NUMERICAL METHODS OF HIGH-ORDER ACCURACY FOR NON-LINEAR TWO-POINT BOUNDARY VALUE PROBLEMS

P.G. CIARLET, M.H. SCHULTZ and R.S. VARGA

Case Institute of Technology, Cleveland (U.S.A.)

RESUME

Soit le problème : $Lu(x) = f(x,u)$, $0 < x < 1$, $u^{(k)}(0) = u^{(k)}(1) = 0$, $0 \leqslant k \leqslant n-1$, où $L = \sum_{j=0}^{n} D^j(p_j(x)D^j)$. On suppose de plus que $f_u(x,u) \geqslant -\gamma > -\lambda$, γ étant une constante, et où λ est la première valeur propre de L associé aux conditions aux limites ci-dessus. Ce problème équivaut à rendre minimum une certaine fonctionnelle $F(\omega)$ sur un espace S de fonctions assez "régulières". La minimisation de cette fonctionnelle sur des sous-espaces S_M de S convenablement choisis conduit à des "approximations" $\hat{\omega}_M(x)$, qui convergent vers l'unique solution $\varphi(x)$ du problème original, et ce, avec des hypothèses très faibles sur le comportement "asymptotique" des sous-espaces S_M. De plus, avec un choix particulier des sous-espaces S_M (que l'on associe alors à un maillage de pas h sur $[0,1]$) l'erreur est de la forme $0(h^{2m+1})$ dans la norme du Sup, et m peut être choisi aussi grand qu'on le veut.

SUMMARY

Let $Lu(x) = f(x,u)$, $0 < x < 1$; $u^{(k)}(0) = u^{(k)}(1) = 0$, $0 \leqslant k \leqslant n-1$, where $L = \sum_{j=0}^{n} D^j(p_j(x)D^j)$. Assume further that $f_u(x,u) \geqslant -\gamma > -\lambda$, for some constant γ, where λ is the first eigenvalue of L associated with the above boundary conditions. This problem is equivalent to minimizing a certain functional $F[\omega]$ over a space S of smooth functions. Minimizing $F[\omega]$ over appropriate finite-dimensional subspaces S_M of S leads to "approximations" $\hat{\omega}_M(x)$, which are shown to converge to the unique solution $\varphi(x)$ of the above problem, under mild asymptotic properties of the subspaces S_M. Moreover, with an appropriate choice of the subspaces S_M (associated with a mesh spacing h on $[0,1]$) the error is of order $0(h^{2m+1})$ in the sup-norm, where m can be chosen arbitrarily large.

1. INTRODUCTION

This paper is the extension to a nonlinear case of the results contained in Varga (1966), where the following type of problem was considered (p. 365) :

$$u'' = \sigma(x)u - f(x), \; 0 < x < 1 \; ; \; u(0) = u(1) = 0 \; , \tag{1}$$

under the major assumption :

$$\sigma(x) \geqslant 0 \; . \tag{2}$$

Otherwise, $\sigma(x)$ and $f(x)$ were assumed to be sufficiently smooth functions. Thus, we will link our results with those in the above reference, and we will adopt the same notations, whenever it is possible.

We now consider more general problems of the following form :

$$u'' = f(x,u), \; 0 < x < 1 \; ; \; u(0) = u(1) = 0 \; , \tag{3}$$

under the *major assumption* that there exists a constant γ such that :

$$\frac{\partial f}{\partial u}(x,u) \geqslant -\gamma > -\pi^2 \; ; \; 0 \leqslant x \leqslant 1 \; ; \; -\infty < u < +\infty \; . \tag{4}$$

For convenience, we will assume that $f(x,u)$ is of class C^1 on the strip $([0,1] \times R)$, although the following hypotheses are also sufficient :

$$f(x,u) \in C^0([0,1] \times R) \tag{4'}$$

$$\frac{f(x,u_1) - f(x,u_2)}{u_1 - u_2} \geqslant -\gamma > -\pi^2 \; ; 0 \leqslant x \leqslant 1 ; -\infty < u_1, u_2 < +\infty, \; u_1 \neq u_2 \; , \tag{4''}$$

and there exists a constant $M = M(c)$ depending only on c such that (4'') $u_1 \neq u_2$, $|u_1| \leqslant c$, $|u_2| \leqslant c$ implies :

$$\frac{f(x,u_1) - f(x,u_2)}{u_1 - u_2} \leqslant M(c) < +\infty \; , \tag{4'''}$$

for all $x \in [0,1]$.

It is known (Lees (1965)) that under the hypothesis (4) or the weaker hypotheses (4') – (4'') – (4'''), the problem (3) has a unique solution.

2. FORMULATION AS A VARIATIONAL PROBLEM

DEFINITION 1.– Let S be the class of piecewise C^1-functions defined over [0,1] and which moreover vanish at the end points.

THEOREM 1.– Let $\varphi(x)$ be the unique solution of (3). Then, $\varphi(x)$ minimizes strictly the following functional :

$$F[w] = \int_0^1 \left\{ \frac{1}{2} [w'(t)]^2 + \int_0^{w(t)} f(t,\eta)d\eta \right\} dt , \qquad (5)$$

over the space S.

Proof.– The proof is achieved by showing that :

$$F[w] - F[\varphi] \geqslant \frac{\pi^2 - \gamma}{2} \int_0^1 \{w(t) - \varphi(t)\}^2 dt ,$$

for any function $w(x) \in S$.

3. APPROXIMATION SCHEME

It is then quite natural to define an approximation scheme as follows :

Take any *finite* dimensional subspace S_M of S, of dimension M, that is, we are given M linearly independent functions $\{w_i(M,x)\}_{i=1}^M$ which are in S. Then we try to minimize the functional F[w] *over* S_M and the first result is the following, which is somehow the finite-dimensional equivalent of Theorem 1 :

THEOREM 2.– In S_M, there exists one and only one function which minimizes the functional (5) over S_M.

Proof.– Any function in S_M can be written as $w_M(x) = \sum_{j=1}^{M} u_j w_j(M,x)$. Hence, the functional (5), when expressed over S_M, becomes a functional *over* R^M, which we denote $F[\underline{u}] = F[u_1, u_2, \ldots, u_M]$.

First, it is clear that this functional $F[u]$ is bounded below, since $F[\underline{u}] \geqslant F[\varphi]$ by Theorem 1. In fact, letting :

$$\mathcal{M} = \sup_{[0,1]} |f(x,0)| ,$$

it can be proved directly that :

$$F[\underline{u}] \geqslant \frac{-\mathcal{M}^2}{2(\pi^2 - \gamma)}, \ \underline{u} \in R^M .$$

Next, it can be shown that, given any norm $\|\underline{u}\|$ over R^M :

$$\lim_{\|\underline{u}\| \to +\infty} F[\underline{u}] = +\infty .$$

Therefore, the two previous facts imply that $F[\underline{u}]$ attains its minimum over R^M for *at least one* vector.

Finally, $F[\underline{u}]$ attains its minimum for a *unique* vector, denoted $\hat{\underline{u}}$. This is achieved by proving that the matrix $B(\underline{u}) = (b_{ij}(\underline{u}))$, where :

$$b_{ij}(\underline{u}) = \frac{\partial^2 F}{\partial u_i \partial u_j} (\underline{u})$$

is uniformly positive definite, which in turn implies that $F[\underline{u}]$ represents a strictly convex surface, thus implying uniqueness of the minimizing vector.

Our approximating problem reduces now to writing the equations :

$$\frac{\partial F}{\partial u_i} = 0, 1 \leqslant i \leqslant M .$$

By Theorem 2, this system of (nonlinear) equations has a unique solution $\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_M\}$, to which is associated a *function* :

$$\hat{w}_M(x) = \sum_{j=1}^{M} \hat{u}_j w_j(M,x) .$$

An interesting feature is that the above system of nonlinear equations satisfies all the conditions required to apply the Gauss-Seidel-type or SOR-type methods described in Schechter (1962).

4. CONVERGENCE

We have seen that given any finite-dimensional subspace S_M of S, we can find in S_M an element $\hat{w}_M(x)$ which is the best approximation to the solution $\varphi(x)$, in the sense of minimizing the functional (5). It is then quite natural to expect that the difference $[\varphi(x) - \hat{w}_M(x)]$ might converge to zero (in some topology) if we have a *sequence* of subspaces $\{S_{M_i}\}_{i=1}^{\infty}$ (where $\lim_{i\to\infty} S_{M_i} = +\infty$) satisfying appropriate asymptotic properties.

It turns out that the topology most appropriate for our problem is induced by the Sobolev-type norm :

$$\|w\|_D = \{\textstyle\int_0^1 [w'(t)]^2 \, dt\}^{1/2} \quad . \tag{7}$$

Notice that this is indeed a norm on the space S since the boundary values are zero.

Our aim now is to derive an upper bound for $\|\hat{w}_M - \varphi\|_D$. To do this, we need first prove two basic lemmas :

LEMMA 1.– Let p(x) be a continuous function defined on [0,1] such that $p(x) \geqslant -\gamma > -\pi^2$, for all $x \in [0,1]$. Then, the following quantity :

$$\|w\|_{\{P\}} = \{\textstyle\int_0^1\{[w'(t)]^2 + p(t)\,[w(t)]^2\}\,dt\}^{1/2}$$

is a norm (in the space S) which is equivalent to the norm (7).

Proof.– The proof is very simple and left to the reader. Notice the exact similarity with the norm introduced in Varga (1966) in equation (3)p.366.

LEMMA 2.– The solution $\varphi(x)$ and the approximate solution $\hat{w}_M(x)$ both verify the same a priori bound :

$$\sup_{[0,1]} |\hat{w}_M(x)| \leqslant D(\gamma)\mathcal{M} \ , \tag{9}$$

$$\sup_{[0,1]} |\varphi(x)| \leqslant D(\gamma)\,\mathcal{M} \ , \tag{10}$$

where $\mathcal{M}$ was defined in (6) and $D(\gamma)$ is a constant depending only on γ.

Proof.– This is easily derived from the fact that :

$$F[\hat{w}_M],\ F[\varphi] \leqslant F[0] = 0 \ .$$

Now, we come to the key fact : in Varga (1966) it was shown that the element of best approximation $\hat{w}_M(x)$ in S_M could be considered as a *projection* of the solution $\varphi(x)$ on S_M, in the sense of an inner-product associated with a norm of type (8). Basically, this was possible because the problem was linear. However, we are still able now to view $\hat{w}_M(x)$ as a *projection* of $\varphi(x)$ over S_M in the sense of an inner-product of the form (8) *which will now vary with each* S_M, and this is achieved in the following fashion : The vector $\underline{\hat{u}} = \{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_M\}$ is the unique solution of the M equations :

$$\frac{\partial F}{\partial u_i} = 0, \; 1 \leqslant i \leqslant M \; .$$

Equivalently the function :

$$\hat{w}_M(x) = \sum_{j=1}^{M} \hat{u}_j w_j(M,x)$$

satisfies the M equations :

$$\int_0^1 \{\hat{w}'_M(t) w'_i(M,t) + f(t, \hat{w}_M(t)) \; w_i(M,t)\} \, dt = 0, \; 1 \leqslant i \leqslant M \; . \tag{11}$$

Similarly, it is readily seen that :

$$\int_0^1 \{\varphi'(t) w'_i(M,t) + f(t, \varphi(t)) \; w_i(M,t)\} \, dt = 0, \; 1 \leqslant i \leqslant M \; . \tag{12}$$

Thus, by simply subtracting (12) to (11), we obtain :

$$\int_0^1 [\hat{w}_M(t) - \varphi(t)]' w'_i(M,t) + p_M(t) \, [\hat{w}_M(t) - \varphi(t)] \, w_i(M,t) \; dt = 0, \; 1 \leqslant i \leqslant M, \tag{13}$$

where :

$$p_M(t) = \frac{\partial f}{\partial u}(t, \theta_t \varphi(t) + (1 - \theta_t) w_M(t)) , \qquad \text{with} \qquad 0 < \theta_t < 1 \; .$$

Thus by hypothesis (4), it follows that this is now an *orthogonality relation* in the sense of the norm $\|w\|_{\{p_M(x)\}}$, as defined in Lemma 1. As we noted earlier, the inner-product varies which each subspace S_M, but by Lemma 2, $p_M(x)$ satisfies bounds of the form :

$$-\pi^2 < -\gamma \leqslant p_M(x) \leqslant K(\gamma) , \quad 0 \leqslant x \leqslant 1 \; , \tag{14}$$

where $K(\gamma)$ is a constant *depending only on* γ.

Thus, we have by (13) :

$$\|\varphi - \hat{w}_M\|_{\{p_M\}} = \operatorname*{Inf}_{w \in S_M} \|\varphi - w\|_{\{p_M\}} , \tag{15}$$

(Notice the similarity with equation (8) p. 366 in Varga (1966)). Moreover, by Lemma 1 and inequality (14), there exists a constant $L(\gamma)$ depending only on γ such that :

$$\|\varphi - \hat{w}_M\|_{\{p_M\}} \leqslant L(\gamma) \operatorname*{Inf}_{w \in S_M} \|\varphi - w\|_D . \tag{16}$$

Finally, since there exists a constant $C(\gamma)$ depending only on γ such that :

$$\operatorname*{Sup}_{[0,1]} |w(x)| \leqslant C(\gamma) \|w\|_D, \text{ for all } w \in S, \tag{17}$$

we have proved

THEOREM 3.– The following error bound holds :

$$\operatorname*{Sup}_{[0,1]} |\varphi(x) - \hat{w}_M(x)| \leqslant M(\gamma) \operatorname*{Inf}_{w \in S_M} \|\varphi - w\|_D , \tag{18}$$

where the constant $M(\gamma)$ depends only on γ and $\|w\|_D$ is the norm defined in (7).

Remark.– The constant $M(\gamma)$ can be explicitly determined a priori.

As an immediate consequence, we have now.

THEOREM 4.– Given a (non-necessarily nested) sequence $\{S_{M_i}\}_{i=1}^{M}$, the best approximations $\hat{w}_{M_i}(x)$ converge *uniformly* to the solution if :

$$\lim_{i \to \infty} \{ \inf_{w \in S_{M_i}} \|w - g\|_D \} = 0$$

for all $g(x)$ in S.

5. HIGH-ORDER ACCURACY METHOD

As an example of subspaces satisfying the sufficient condition of Theorem 4, consider the spaces $S_M = H_N^{(m)}$ (Hermite Interpolation spaces), for $m \geqslant 0$, introduced in Varga (1966) (p. 365). By the fundamental inequality (18), we will obtain an error bound by plugging in the right-side of (18) *any trial* function. In Varga (1966), p. 368, inequality (20), an element $\tilde{w}(x)$ in $H_N^{(m)}$ is constructed such that :

$$\|\varphi - w\|_D \leqslant K\, M_{2m+2}\, h^{2m+1} , \tag{19}$$

where :

$$M_{2m+2} = \sup_{[0,1]} |\varphi^{(2m+2)}(x)| \tag{20}$$

and K is a constant depending only on m.

It is clear that given the problem (3), it is possible to see the order of differentiability of $\varphi(x)$; moreover, as was proved in Lees (1965), it is possible to derive a complete a priori bound for M_{2m+2}.

Thus, we have proved.

THEOREM 5.– When $S_M = H_N^{(m)}$, the following inequality hold :

$$\sup_{[0,1]} |\varphi(x) - \hat{w}_M(x)| \leqslant N(\gamma,m)\, h^{2m+1} , \tag{21}$$

where $N(\gamma,m)$ is a constant which can be *completely determined a priori.*

Remark.– Notice the exact similarity with Varga (1966), Theorem 1, p. 368.

Remark.– From Theorem 5, the error is at least $0(h^{2m+1})$ when the Hermite. Interpolation spaces $H_N^{(m)}$ are used. In fact, for $m = 0$ $\left(H_N^{(0)} \right.$ coincides with the space of piecewise linear continuous functions associated with a mesh spacing $\left. h = \frac{1}{N+1} \right)$, the error is $0(h^2)$ (instead of $0(h^1)$ as expected) whenever $\frac{\partial f}{\partial u}(x,u) \geqslant -\gamma > -8$, and at least $0(h^{3/2})$ whenever $-8 \geqslant -\gamma \geqslant -\pi^2$. A similar improvement might extend to higher values of m, although it has not been proved yet.

Remark. – *All* of the above results extend to more general problems such as :

$$\frac{d}{dx}\left\{p(x)\frac{du}{dx}\right\} = f(x,u),\ 0 < x < 1\ ;\quad \sigma_0 u(0) - u'(0) = 0,\quad \sigma_0 > 0\ ;$$
$$\sigma_1 u(1) + u'(1) = 0,\ \sigma_1 > 0,$$

where p(x) is only piecewise continuous and $p(x) \geqslant \omega > 0$, $0 \leqslant x \leqslant 1$, for some constant ω.

Then the functional (5) is to be replaced by :

$$F[w] = \int_0^1 \left\{ \frac{p(t)}{2}[w'(t)]^2 + \int_0^{w(t)} f(t,\eta)d\eta \right\} dt + \sigma_0 \frac{p(0)}{2}[w(0)]^2 + \sigma_1 \frac{p(1)}{2}[w(1)]^2.$$

Similarly, $\frac{\partial f}{\partial u}(x,u)$ has to be uniformly bounded away (as in (4)) from the first eigenvalue of the operator $\frac{d}{dx}\left\{p(x)\frac{d}{dx}\right\}$ associated with the boundary conditions $\sigma_0 u(0) - u'(0) = \sigma_1 u(1) + u'(1) = 0$. As in Section 5, we can get u.c.(h^{2m+1}) accuracy provided the mesh spacing is conveniently chosen (with respect to the discontinuities of p(x)) and the basis functions of $H_N^{(m)}$ are slightly modified.

BIBLIOGRAPHY

[1] LEES.M.– *"Two-Point Boundary-Value Problems"*, 1965, lecture notes.

[2] SCHECHTER S.– "Iteration methods for nonlinear problems", *Trans. Amer. Math. Soc.*, 1962, **104**.

[3] VARGA R.S.– "Hermite Interpolation-Type Ritz Methods for Two-Point Boundary Value Problems", *Numerical Solution of Partial Differential Equations*, 1966, Edited by J.H. Bramble, Academic Press, pp. 365-373.