NW

stichting
mathematisch
centrum

MC

NW

AFDELING NUMERIEKE WISKUNDE NW 24/75 SEPTEMBER

P.W. HEMKER

GALERKIN'S METHOD AND LOBATTO POINTS

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

AMS(MOS) subject classification scheme (1970): 65L10

Galerkin's method and Lobatto points

by

P. W. Hemker

ABSTRACT

An efficient implementation of Galerkin's method for the solution of a two-point boundary-value problem is described. Using the space $M^{0,k}$ of continuous piecewise polynomials of degree $\leq k$, an approximation is obtained that is pointwise accurate $O(h^{2k})$ on a quasi uniform grid. By selecting a particular set of basis functions in $M^{0,k}$, the resulting scheme has a striking resemblance with collocation, but in contrast with the corresponding collocation at Gaussian points, the piecewise polynomials have discontinuous derivatives.

A theorem by Douglas and Dupont concerning the relation between the degree of the quadrature rule and the pointwise error-bound is slightly generalized in order to deal with non-symmetric operators.

KEY WORDS & PHRASES: *Galerkin's method, collocation method, Lobatto quadrature*

1. GALERKIN'S METHOD AND LOBATTO POINTS

Consider the two-point boundary-value problem on $[a,b]$

(1) $$Ly \equiv -(py')' + qy' + ry = s;$$
$$p,q,r,s \in C^{t+1}[a,b],\ t \geq 2k-1;$$
$$0 < p_0 \leq p(x) \text{ on } [a,b];$$

(2) $$y(a) = \alpha,\ y(b) = \beta.$$

In this paper we construct a Galerkin method for the numerical approximation of the solution to this problem. Hence, the analytical solution $y(x)$ is approximated by a function $y_h(x)$ of the form

(3) $$y_h(x) = \sum_i a_i\phi_i(x);$$

(4) $$y_h(a) = \alpha;\ \ y_h(b) = \beta.$$

Here, $\{\phi_j\}_{j=0}^{M}$ is a set of continuous functions on $[a,b]$. The coefficients a_j are computed from the linear system

(5) $$\sum_j a_j \int_a^b p(x)\phi_j'(x)\phi_i'(x) + q(x)\phi_j'(x)\phi_i(x) + r(x)\phi_j(x)\phi_i(x)\,dx = \int_a^b s(x)\phi_i(x)\,dx, \qquad 1 \leq i \leq M-1,$$

and the constraints (4).

The $M-1$ functions $\{\phi_i\}_{i=1}^{M-1}$ are a subset from $\{\phi_j\}_{j=0}^{M}$ such that $\phi_i(a) = \phi_i(b) = 0$, $i = 1,\ldots,M-1$. In shorthand, we write instead of eq. (5):

(6) $$\sum_j a_j\, B(\phi_j,\phi_i) = (s,\phi_i), \qquad i = 1,\ldots,M-1.$$

It is well known that a set $\{\phi_j\}_{j=0}^{M}$ of piecewise polynomials has many computational advantages. In order to define $M^{0,k}$, the space of continuous k-th degree piecewise polynomials, we introduce a grid

$\{a = x_0 < x_1 < \dots < x_N = b\}$. The base-functions ϕ_j in $M^{0,k}$ are selected such that they are continuous on $[a,b]$ and identical to zero on $[a,b]$ except on at most two intervals $[x_{i-1},x_i]$. (This yields the band-matrix structure in the resulting discrete operator.) On each interval $[x_{i-1},x_i]$, $i = 1,2,\dots,N$, a function $v_h \in M^{0,k}$ is a polynomial of degree less or equal to k. What particular basis functions in $M^{0,k}$ are selected is given by eq. (12). We will motivate this choice by the following arguments.

It has been shown by DOUGLAS & DUPONT [1974] that, if the set $\{\phi_j\}_{j=0}^{M}$ allows for discontinuities in the derivatives of the elements of $M^{0,k}$ at the gridpoints $\{x_i\}$, the error of approximation at the gridpoints is of order $k + r$, $r \le k$, i.e.

$$(7) \qquad |y(x_i) - y_n(x_i)| = O(h^{k+r})$$

as long as $y \in H^{r+1}[a,b]$. Hence, at the gridpoints we permit discontinuities of $\phi_j'(x)$.

Setting up the discrete system of equations (5) requires the evaluation of a number of integrals. The integrals can be computed by the use of a fixed quadrature rule (of degree t) on each subinterval $[x_{i-1},x_i]$ of $[a,b]$. Hence, the linear system that is actually solved reads

$$(8) \qquad \sum_j a_j^\star B^\star(\phi_j,\phi_i) = (s,\phi_i)^\star,$$

where $B^\star(\phi_j,\phi_i)$ and $(s,\phi_i)^\star$ represent respectively $B(\phi_j,\phi_i)$ and (s,ϕ_i) modified by quadrature errors. The approximation actually obtained is

$$(9) \qquad y_h^\star = \sum_j a_j^\star \phi_j .$$

For the selfadjoint equation (i.e. problem (1)-(2) where $q(x) \equiv 0$), it has been indicated by DOUGLAS & DUPONT [1974], that there exists a unique solution to (8), provided that the grid $\{a = x_0 < x_1 < \dots < x_N = b\}$ is fine enough and $t \ge 2k - 2$. Moreover, they obtain the error-bound

$$(10) \qquad |y(x_i) - y_h^\star(x_i)| = O(h^{2k}) \qquad \text{if } y \in H^k[a,b] \text{ and if } t \ge 2k - 1.$$

Douglas and Dupont already noted that a k-point Gauss quadrature rule is sufficient in order to obtain the required accuracy in the errorbound (10). However, in order to obtain an efficient algorithm we advocate the use of a k+1-point quadrature rule.

Let $0 = \xi_0 < \xi_1 < \dots < \xi_k = 1$ be the family of base points of the k+1-point Lobatto quadrature rule on $[0,1]$ (see DAVIS & RABINOWITZ [1967]), and let $\{w_0, w_1, \dots, w_k\}$ be the corresponding set of weights. Using the Lobatto points $\{\xi_i\}$ we can now introduce our basis functions in $M^{0,k}$. Set

$$\xi_{i,\ell} = x_{i-1} + \xi_\ell (x_i - x_{i-1}), \tag{11}$$

then functions ϕ in $M^{0,k}$ are defined by their values at the Lobatto points $\{\xi_{i,\ell}\}$. We define our set of basis functions $\{\phi_j\}_{j=0}^{Nk}$ such that

$$\phi_{ik+\ell}(\xi_{m,n}) = \delta_{im}\,\delta_{\ell n} \tag{12}$$

$i = \ell = 0$ or $i = 0,1,\dots,N-1$; $\ell = 1,\dots,k$.

Note: It is convenient to identify $\phi_{i\ell} \equiv \phi_{ik+\ell}$; thus we can consider the set of $Nk + 1$ basis functions $\{\phi_{i,\ell}\}_{i=0,\dots,N;\ \ell = 0,\dots,k}$.

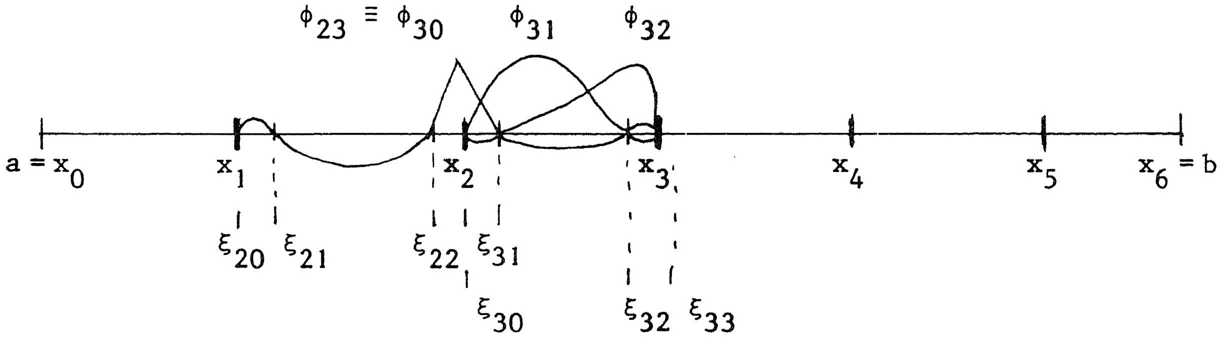

Figure 1. Basis functions in $M^{0,3}$

If this set of basis functions $\{\phi_{i,\ell}\}$ is used for the construction of an approximation (3) and if the k+1-point Lobatto quadrature rule is used on each interval $[x_{i-1}, x_i]$, then the elementary contributions to the entries of the discrete equation are

$$(13)\qquad \frac{1}{w_m}\int_{x_{i-1}}^{x_i} p(x)\phi'_{i\ell}(x)\phi'_{im}(x)dx \approx (x_i-x_{i-1})\sum_{n=0}^{k} w_n p(\xi_{in})\phi'_{i\ell}(\xi_{i,n})\phi'_{im}(\xi_{i,n})/w_m$$

$$= (x_i-x_{i-1})^{-1}\sum_{n=0}^{k} p(\xi_{in})\Phi'_\ell(\xi_n)\Phi'_m(\xi_n)w_n/w_m;$$

$$(14)\qquad \frac{1}{w_m}\int_{x_{i-1}}^{x_i} q(x)\phi'_{i\ell}(x)\phi_{im}(x)dx \approx q(\xi_{im})\Phi'_\ell(\xi_m);$$

$$(15)\qquad \frac{1}{w_m}\int_{x_{i-1}}^{x_i} r(x)\phi_{i\ell}(x)\phi_{im}(x)dx \approx \delta_{\ell m}(x_i-x_{i-1})r(\xi_{im});$$

$$(16)\qquad \frac{1}{w_m}\int_{x_{i-1}}^{x_i} s(x)\phi_{im}(x)dx \qquad \approx (x_i-x_{i-1})s(\xi_{im}).$$

Here, w_m and $\Phi'_m(\xi_n)$ $m,n = 0,1,\ldots,k$ are constants that are computed in advance. We see that the computation of the integrals involves a summation only in (13). At the other places a simple function evaluation suffices. Moreover, the integral (15) only contributes to a single entry in each row, viz. the entry on the main diagonal.

Even the summation in (13) can be circumvented. Since any problem of the form (1)-(2) can be rewritten in the same form with constant p, we can restrict the computational scheme to this case. Hence, also the sum

$$\sum_{n=0}^{k} \Phi'_\ell(\xi_n)\Phi'_m(\xi_n)w_n/w_m$$

is a constant number that can be computed in advance. We note that the transformation that makes p a constant number possibly will disturb the selfadjointness of the equation.

As will be shown in the theorem at the end of this paper, we obtain pointwise accuracy of order 2k by the use of the k+1-point Lobatto quadrature, and by the particular choice of $\{\phi_j\}_{j=0}^{Nk}$ we overcome the disadvantage of Galerkin's method the laborious evaluation of integrals.

Since the integrals in (13)-(16) have been divided by w_m, there is a striking resemblance with the collocation method, as far as the discretization of $q(x)\ y'(x)$, $r(x)\ y(x)$ and $s(x)$ are concerned. Hence we compare our method with collocation at Gaussian points (cf. DE BOOR & SCHWARTZ [1973]) which method attains accuracy of order $O(h^{2k})$ by collocation at only k points on each interval $[x_{i-1},x_i]$. The order of the resulting linear systems are the same for both collocation at Gaussian points and Galerkin at Lobatto points, since in the latter method each internal gridpoint x_i is a Lobatto-point on $[x_{i-1},x_i]$ as well as on $[x_i,x_{i-1}]$. The Galerkin scheme has the additional advantage that the discrete operator $B^*(\phi_j,\phi_i)$ is symmetric if the analytical operator is (i.e. if $q(x) = 0$). In contrast with the Galerkin method collocation requires an approximating function "y_h" that has a continuous derivative. This can be considered as an advantage if the solution y is a smooth function and if y' should be approximated, but it is a disadvantage if y varies rapidly.

Computational remark: The system (8) consists of N + 1 (k+1)×(k+1)-blocks on the main diagonal, with a single entry overlap between each two neighboring blocks. This can be used to reduce the system to tridiagonal form during its construction; each time when a (k+1)×(k+1)-block is computed the k-1 inner rows and columns of this block can be eliminated.

2. A SUPERCONVERGENCE THEOREM

In the following theorem we prove that, also for a non-symmetric, strongly coercive operator B, a (2k-1)-th degree quadrature rule is sufficiently accurate to obtain the pointwise errorbound in eq. (10).

THEOREM. *Let the operator* B *be strongly coercive, i.e. let* B *satisfy*

$$\exists\sigma > 0 \quad \forall v \in H_0^1[a,b] \quad \sigma\|v\|_1^2 \leq |B(v,v)|$$

and let the grid $\{a = x_0 < x_1 < \dots < x_N = b\}$ *satisfy the uniformity condition*

$$h = \max_{i=1,\dots,N} (x_i - x_{i-1}) \le \lambda \min_{i=1,\dots,N} (x_i - x_{i-1}).$$

If the solution of the problem (1)-(2) is approximated by y_h^* *(cf. eq. 9), which is a piecewise polynomial of degree* k*, and if* $B^*(\cdot,\cdot)$ *and* $(\cdot,\cdot)^*$ *are computed by a quadrature rule of degree* t*, then the pointwise errorbound*

$$|y(x_i) - y_h^*(x_i)| = O(h^{2k})$$

holds, if $t \ge 2k - 1$ *and if* h *is sufficiently small.*

PROOF. *) Let $G(x,\xi)$ be Green's function corresponding to the operator L and let V_h be the space of all continuous k-th degree piecewise polynomials on the grid $\{a = x_0 < x_1 < \dots < x_N = b\}$. Let G_i denote $G_i = G(x_i,\cdot)$, then for all $v \in V_h$

$$(17) \qquad \begin{aligned} |y_h(x_i) - y_h^*(x_i)| &\le |B(y_h - y_h^*, G_i)| \le \\ &\le |B(y_h - y_h^*, G_i - v)| + |B(y_h, v) - B^*(y_h^*, v)| + |B(y_h^*, v) - B^*(y_h^*, v)| \\ &\le K\|y_h - y_h^*\|_1 \, \|G_i - v\|_1 + |(s,v) - (s,v)^*| + |B(y_h^*, v) - B^*(y_h^*, v)| \end{aligned}$$

On the space of functions that have finite norms in $H^1[a,b]$ and $H^{t+1}[x_{i-1},x_i]$, $i = 1,2,\dots,N$, we introduce the norm $\|\cdot\|_{\pi,k}$ defined by

$$\|z\|_{\pi,k}^2 = \sum_{i=1,\dots,N} \|z\|_{H^k[x_{i-1},x_i]}^2 .$$

Note that $\|z\|_{\pi,k} = \|z\|_k$ if $z \in H^k[a,b]$ and that, by the Cauchy-Schwartz inequality

---

*) Throughout the proof C denotes a generic constant, that means that it is a constant of which the value may be different on each appearance.

$$\sum_i \|s\|_{H^m[x_{i-1},x_i]} \|v\|_{H^k[x_{i-1},x_i]} \leqq \|s\|_{\pi,m} \|v\|_{\pi,k}.$$

It is also easily verified that, if $k \geq 1$,

$$\|v\|_{\pi,k}\, h^{k-1} \leq C\, \|v\|_{\pi,1} = C\, \|v\|_1 \qquad \text{for all } v \in V_h.$$

By means of the newly defined norm we obtain the following errorbounds

$$(18) \qquad |(s,v)-(s,v)^*| \leqq \int_a^b |(sv)-\Pi(sv)|dx = \sum_i \int_{I_i} |(sv)-\Pi(sv)|dx$$

$$\leqq C \sum_i \int_{I_i} |D^{t+1}(sv)|dx \cdot h^{t+1}$$

$$\leq C \sum_i \int_{I_i} \sum_{j=0,\dots,t+1} |D^{t+1-j}s|\; |D^j v|dx.h^{t+1}$$

$$\leq C \sum_{i,j} \|D^{t+1-j}(s)\|_{L^2(I_i)} \|D^j u\|_{L^2(I_i)} h^{t+1}$$

$$\leq C \sum_i \|s\|_{H^{t+1}(I_i)} \|v\|_{H^k(I_i)} h^{t+1}$$

$$\leq C\, \|s\|_{\pi,t+1} \|v\|_{\pi,k}\, h^{t+1}.$$

Here Π denotes some interpolation operator from $H^{t+1}[a,b]$ into the set of piecewise polynomials of degree less or equal to t on $[a,b]$; Π is such that each polynomial of degree $\leq t$ remains unchanged. For each t-th degree quadrature rule a Π exists, such that

$$\int_0^1 f(x)dx \approx \sum_i w_i\, f(\xi_i) = \int_0^1 (\Pi f)(x)dx.$$

By theorem 5 from CIARLET & RAVIART [1972] we know that

$$\|u-\Pi u\|_{W^{p,m}[a,b]} \leq K(t)\, \|D^{t+1}u\|_{W^{p,0}[a,b]} h^{t+1-m}$$

if $u \in W^{t+1,p}[a,b]$, $1 \le p \le \infty$, $0 \le m \le t+1$.

Analogous to inequality (18) we obtain

$$(19)\qquad |B(y_h^*,v) - B^*(y_h^*,v)| \le C\{\|p\|_{W^{t+1,\infty}(I)} + \|q\|_{W^{t+1,\infty}(I)} + \|r\|_{W^{t+1,\infty}}\} \cdot \|y_h^*\|_{\pi,k} \cdot \|v\|_{\pi,k} \cdot h^{t+1}$$

In eq. (17), if h is small enough, v can be selected such that

$$\|G-v\|_1 < \|D^{k+1}G\|_{\pi,0}\, h^k \text{ and } \|v\|_{\pi,k} \le \|G_i-v\|_{\pi,k} + \|G_i\|_{\pi,k} \le 2\|G_i\|_k$$

In order to complete the proof of the lemma we now have to show that $\|y_h-y_h^*\|_1 \le C\, h^k$ and that $\|y_h^*\|_{\pi,k}$ is bounded by a constant independent of π, if h is small enough.

By the definitions of y_h^*, $(\cdot,\cdot)^*$ and $B^*(\cdot,\cdot)$ we have for all $v \in V_n$

$$\begin{aligned}(20)\qquad B(y_h-y_h^*,v) &= (s,v) - (s,v)^* + B^*(y_h^*,v) - B(y_h^*,v)\\ &\le |(s,v)-(s,v)^*| + |B(y_h^*,v)-B^*(y_h^*,v)|\\ &\le S\,\|v\|_{\pi,k}\, h^{t+1} + P\,\|y_h^*\|_{\pi,k}\,\|v\|_{\pi,k}\, h^{t+1}.\end{aligned}$$

Taking $v = y_h - y_h^*$, we have by the coercivity

$$\begin{aligned}(21)\qquad \sigma\|v\|_1^2 \le |B(v,v)| &\le S\,\|v\|_{\pi,k}\, h^{t+1} + P\,\|y_h^*\|_{\pi,k}\,\|v\|_{\pi,k}\, h^{t+1}\\ &\le C.S\,\|v\|_1\, h^{t+2-k} + C.P\,\|y_h^*\|_{\pi,k}\,\|v\|_1\, h^{t+2-k}\end{aligned}$$

Hence

$$\begin{aligned}(22)\qquad \sigma\,\|y_h-y_h^*\|_1 &\le \{C.S + C.P.\|y_h^*\|_{\pi,k}\}\, h^{t+2-k}\\ &\le \{C.S + C.P\|y_h\|_{\pi,k}\}\, h^{t+2-k} + C.P\,\|y_h-y_h^*\|_{\pi,k}\, h^{t+2-k}\\ &\le \{C.S + C.P\|y_h\|_{\pi,k}\}\, h^{t+2-k} + C.P\,\|y_h-y_h^*\|_1\, h^{t+3-2k}.\end{aligned}$$

If $t+3-2k > 0$ then

$$\sigma - C\,P\,h^{t+3-2k} > 0$$

if h is small enough and

$$(23)\qquad 0 < (\sigma - C.Ph^{t+3-2k})\,\|y_h - y_h^*\|_1 \le \{C.S + C.P\|y_h\|_{\pi,k}\}\,h^{t+2-k}$$

Since in the norm $\|\cdot\|_{\pi,k}$ the Galerkin solution y_h converges to the solution y

$$\|y_h\|_{\pi,k} \le \|y\|_{\pi,k} + \|y-y_h\|_{\pi,k} = \|y\|_k + \|y-y_h\|_{\pi,k} \le 2\,\|y\|_k$$

if k is small enough. Hence, in order to obtain convergence for h → 0, $t \ge 2k - 2$ is necessary. Moreover, if $t \ge 2k - 2$

$$(24)\qquad \|y_h - y_h^*\|_1 < C\,h^k$$

and

$$(25)\qquad \|y_h^*\|_{\pi,k} \le \|y_h\|_{\pi,k} + \|y_h - y_h^*\|_{\pi,k} < 2\,\|y_h\|_{\pi,k} < 4\,\|y\|_k$$

if h is small enough.

From the inequalities (17), (18), (19), (24) and (25) it now easily follows that

$$|y_h(x_i) - y_h^*(x_i)| \le C\,h^{2k}$$

provided that $t \ge 2k - 1$ and h is small enough. Since, by theorem 1 in DOUGLAS & DUPONT [1974]

$$|y_h(x_i) - y(x_i)| \le C\,k^{2k}$$

if k is small enough, the lemma is completed by combining both inequalities.

REFERENCES

CIARLET, P.G. & P.A. RAVIART [1972]. *General Lagrange and Hermite interpolation in* R^n *with applications to the finite element method.* Arch. Rat. Mech. Anal. 46, 177-199.

DE BOOR, C. & B. SCHWARTZ [1973]. *Collocation at Gaussian points.* SIAM J. Num. Anal. 10, 582-606.

DOUGLAS, J. Jr. & T. DUPONT [1974]. *Galerkin approximations for the two-point boundary-value problem using continuous piecewise polynomial spaces.* Numer. Math. 22, 99-109.

DAVIS, P.J. & P. RABINOWITZ [1967]. *Numerical integration.* Blaisdell.