Des Algorithmes pour le Problème Inverse des Valeurs Propres

Paul Morel
Université de Bordeaux I

Recommandé par N. Gastinel

RESUME

Le problème inverse des valeurs propres est la recherche d'une matrice diagonale X, telle que A étant une matrice donnée, la matrice $A+X$ à un spectre fixé. On rappelle d'abord certains résultats connus relatifs à ce problème. On démontre ensuite un résultat d'existence, puis des résultats de convergence et de stabilité numérique pour des algorithmes de résolution approchée du type approximations successives et de Newton. On obtient des interprétations géométriques et algébriques d'un algorithme de minimisation dû à O. Hald, ce qui permet, d'une part de faire une démonstration convaincante de sa convergence, et d'autre part d'envisager des variantes numériquement intéressantes. On compare enfin l'efficacité des divers algorithmes étudiés sur des problèmes tests.

1. INTRODUCTION

On appellera problème (P_s) la recherche d'une matrice diagonale réelle $X=(x_i\delta_{ij})\in\mathfrak{M}_{nn}(\mathbf{R})$ telle que A étant une matrice donnée symétrique de $\mathfrak{M}_{nn}(\mathbf{R})$, le spectre de $A+X$: $\mathrm{Sp}(A+X)$ soit égal au spectre de la matrice fixée $S=(s_i\delta_{ij})$ de $\mathfrak{M}_{nn}(\mathbf{R})$.

Le but de ce travail est double. Premièrement nous voulons étudier, mettre en oeuvre et comparer divers algorithmes pour résoudre le problème (P_s). Deuxièmement, nous voulons obtenir des interprétations géométriques et algébriques d'un algorithme dû à O. Hald pour, d'une part avoir une démonstration simple et rigoureuse de sa convergence et, d'autre part, en déduire des variantes numériquement intéressantes.

Le problème (P_s) est la version discrétisée de la recherche d'une fonction q de $\mathbf{R}$ dans $\mathbf{R}$ telle que l'opérateur différentiel $L(y)=-y''+q(x)y$ auquel on joint des conditions aux limites du type $y(0)=y(1)=0$ admette un spectre fixé à l'avance. On peut faire deux hypothèses ne diminuant en rien la généralité du problème. On peut supposer que la diagonale de la matrice A

est nulle. En effet $\mathrm{Sp}(A+X)=\mathrm{Sp}(X)$ si et seulement si $\mathrm{Sp}(A'+X')=\mathrm{Sp}(S)$ où $A'=A-\mathrm{diag}A$ et $X'=X+\mathrm{diag}A$.

On peut supposer que le spectre visé est positif. En effet $\mathrm{Sp}(A+X)=\mathrm{Sp}(X)$ si et seulement si $\mathrm{Sp}(A+X+tI)=\mathrm{Sp}(S+tI)$ où $t\in\mathbf{R}$ et I est la matrice identité. Nous incluons donc ces deux hypothèses, non restrictives, dans la formulation du problème (P_s).

Dans le paragraphe 2 nous reprendrons une démonstration de Oliveira [1] pour obtenir un résultat d'existence et unicité et nous le situerons par rapport aux autres résultats du même type obtenus par Haldeler [2, 3], Laborde [4, 5] et Morel [6, 7].

Le paragraphe 3 montre que les hypothèses de Laborde [4] et Haldeler [2] assurant la convergence des approximations successives impliquent aussi la convergence de l'algorithme Newton.

Dans le paragraphe 4 nous prouvons la convergence d'algorithmes de minimisation et nous interprétons un algorithme dû à O. Hald [8].

Dans le dernier paragraphe nous comparerons sur des problèmes tests, bâtis à partir de la discrétisation de problèmes de Sturm-Liouville, l'efficacité numérique des algorithmes présentés.

2. CONDITIONS SUFFISANTES ET CONDITIONS NECESSAIRES

S. Friedland [9, 10] démontre que pour toutes matrices A de $\mathfrak{M}_{nn}(\mathbf{C})$ et $S=(s_i\delta_{ij})$ de $\mathfrak{M}_{nn}(\mathbf{C})$, il existe au moins une et au plus n ! matrice diagonale $X=(x_i\delta_{ij})$ de $\mathfrak{M}_{nn}(\mathbf{C})$ telle que $\mathrm{Sp}(A+X)=\mathrm{Sp}(X)$.

Sa démonstration dans [9], comporte plusieurs étapes. En employant le *Nullstellensatz* de Hilbert, il montre que le système

$$x_i^{m_i}+P_i(x_1,x_2,\dots,x_n)=0 \qquad i=1,2,\dots,n,$$

où le degré de chaque polynôme P_i est inférieur à m_i possède toujours des solutions dans $\mathbf{C}$. Il remarque ensuite que son problème se réduit à la résolution d'un système d'équations polynomiales justiciable du résultat précédant. Cette démonstration faisant appel à des propriétés des idéaux de l'anneau des polynomes à n indéterminées n'est pas constructive.

L'étude en dimension 2×2 montre immédiatement que le problème (P_s) ne possède pas toujours de solution. K. Haldeler [2], Laborde [4] et P. Morel [6, 7], ont donné des conditions nécessaires de plus en plus précises pour que le problème (P_s) possède des solutions. On montre dans Morel [7] que nécessairement $S=(s_i\delta_{ij})$ doit vérifier

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

ce qui met en évidence une nécessaire séparation du spectre visé. 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 $x \to \mu(A + (x_i\delta_{ij}))$, 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

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

où $(\mu^{<}(S) \mid \mu(A))$ est le produit scalaire entre les vecteurs $\mu^{<}(S)$ et $\mu(A)$ qui désignent respectivement les valeurs propres de S dans l'ordre non décroissant et les valeurs propres de A dans l'ordre non croissant.

Ces conditions nécessaires, qui peuvent servir à choisir correctement une approximation initiale lors de la mise en oeuvre d'un algorithme ne sont suffisantes que si la dimension est inférieure ou égale à 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],

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

Elle recouvre celle de Laborde ($p = +\infty$) et rappelle sans être pourtant identique la première connue, celle de K. Hadeler ($p = 2$) et celle de Oliveira ($p = 1$).

Adaptons une démonstration de Oliveira [1] pour obtenir le résultat d'existence et en quelque sorte d'unicité suivant

Proposition 1. *Si*

$$d(s) = \min_{i \neq j} |s_i - s_j| > l(A) = \max_{i \neq j} |\lambda_i(A) - \lambda_j(A)|,$$

alors le problème (P_s) *possède*

(i) $n!$ *solutions,*
(ii) *une et une seule solution* $X = (x_i\delta_{ij})$ *vérifiant*

$$x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n > 0.$$

Démonstration. Soit $\lambda(S)=\{\lambda_1(S),\lambda_2(S),\dots,\lambda_n(S)\}$ le vecteur de $\mathbf{R}^n$ obtenu en adoptant une numérotation quelconque des éléments de $\mathrm{Sp}(s_i\delta_{ij})$, et soit P la matrice de permutation telle que $P\lambda(S)=\mu(S)$, c'est-à-dire qui ordonne dans l'ordre noncroissant. La condition $d(s)>l(A)$ impose que les intervalles $J_i=[\lambda_i(S)-\mu_i(A),\lambda_i(S)-\mu_n(A)]$ sont deux à deux disjoints.

Posons $\pi=J_1\times J_2\times\cdots\times J_n$. π est convexe compact. Si $x\in\pi$ on aura d'une part $Px=\mu(x)$, et d'autre part $P^{-1}\mu(x)=x$, où $\mu(x)$ représente le vecteur dont les composantes sont celles de x mais dans l'ordre non croissant. Pour $A+X=A+\mathrm{diag}(x_1,x_2,\dots,x_n)$ d'après le théorème du min-max (cf. Wilkinson [19, p. 102]), on a

$$\mu_i(x)-\mu_n(A)\leqslant\mu_i\big(A+(x_i\delta_{ij})\big)\leqslant\mu_i(x)+\mu_1(A)\qquad i=1,2,\dots,n,$$

d'où

$$\mu_n(A)\leqslant\mu_i(A+X)-\mu_i(x)\leqslant\mu_i(A)\qquad i=1,2,\dots,n.$$

Soit alors Tp l'opérateur analogue de celui employé pour la première fois par K. Hadeler [2], défini par

$$\mathrm{Tp}:x\to\mathrm{Tp}(x)=x+\lambda(S)-P^{-1}\mu\big(A+(x_i\delta_{ij})\big).$$

Tp est continu, car d'après Kato [12, p. 124] $x\to\mu(A+(x_i\delta_{ij}))$ est continue. $\mathrm{Tp}(\pi)\subset\pi$ car

$$(\lambda(S)-\mathrm{Tp}(x))_i=P^{-1}\mu\big(A+(x_i\delta_{ij})\big)-x_i=\big(P^{-1}\big[\mu\big(A+(x_i\delta_{ij})\big)-\mu(x)\big]\big)_i.$$

Ce qui prouve que $(\mathrm{Tp}(x))_i\in J_i$ d'après les inégalités du min-max citées.

Le théorème de Brouwer permet alors d'affirmer qu'il existe $x_p\in\pi$ point fixe de Tp. Il est clair que la condition $d(s)>l(A)$ est indépendante de la numérotation des éléments de $\mathrm{Sp}(S)$ et la démonstration donnée ici prouve qu'il y a au moins une solution pour chaque numérotation, chacune étant en quelque sorte indicée par la permutation qui l'envoie dans le cône

$$C=\{x\in\mathbf{R}^n|x_1\geqslant x_2\geqslant\cdots\geqslant x_n>0\}.$$

L'unicité est obtenue en se rappelant que Friedland a montré qu'il existait $n!$ solutions au plus. ■

3. UN ALGORITHME DU TYPE DES APPROXIMATIONS SUCCESSIVES ET UN ALGORITHME DU TYPE DE NEWTON

Pour $n \geqslant 2$ toutes les conditions suffisantes assurant l'existence proviènnent 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.

Soit $\Omega \subset \mathbf{R}^n$ tel que pour $x \in \Omega$ la matrice $A+(x_i\delta_{ij})$ ne possède que des valeurs propres simples; alors d'après Lancaster [13], l'application $x \to \mu(A+X)=\mu(A+(x_i\delta_{ij}))$ est indéfiniment différentiable dans Ω. Supposons que A soit symétrique, tridiagonale et que $a_{i-1,i} \neq 0$, $i=2,3,\ldots,n$. Alors d'après Wilkinson [11, p. 300] on sait que pour tout $x \in \mathbf{R}$, la matrice $A+(x_i\delta_{ij})$ n'a que des valeurs propres simples; donc $x \to \mu(A+x_i\delta_{ij})$ appartient à $C^{\infty}(\mathbf{R}^n)$. Notons que si le problème (P_s) provient de la discrétisation d'un problème de Sturm-Liouville, nous sommes dans cette situation. Supposons que

$$x \in \Omega = \left\{ x \in \mathbf{R}^n \min_{i \neq j} |x_i - x_j| > \max_{i \neq j} |\lambda_i(A) - \lambda_j(A)| \right\},$$

alors d'après les inégalités du min-max il est facile de voir que dans ce cas $A+x_i\delta_{ij}$ n'a que des valeurs propres simples et donc que $x \to \mu(A+x_i\delta_{ij})$ appartient à $C^{\infty}(\Omega)$.

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:

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

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

C'est Hadeler [2] qui le premier a appliqué l'algorithme des approximations successives

$$\text{Alg. 1.} \qquad x^{n+1} = x^n + \mu(S) - \mu(A+(x_i^n\delta_{ij})) \qquad n \geqslant 0$$

à la résolution du problème (P_s).

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

PROPOSITION 2. *Soit A appartenant à* $\mathfrak{M}_{nn}(\mathbf{R})$, *symétrique à diagonale nulle. Si*

$$\min_{i\neq j}|s_i-s_j|\geqslant 4\max_i\sqrt{\sum_j a_{ij}^2}$$

alors quelque soit $\omega\in]0,1]$ *l'application* $T:x\to x+\omega(\mu(S)-\mu(A+X))$ *est k-lipschitzienne de constante* $k\leqslant\frac{13}{18}$ *de la boule* $B(s,d(s)/12)$ *dans elle même.*

Laborde [4] a démontré également que sous l'hypothèse $\min_{i\neq j}|s_i-s_j|>2\max_i|\lambda_i(A)|$ la solution était un point attractif (cf. Ortega [16, p. 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.

L'avantage majeur de cet algorithme est le fait qu'il n'utilise pas les vecteurs propres; l'unique opération coûteuse est l'extraction des valeurs propres de $A+X^n$ ce que l'on réalise par une méthode du type Q.R. Un autre avantage est sa tendance à converser l'invariant important pour le problème, qu'est la trace. Appelons défaut de trace à l'itération k le nombre $e_k=\sum_{i=1}^n(x_i^k-s_i)$.

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

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

alors si

(i) $e_0=0\Rightarrow\forall k\geqslant 0,\ e_k=0,$
(ii) $e_0\neq 0\Rightarrow\lim_{k\to\infty}e_k=0.$

Démonstration. Cela n'est rien d'autre qu'une propriété de la trace, i.e.,

$$\sum_i\mu_i\big(A+\big(x_i^n\delta_{ij}\big)\big)=\sum_i x_i^n,$$

d'où

$$\sum_i x_i^{n+1}=\sum_i x_i^n+\omega\Big(\sum_i s_i-\sum_i\mu_i(A+X^n)\Big),$$

donc

$$e_{n+1}=(1-\omega)e_n.$$

On constatera sur les essais numériques ce bon comportement. ■

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

PROPOSITION 4. Supposons

(1) qu'il existe une solution $\bar{x}$, i.e. $\mu(A+(\bar{x}_i\delta_{ij}))=\mu(S)$,
(2) $f:x\to\mu(A+X)$ soit de classe C^1 dans un voisinage Ω de $\bar{x}$,
(3) $\forall i=1,2,\dots,n$, $\operatorname{Re}\rho_i>0$ 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. $$x^{n+1}=x^n-\omega\left[J(x^n)+\lambda I\right]^{-1}\left[\mu(A+(x_i^n\delta_{ij}))-\mu(S)\right]$$

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

Démonstration. Considérons l'application g de Ω dans $\mathfrak{M}_{nn}(\mathbf{R})$ définie par $g:x\to(1/\omega)[J(\bar{x})+\lambda I]$. $J(\bar{x})$ est inversible d'après (3), ainsi $g(\bar{x})$ est inversible pour $\lambda>0$ et $\omega>0$. L'application

$$G:x\to G(x)=x-\omega\left[J(x)+\lambda I\right]^{-1}\left[\mu(A+X)-\mu(S)\right]$$

est donc définie dans un voisinage Ω' de $\bar{x}$. Comme $f:x\to\mu(A+X)$ est de classe C^1, alors G est dérivable en $\bar{x}$ de dérivée

$$G'(\bar{x})=I-\omega\left[J(\bar{x})+\lambda I\right]^{-1}J(\bar{x}).$$

Il est facile de vérifier que les

$$\nu_i=\frac{\lambda+(1-\omega)\rho_i}{\lambda+\rho},\qquad i=1,2,\dots,n$$

sont les valeurs propres de $G'(\bar{x})$. Une condition suffisante pour que $\bar{x}$ soit un point attractif est que le rayon spectral de $G'(\bar{x})$ soit inférieur à un, ce qui s'écrit

$$|\nu_i|^2<1,\qquad i=1,2,\dots,n,\qquad \text{i.e.}\quad -(2-\omega)\omega|\rho_i|^2<2\lambda\omega\operatorname{Re}\rho_i$$

et il est clair que ces conditions sont réalisées dès que $\lambda>0$, $\omega\in]0,2[$ et $\mathrm{Re}\,\rho_i>0$. ■

Remarquons que s'il existe une solution $\bar{x}$ pour un spectre visé $\mathrm{Sp}(s_i\delta_{ij})$ 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). $JC(\bar{x})=(u_{ji}^2)_{ij}$ 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 $i=1,2,\dots,n$ des disques centrés en u_{ii}^2 et de rayon $1-u_{ii}^2$. Tous ces disques seront contenus dans le $\frac{1}{2}$ plan $\mathrm{Re}\,z>0$ dès que $\forall i=1,2,\dots,n$, $u_{ii}^2>\frac{1}{2}$. 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*:

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

Une autre façon d'obtenir que tous les disques de centre u_{ii}^2 et de rayon $1-u_{ii}^2$ soient dans $\mathrm{Re}\,z>0$ et d'imposer que $1-u_{ii}^2=\sum_{j=1,j\neq i}^{n}u_{ij}^2<\frac{1}{2}$, puisqu' ils passent tous par le point 1. Or Hadeler [2] dans le cours de sa démonstration obtient que si $\min|x_i-x_j|\geqslant d$ avec $d\geqslant 2K(A)=2\max_i\sqrt{\sum a_{ij}^2}$ alors

$$\sum_{\substack{i=1\\ i\neq j}} u_{ij}^2<\frac{d-\sqrt{d^2-4K(A)^2}}{d+\sqrt{d^2-4K(A)^2}}.$$

Si alors x vérifie $\min|x_i-x_j|\geqslant(3/\sqrt{2}\,)K(A)$ le jacobien $J(x)=(u_{ji}^2)_{ij}$ est à diagonale dominante.

Supposons que $d(s)=\min_{i\neq j}|s_i-s_j|\geqslant 2\sqrt{3k(A)}$ ce qui implique l' existence d'une solution dans $B\big(s,\frac{1}{12}d(s)\big)$. Pour $x\in B\big(s,\frac{1}{12}d(s)\big)$ on aura

$$d(x)=\min_{i\neq j}|x_i-x_j|\geqslant\tfrac{5}{3}k(A)\geqslant\frac{3}{\sqrt{2}}k(A).$$

On peut donc énoncer le corollaire suivant.

Corollaire 2. *Si A est symétrique à diagonale nulle et si*

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

alors

(1) *il existe $\bar{x}$ solution de (P_s),*
(2) *$\bar{x}$ est un point attractif pour l'algorithme de Newton*, Alg. 2.

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 tql2 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

(i) $e_0=0\Rightarrow\forall k,\ e_k=0$,
(ii) $e_0\neq 0\Rightarrow\lim_{k\to\infty}e_k=0$.

Démonstration. Remarquons que si $y=J(x)z$ et que $J(x)$ est doublement stochastique, alors $\sum y_i=\sum z_i$. Il vient alors

$$[J(x^n)+\lambda I](x^{n+1}-x^n)=-\omega[\mu(A+x_i^n\delta_{ij})-\mu(S)],$$

$$\sum_i\left([J(x^n)+\lambda I](x^{n+1}-x^n)\right)_i=(1+\lambda)\sum_i(x^{n+1}-x^n)_i$$

$$=(1+\lambda)(e_{n+1}-e_n)=-\omega e_n,$$

d'où

$$(1+\lambda)e_{n+1}=(1+\lambda-\omega)e_n.$$

Il est alors immédiat de voir que pour $\omega\in]0,2[$ et $\lambda\geqslant 0$ on a $|1-\omega/(1+\lambda)|<1$. ∎

Notons que la méthode de Newton traditionnelle, qui correspond aux paramètres $\omega=1$ et $\lambda=0$, assure dès la seconde itération que la trace visée est atteinte, i.e., $e_1=0$ quelque soit e_0.

4. ALGORITHMES DE MINIMISATION

Dès que l'on sait calculer la dérivée de $x\to\mu(A+(x_i\delta_{ij}))$ pour approximer la solution de l'équation $\mu(A+x_i\delta_{ij})=\mu(S)$ on peut penser à minimiser

$$f(x)=\tfrac{1}{2}\|\,\mu(A+x_i\delta_{ij})-\mu(S)\|_2^2.$$

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

La fonction $f:x\to f(x)$ 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

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

où

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

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

$$A+(x_i\delta_{ij})=U\mathrm{Diag}[\,\mu(A+x_i\delta_{ij})\,]\,U^T$$

Notons également que $\lim_{x\to+\infty}f(x)=+\infty$. En effet d'après le théorème de Gerchgorin pour tout i, on a $\mu_i(A+x_i\delta_{ij})$ contenu dans la réunion pour $k=1,2,\ldots,n$ des boules de centre x_k et de rayon $\sum_{j\neq k}|a_{kj}|$. Ainsi f tend vers l'infini avec $\mu(A+x_i\delta_{ij})$ quand $x\to+\infty$.

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

$$\text{Alg. 3.} \qquad x^{n+1}=x^n-\rho_n\nabla f(x^n), \qquad n\geqslant 0, \quad \rho_n\geqslant 0.$$

Pour assurer la convergence de la suite $\{x^n\}$ on sait d'après J. Cea [18, p. 79] qu'il suffit que le coefficient ρ_n vérifie

$$f(x^{n+1})-f(x^n)>0,$$

$$\lim_{n\to\infty}\left[f(x^{n+1})-f(x^n)\right]=0 \quad\Rightarrow\quad \lim_{n\to\infty}\nabla f(x^n)=0.$$

On nomme un tel choix, un choix convergent de ρ.

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

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

$$\sum_1^n\mu_i(x)=\sum_1^n\mu_i(y).$$

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* $\mathrm{Sp}(A+X)=(s_1,s_2,\dots,s_n)$ *est que $x\leqslant s$.*

Notons alors w_s l'ensemble des $x\in\mathbf{R}^n$ tels que $x\leqslant s$. Horn [14] reprenant Hardy-Littlewood-Polya montre que w_s peut encore s'écrire $W_s=\{x\in\mathbf{R}^n|x=Ms,\ M$ matrice doublement stochastique$\}$ ce qui prouve 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.

On notera W_s l'ensemble des matrices diagonales $X=(x_i\delta_{ij})$ telles que $x\in w_s$.

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

$$\mathcal{O}(S)=\{B\in\mathfrak{M}_{nn}(\mathbf{R})|B=USU^T,\ U\text{ unitaire}\},$$

est un ensemble compact, mais non convexe. Il est clair que si x est une solution de (P_s) alors $A+x_i\delta_{ij}\in\mathcal{O}(S)$. Sur l'ensemble des matrices symétriques considérons le produit scalaire $(A,B)=\mathrm{tr}(AB)$. La norme induite est la norme de Schur. 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 $\frac{1}{2}$ près le carré de la distance de $A+X$ à $\mathcal{O}(S)$.

$A+W_s=\{A+x_i\delta_{ij}|x\in w_s\}$ est un ensemble convexe et compact 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] et 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.

La projection de $A^k=A+X^k=A+(x_i^k\delta_{ij})$ sur l'ensemble $\mathcal{O}(S)$ bien qu'il soit non convexe est facile à déterminer. On a

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

Soit M^k la matrice orthogonale définie par $A^k=M^k\mathrm{Diag}\mu(A+X^k)M^{kT}$. Alors la matrice $B^k=M^k\mathrm{Diag}\mu(S)M^{kT}$ vérifie

$$\|A^k-B^k\|_s^2=\|M^k\big[\mathrm{Diag}\mu(A+X^k)-\mathrm{Diag}\mu(S)\big]M^{kT}\|_s^2=\|\mu(A+X^k)-\mu(S)\|_2^2$$

car la norme de Schur est unitairement invariante. C'est-à-dire que B^k est une projection de A^k sur $\mathcal{O}(S)$. Il peut exister plusieurs projections de B^k sur $\mathcal{O}(S)$, mais notre façon de procéder en détermine une seule.

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 décomposition orthogonale de l'espace il suffit de projeter B^k sur W_s puis d'ajouter A. Or $\mathrm{Diag}B^k$ est la projection de B^k sur W_s car par définition de B^k on a

$$(\mathrm{Diag}B^k)_i=\sum_{j=1}^n m_{ij}^2\mu_j(S).$$

C'est-à-dire que $\mathrm{diag}B^k$ apparait comme l'image de $\mu(S)$ par la matrice doublement stochastique $(m_{ij}^2)_{ij}$ d'où le résultat.

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

(i) $B^k = M^k \operatorname{Diag}\mu(S) M^{kT}$ si $A^k = A + X^k \operatorname{Diag}\mu(A + X^k) M^{kT}$,
(ii) $A^{k+1} = A + X^{k+1} = A + \operatorname{Diag} B^k$. 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*

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

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

$$A + X^n = M^n \operatorname{Diag}\mu(A + X^n) M^{nT}$$

d'où

$$X^n = \operatorname{Diag}\{ M^n \operatorname{diag}\mu(A + X^n) M^{nT} \}$$

or

$$X^{n+1} = \operatorname{Diag}\{ M^n \operatorname{Diag}\mu(S) M^{nT} \}$$

donc

$$X^{n+1} = X^n + \operatorname{diag}\{ M^n [\operatorname{diag}\mu(S) - \operatorname{diag}\mu(A + X^n)] M^{nT} \}.$$

Cette formule itérative est celle obtenue par O. Hald [8]. Alg. 4 est l'interprétation de l'algorithme de O. Hald. Soit $e^T = (1, 1, \dots, 1)$; il vient

$$X^{n+1}e = x^{n+1} = x^n + \operatorname{diag}\{ M^n [\operatorname{diag}\mu(S) - \operatorname{diag}\mu(A + X^n)] M^{nT} \}e.$$

Or il est facile de vérifier que pour toutes matrices symétriques $R = V \operatorname{Diag}\mu(R) V^T$ on a

$$(\operatorname{diag} R)_i = \sum_{k=1}^{n} v_{ik}^{\cdot}\mu_k(R), \qquad i = 1, 2, \dots, n.$$

D'où avec les notations déjà introduites, bien qu'il n'y ait aucune considération de différentiabilité

$$\operatorname{Diag}(A + x_i\delta_{ij})e = J(x)^T \mu(A + x_i\delta_{ij}).$$

On a

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

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

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

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

Démonstration. On a

$$X^{n+1}=X^n+\mathrm{Diag}\left\{M^n\left[\mathrm{Diag}\,\mu(S)-\mathrm{Diag}\,\mu(A+X^n)\right]M^{nT}\right\},$$

d'où

$$B^{n+1}=M^n\,\mathrm{Diag}\,\mu(S)M^{nT}=M^n\,\mathrm{Diag}\,\mu(A+X^n)M^{nT}$$
$$+M^n\left[\mathrm{Diag}\,\mu(S)-\mathrm{Diag}\,\mu(A+X^n)\right]M^{nT},$$
$$B^{n+1}=A+X^n+(X^{n+1}-X^n)$$
$$+\mathrm{Hors\,diag}\left\{M^n\left[\mathrm{Diag}\,\mu(S)-\mathrm{Diag}(A+X^n)\right]M^{nT}\right\},$$
$$B^{n+1}=A+X^{n+1}+\mathrm{Hors\,diag}\left\{M^n\,(\mathrm{Diag}\,\mu(S)-\mathrm{Diag}\,\mu(A+X^n))M^{nT}\right\}.$$

En utilisant les propriétés de la norme de Schur on aura

$$\|M^n[\mathrm{Diag}\,\mu(S)-\mathrm{Diag}\,\mu(A+X^n)]M^{nT}\|_s^2=\|\mathrm{Diag}\,\mu(S)-\mathrm{Diag}\,\mu(A+X^n)\|_s^2$$
$$=\|X^{n+1}-X^n\|_s^2+\|\mathrm{Hors\,diag}\{M^n[\mathrm{Diag}\,\mu(S)-\mathrm{Diag}\,\mu(A+X^n)]M^{nT}\}\|_s^2,$$
$$\|\mathrm{Diag}\,\mu(S)-\mathrm{Diag}\,\mu(A+X^n)\|_s^2=\|X^{n+1}-X^n\|_s^2+\|B^{n+1}-(A+X^{n+1})\|^2,$$
$$\|\mu(S)-\mu(A+X^n)\|_2^2\geqslant\|X^{n+1}-X^n\|_s^2+\|\mu(S)-\mu(A+X^{n+1})\|_2^2.$$

Par sommation sur tous les $n \geqslant 0$ il vient

$$\lim_{n\to\infty} \| \mu(S) - \mu(A + X^n)\|_2^2 + \sum_{n>0} \|X^{n+1} - X^n\|_S^2 \leqslant \| \mu(S) - \mu(A + X^0)\|_2^2.$$

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* $A \in \mathfrak{M}_{nn}(\mathbf{R})$ *symétrique à diagonale nulle. Soit* $S = (s_i \delta_{ij})$ *fixée. Alors*

(1) *il y a coincidence entre les trois algorithmes de Hald, des projections successives*: *Alg.* 4, *de gradient à pas fixe* $\rho_n = 1$: *Alg.* 3, *i.e.*,

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

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

Démonstration. Il reste à prouver le (2). D'après le résultat du lemme précédent dû à O. Hald on a

$$\sum_{n>0} \|X^{n+1} - X^n\|_S^2 = \sum_{n>0} \|x^{n+1} - x^n\|_2^2 = \sum_{n>0} \|\nabla f(x^n)\|_2^2 < +\infty$$

d'où

$$\lim_{n\to\infty} \nabla f(x^n) = 0.$$

Le choix de $\rho_n = 1$ est donc un choix convergent, dans la terminologie de J. Cea [18] pour l'algorithme de descente: Alg. 3. D'où le résultat, d'après Cea [18, Théorème 5.1, p. 90]. ∎

On trouvera dans Cea [18] et Ortega-Rheinboldt [16] des méthodes pour déduire d'autres choix convergents de ρ à partir du choix convergent initial.

La formule itérative commune aux trois algorithmes s'écrit

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

En utilisant la relation entre le spectre et la diagonale d'une matrice vue plus haut, elle se réduit à

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

On peut donc considérer que nous recherchons par approximations successives les points fixes de l'application

$$g:x\rightarrow g(x)=J(x)^{T}\mu(S).$$

Faisons l'hypothèse que A est tridiagonale symétrique et que $a_{i-1,i}\neq 0$ pour $i=2,3,\ldots,n$. Alors $x\rightarrow\mu(A+x_i\delta_{ij})$ appartient à $C^{\infty}(\mathbf{R}^{n})$, donc $x\rightarrow g(x)$ est continue. Comme $J(x)$ est doublement stochastique g applique le convexe compact w_s dans lui-même. D'après Brouwer g possède toujours au moins un point fixe dans w_s. On vérifie d'une part que l'ensemble des points fixes de g coincide avec l'ensemble des points stationnaires de $f(x)=\frac{1}{2}\|\mu(A+x_i\delta_{ij})-\mu(S)\|_2^2$ et que d'autre part, pour qu'un point fixe z de g soit une solution, il suffit que $J(z)$ soit inversible.

Du point de vue numérique c'est la formule la plus simple, c'est-à-dire $x^{n+1}=J(x^{n})^{T}\mu(S)$ que l'on met en oeuvre. Remarquons que cela représente un gain notoire par rapport à la mise en oeuvre de l'algorithme de O. Hald. L'inconvénient majeur est que pour obtenir $J(x^{n})$ il faut calculer tous les vecteurs propres de $A+x_i^n\delta_{ij}$. L'avantage principal est que tous les itérés x^{n} appartiennent à w_s. C'est-à-dire qu'ils vérifient des conditions nécessaires assez strictes; en particulier la trace est respectée.

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

De la relation

$$A+X=M\operatorname{diag}\mu(S)M^{T}$$

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 (P_s) existe alors, nécessairement, $x\in S(0,r)\subset B(0,r)$. Posons

$$A+B(0,r)=\left\{M\in\mathfrak{M}_{nn}(\mathbf{R})\,\middle|\,M=A+(x_i\delta_{ij})\sum_{i=1}^{n}x_i^2\leqslant r^2\right\}.$$

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:

$$A=\frac{1}{h^2}\begin{bmatrix} C_a+(2+h^2q_1) & -1 & & & \\ -1 & 2+h^2q_2 & -1 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & -1 & 2+h^2q_{n-1} & -1 \\ & & & -1 & (2+h^2q_n)+C_b \end{bmatrix},$$

avec $C_a=-\beta_1/(\alpha_1h+\beta_1)$ et $C_b=-\beta_2/(\alpha_2h+\beta_2)$. La procédure est alors la suivante. On se donne une fonction q et les constantes α_1; α_2, β_1, β_2. Si $n=10$ alors $h=\frac{1}{11}$ et on obtient des matrices tridiagonales symétriques de taille 10×10. On calcule alors le spectre de cette matrice par l'algorithme du type Q.R. que Wilkinson-Reinsh nomme tql2 (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 tql2. Dans la méthode de Newton: Alg. 2 la résolution des systèmes linéaires est effectuée par un sous programme de Gauss avec pivot maximal.

Le temps d'une itération pour chacun des algorithmes est d'environ 7 millisecondes.

il vient

$$\Delta(k,k+1) \geqslant \|x^{k+1}-x^k\|_2^2 - \tfrac{1}{2}K\|x^{k+1}-x^k\|_2^2 = (1-\tfrac{1}{2}K)\|x^{k+1}-x^k\|_2^2.$$

Si $K<2$, comme $\lim_k \Delta(k,k+1)=0$ car $f(x^k)$ converge, cela implique

$$\lim_{k\to\infty} x^{k+1}-x^k=0.$$

$$\|x^{k+1}-x^k\| = \|\mathrm{proj}_B[x^k-\nabla f(x^k)]-x^k\|$$

or la suite des $\{x^k\}$ appartient au compact B, on peut donc en extraire une sous suite convergente disons encore x^k qui converge vers z d'où

$$\mathrm{proj}_B[x^k-\nabla f(x^k)]-x^k \text{ converge vers } \mathrm{proj}_B[z-\nabla f(z)]-z=0$$

d'après la continuité de $x\to\mathrm{proj}_B[x-\nabla f(x)]$.

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

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

alors si

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

où

$$f(x)=\tfrac{1}{2}\|\mu(A+x_i\delta_{ij})-\mu(S)\|_2^2,$$

l'algorithme $x^{n+1}=\mathrm{proj}_B[x^n-\nabla f(x^n)]$ *produit une suite dont tous les points adhérents sont des points stationnaires de f.*

5. ESSAIS NUMERIQUES

Nous avons fait de nombreuses expériences numériques, aussi nous n'en présentons ici qu'une partie. Nous ne présenterons pas de résultats numériques sur les expériences menées en faisant varier les constantes ω dans Alg. 1, et λ et ω dans Alg. 2; disons seulement que cela n'apporte pas de changements fondamentaux dans le comportement numérique de ces algorithmes. Dans les essais suivants nous aurons $\omega=1$ dans Alg. 1, $\omega=1$ et $\lambda=0.1/(\text{n}^\circ$ de l'itération) dans Alg. 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 $[A+B(0,r)]\cap \mathcal{O}(S)$; il est facile d'adapter l'algorithme des projections successives. Soit

$$A^k = A + X^k = M^k = M^k \operatorname{Diag}\mu(A+X^1)M^{kT}$$

la matrice $C^k = M^k \operatorname{Diag}\mu(S)M^{kT}$ réalise le minimum de la distance entre A^k et $\mathcal{O}(S)$. La projection de C^k sur le convexe compact $A+B(0,r)$ est facile à obtenir; c'est $A^{k+1} = A + X^{k+1} = A + r\operatorname{Diag}C^k/\|\operatorname{diag}C^k\|_2$. 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 définit 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 $x\in B=B(0,r)=\{x|\,\|x\|\leqslant r\}$ par une méthode de gradient projeté. En effet, pour le gradient projeté il vient:

$$x^{n+1} = \operatorname{proj}_B\{x^n - \rho_n \nabla f(x^n)\},$$

avec pour $\operatorname{proj}_B y = (r/\|y\|)\,y$. L'algorithme précédent est obtenu lorsque le pas ρ_n est constamment égal à 1.

En utilisant la formule de Taylor à l'ordre 2, il vient

$$\begin{aligned}\Delta(k,k+1) &= f(x^k) - f(x^{k+1})\\ &= -\nabla f(x^k)(x^{k+1}-x^k) - \tfrac{1}{2}f''(\alpha)(x^{k+1}-x^k)(x^{k+1}-x^k).\end{aligned}$$

Or x^{k+1} est la projection de $x^k - \nabla f(x^k)$ sur le convexe formé par l'ensemble des matrices diagonales de norme inférieure à r. On sait (cf. Golstein [22, p. 123]) que l'on peut caractériser cette projection par

$$\left(x^{k+1}-q|x^k - f(x^k) - q\right) \geqslant \|x^{k+1}-q\|_2^2$$

pour toute diagonale q de norme inférieure à r. En prenant $q=x^k$, il vient

$$-\nabla f(x^k)(x^{k+1}-x^k) \geqslant \|x^{k+1}-x^k\|_2^2.$$

Soit

$$K = \sup\|f''(\xi)\|;$$

il vient

$$\Delta(k,k+1) \geqslant \|x^{k+1}-x^k\|_2^2 - \tfrac{1}{2}K\|x^{k+1}-x^k\|_2^2 = \left(1-\tfrac{1}{2}K\right)\|x^{k+1}-x^k\|_2^2.$$

Si $K<2$, comme $\lim_k \Delta(k,k+1)=0$ car $f(x^k)$ converge, cela implique

$$\lim_{k\to\infty} x^{k+1}-x^k=0.$$

$$\|x^{k+1}-x^k\| = \|\mathrm{proj}_B\left[x^k-\nabla f(x^k)\right]-x^k\|$$

or la suite des $\{x^k\}$ appartient au compact B, on peut donc en extraire une sous suite convergente disons encore x^k qui converge vers z d'où

$$\mathrm{proj}_B\left[x^k-\nabla f(x^k)\right]-x^k \text{ converge vers } \mathrm{proj}_B\left[z-\nabla f(z)\right]-z=0$$

d'après la continuité de $x \to \mathrm{proj}_B[x-\nabla f(x)]$.

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

$$a_{i-1,i}\neq 0, \qquad i=2,3,\dots,n,$$

alors si

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

où

$$f(x)=\tfrac{1}{2}\|\mu(A+x_i\delta_{ij})-\mu(S)\|_2^2,$$

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

5. ESSAIS NUMERIQUES

Nous avons fait de nombreuses expériences numériques, aussi nous n'en présentons ici qu'une partie. Nous ne présenterons pas de résultats numériques sur les expériences menées en faisant varier les constantes ω dans Alg. 1, et λ et ω dans Alg. 2; disons seulement que cela n'apporte pas de changements fondamentaux dans le comportement numérique de ces algorithmes. Dans les essais suivants nous aurons $\omega=1$ dans Alg. 1, $\omega=1$ et $\lambda=0.1/(\mathrm{n}^\circ$ de l'itération) dans Alg. 2.

TABLEAU 3

$q(x)=x(1-x)$, $y(0)=y'(0)$, $y(1)=y'(1)$

No.	Alg. 1	Alg. 2	Alg. 3	Alg. 4
0	0.280	0.280	0.280	0.280
5	0.916	0.132	0.610×10^{-1}	0.550×10^{-1}
10	$0.112\times10^{+3}$	0.179	0.427×10^{-1}	0.408×10^{-1}
15	diverge	$0.145\times10^{+1}$	0.395×10^{-1}	0.371×10^{-1}
20		0.128	0.372×10^{-1}	0.336×10^{-1}
25		0.137×10^{-2}	0.339×10^{-1}	0.292×10^{-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

$q(x)=1-x$, $y(0)=y(1)=0$

No.	Alg. 1	Alg. 2	Alg. 3	Alg. 4
0	0.259	0.259	0.259	0.259
5	0.279	0.392×10^{-2}	0.423×10^{-1}	0.371×10^{-1}
10	$0.710\times10^{+1}$	0.282×10^{-4}	0.195×10^{-1}	0.174×10^{-1}
15	$0.113\times10^{+3}$	0.384×10^{-5}	0.134×10^{-1}	0.120×10^{-1}
20	diverge	0.250×10^{-5}	0.106×10^{-1}	0.941×10^{-2}
25		0.377×10^{-4}	0.891×10^{-2}	0.778×10^{-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

$q(x)=1-x$, $y'(0)=0$, $y(1)=0$

No.	Alg. 1	Alg. 2	Alg. 3	Alg. 4
0	0.268	0.268	0.268	0.268
5	0.376	0.139	0.466×10^{-1}	0.551×10^{-1}
10	0.976	0.577	0.235×10^{-1}	0.408×10^{-1}
15	$0.156\times10^{+3}$	0.402	0.185×10^{-1}	0.375×10^{-1}
20	diverge	0.161	0.161×10^{-1}	0.136×10^{-1}
25		$0.948\times10^{+1}$	0.140×10^{-1}	0.292×10^{-1}
50		0.196×10^{-1}	0.573×10^{-2}	0.863×10^{-2}
80		0.145×10^{-9} en No. 60	0.148×10^{-2}	0.351×10^{-2}

TABLEAU 6
$q(x)=1-x$, $y(0)=y'(0)$, $y(1)=y'(1)$

No.	Alg. 1	Alg. 2	Alg. 3	Alg. 4
0	0.279	0.279	0.279	0.279
5	0.519	0.210	0.609×10^{-1}	0.398×10^{-1}
10	$0.140\times10^{+2}$	0.314	0.426×10^{-1}	0.208×10^{-1}
15	$0.112\times10^{+3}$	0.647	0.394×10^{-1}	0.164×10^{-1}
20	diverge	0.219×10^{-1}	0.371×10^{-1}	0.140×10^{-1}
25		0.311×10^{-2}	0.338×10^{-1}	0.120×10^{-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.

La 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% d'erreur sur la plus petite composante et 0.25% 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.

REFERENCES

1 G. De Oliveira, Note on inverse characteristic problem, *Numer. Math.* **15** (1970), 339–341.
2 K. P. Hadeler, Ein inverses Eigenwertproblem, *Linear Algebra Appl.* **1** (1968), 83–101.
3 K. P. Hadeler, Newton-Verfahren für inverse Eigenwertaufgen, *Numer. Math.* **12** (1968), 35–39.
4 F. Laborde, Sur un problème inverse de valeurs propres, *C.R.A.S.* **268** (1969), 153–156.

5 F. Chatelin-Laborde, Thèse, Méthodes numériques de calcul de valeurs propres et vecteurs propres d'un opérateur linéaire, Grenoble, 1971.
6 P. Morel, A propos d'un problème inverse de valeurs propres, *C.R.A.S.* **277** (1973), 125–128.
7 P. Morel, Sur le problème inverse des valeurs propres, *Numer. Math.* **23** (1974), 83–94.
8 O. Hald, On discrete and numerical inverse Sturm-Liouville problems, Rep. 42, Uppsala Univ., Dept. of Computer Sci., 1972.
9 S. Friedland, Matrices with prescribed off diagonal elements, *Isr. J. Math.* **11** (1972), 184–189.
10 S. Friedland, Inverse eigenvalue problems, à paraître.
11 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford U. P., 1965.
12 T. Kato, *Perturbation Theory of Linear Operators, Springer, 1966.*
13 P. Lancaster, On eigenvalues of matrices dependent on a parameter, *Numer. Math.* **6** (1964), 377–387.
14 A. Horn, Doubly stochastic matrices and the diagonal of a rotation matrix, *Amer. J. Math.* **76** (1954), 620–630.
15 Hardy, Littlewood and Polya, *Inequalities*, Cambridge U. P., 1948.
16 J. M. Ortega W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic, 1970.
17 J. H. Wilkinson and C. Reinsch, *Linear Algebra. Handbook for Computations*, Springer (1971).
18 J. Cea, *Optimisation, Théorie et Algorithmes*, Dunod, 1971.
19 Gubin, Polyak and Raik, The method of projections for finding the common point of convex sets, USSR *Comput. Math. Math. Phys.* **6** (1967), 1–24.
20 G. Pierra, Sur le croissement de méthodes de descente, *C.R.A.S.* **277** (1973), 1071–1074.
21 H. W. Wielandt and A. J. Hoffman, The variation of the spectrum of a normal matrix, *Duke J. Math.* **20** (1953), 37–39.
22 A. Golstein, *Constructive Real Analysis*, Harper Int. Ed., 1967.

Received February, 1975