theory of linear operators. Springer Verlag (1966).

[13] LANCASTER P. - On eigenvalues of matrices dependent on a parametor. Numer. Math Vol 6, (1964), 377-387.

[14] HORN A. - Doubly stochastic matrices and the diagonal of a rotation matrice. Amer. J. Math. Vol 76, (1954), 620-630.

[15] HARDY-LITTLEWOOD-POLYA - Inegalities, Cambridge University Press (1948).

[16] ORTEGA-RHEINBOLDT - Iterature solution of non linear equations in several variables, Academic Press (1970).

[17] WILKINSON-REINSCH - Linear algebra. Handbook for compotations. Springer Verlag.

[18] CEA J. - Optimisation, théorie et algorithmes. Dunod 1971.

[19] GUBIN-POLYAK-RAIK - The method of projections for fonding the common point of convex sets. USSR Comp. Math and Math Phys. Vol 6, (1967), 1-24.

[20] PIERRA G. - Sur le croissement de méthodes de descente. CRASS t 277, (1973) 1071-1074.

[21] WIELANDT-HOFFMAN - The variation of the spectrum of a normal matrix, Doke J. of Math Vol 20, (1953), 37-39.

[22] GOLSTEIN A. - Constructive real analysis. Harper International Edition (1967).