stichting
mathematisch
centrum

MC

DEPARTMENT OF NUMERICAL MATHEMATICS NW 27/76 JUNE

M. BAKKER

NUMERICAL SOLUTION OF MILDLY NONLINEAR TWO-POINT BOUNDARY VALUE PROBLEMS BY MEANS OF GALERKIN'S METHOD

2e boerhaavestraat 49 amsterdam

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

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

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

Numerical solution of mildly nonlinear two-point boundary value problems by means of Galerkin's method

by

M. Bakker

ABSTRACT

This paper deals with the numerical solution of certain classes of even-order, self-adjoint, positive-definite, mildly nonlinear two-point boundary value problems, such as those analyzed by CIARLET, SCHULTZ & VARGA [1967]. The solution of the problems is approximated by piecewise polynomials of degree k which are m-1 times differentiable (2m being the order of the boundary value problem). If h is the mesh width of the trial space S_h, then it is proved that the numerical solution has a global error of order $h^{k+1-\ell}$, $\ell = 0,\dots,m$, and at the grid points the first m-1 derivatives have a local error of order $h^{2(k+1-m)}$. In two ways this is an extension of the results reported by DOUGLAS & DUPONT [1972,1974]:

(i) We prove that those results also hold for certain nonlinear problems.

(ii) For linear, and certain nonlinear, self-adjoint, positive-definite boundary value problems of order 2m, we prove that superconvergence generally holds for derivatives up to order m-1.

KEY WORDS & PHRASES: *Galerkin's method, mildly nonlinear boundary value problems, superconvergence.*

1. INTRODUCTION

In this paper we begin by studying a numerical method for solving the nonlinear boundary value problem:

$$(1.1)\qquad Ny \equiv -\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + f(x,y) = 0, \qquad x \in [a,b] = I,$$

$$(1.2)\qquad y(a) = y(b) = 0,$$

where p and f are supposed to be sufficiently differentiable and $p(x) \geq p_0 > 0$, $x \in I$.

The solution of (1.1)-(1.2) belongs to the space $H_0^1(I) \cap H^2(I)$, with

$$H^m(I) = \{v \mid D^j v \in L^2(I),\ \ j = 0,\dots,m\},$$

$$H_0^1(I) = \{v \mid v \in H^1(I),\ \ v(a) = v(b) = 0\},$$

where D^j stands for d^j/dx^j.

In the space $H^m(I)$ we define the Sobolev inner product and Sobolev norm by:

$$(u,v)_m = (u,v)_{H^m(I)} = \sum_{j=0}^{m} (D^j u, D^j v),$$

$$\|u\|_m = \|u\|_{H^m(I)} = \sqrt{(u,u)_m},$$

(.,.) being the inner product in $L^2(I)$.

Since the solution y of (1.1)-(1.2) also satisfies the *weak Galerkin form* (with $f(.,y)\colon I \to \mathbb{R}$ meaning $f(x,y(x))$):

$$(1.3)\qquad (py',w') + (f(.,y),w) = 0, \qquad w \in H_0^1(I),$$

it is reasonable to suppose that y can be approximated in a subspace of $H_0^1(I)$.

<u>DEFINITION</u> 1.1 Let $P_k(E)$ denote the set of polynomials of degree not greater than k restricted to the interval $E \subset I$. Let $\pi\colon a = x_0 < x_1 < \dots < x_M = b$ be

a partition of I with

$$
\begin{aligned}
h_j &= x_j - x_{j-1}, \\
I_j &= [x_{j-1}, x_j], \qquad j = 1,\dots,M, \qquad (1.4)\\
h &= \max h_j.
\end{aligned}
$$

Further we assume that π is quasi-uniform, i.e. $h_j \geq Ch$, where C is a constant independent of h and M. Then we define the space of kth degree piecewise polynomials by

$$
S_h = \{w \mid w \in H_0^1(I), \quad w \in P_k(I_j), \quad j = 1,\dots,M\}. \tag{1.5}
$$

In the following sections we will show how the solution of (1.1)-(1.2) can be approximated in S_h and under what conditions. Throughout this paper C, C_1, C_2, C', etc. will denote generic constants which will not be equal and θ, θ', θ_1, θ_2, etc. will be continuous functions of x on I, not necessarily equal and bounded between -1 and +1.

We conclude this introduction with a lemma which we shall need throughout this paper.

LEMMA 1.1 (Poincaré's inequality). *Let* $w \in H_0^1(I)$; *then*

$$
\begin{aligned}
\|w\|_\infty &\leq C \|Dw\|_0, \\
\|w\|_0 &\leq C \|Dw\|_0, \\
\|w\|_0 &\leq C \|w\|_1,
\end{aligned}
$$

where $\|.\|_\infty$ *denotes the supremum norm on* I.

PROOF

$$
\begin{aligned}
|w(x)| &= \left|\int_a^x 1 \cdot w'(t)dt\right| \\
&\leq \left\{\int_a^x dt\right\}^{\frac{1}{2}} \cdot \left\{\int_a^x [w'(t)]^2\, dt\right\}^{\frac{1}{2}} \quad \text{(Cauchy-Schwartz)} \\
&\leq \sqrt{(b-a)} \cdot \|Dw\|_0, \qquad x \in I,
\end{aligned}
$$

which proves the first inequality; the other inequalities can be proved from the first one. □

2. BACKGROUND MATERIAL

In this section we mention some properties of the operator N defined by (1.1) which play an important role in the uniqueness of the solution of (1.1)-(1.2). A great deal of this section is derived from CIARLET, SCHULTZ & VARGA [1967].

DEFINITION 2.1 The operator N defined by (1.1) is said to be *strictly monotone* if for any $y,z \in H_0^1(I) \cap H^2(I)$ the inequality

$$(2.1) \qquad (Ny-Nz, y-z) \geq C \, \|y-z\|_1^2$$

holds. (This definition is a particular case of strict monotonicity as defined in CIARLET, SCHULTZ & VARGA [1969].)

Next we define, for $p(x) \geq p_0 > 0$,

$$(2.2) \qquad \Lambda = \inf_{w \in H_0^1(I), w \neq 0} \frac{(pDw, Dw)}{\|w\|_0^2} .$$

One can easily recognize that Λ is the smallest eigenvalue of the operator $-D(pD.)$. acting on $H_0^1(I) \cap H^2(I)$. By expanding $w(x)$ into its Fourier series, we obtain

$$\Lambda \geq p_0 \inf_{w \in H_0^1(I)} \frac{\|w'\|^2}{\|w\|^2} = p_0 \left(\frac{\pi}{b-a}\right)^2 .$$

LEMMA 2.1 *Let $\gamma > -\Lambda$, where Λ is defined by (2.2); then for any $w \in H_0^1(I)$*

$$(2.3) \qquad \|w\|_\gamma = \{(pw', w') + \gamma(w,w)\}^{\frac{1}{2}}$$

is a norm equivalent to $\|w\|_1$.

PROOF We distinguish two cases: $\gamma < 0$ and $\gamma \geq 0$.

(i) $\gamma \geq 0$:

$$\begin{aligned}(pw',w') + \gamma(w,w) &= \tfrac{1}{2}(pw',w') + \gamma(w,w) + \tfrac{1}{2}(pw',w') \\ &\geq \tfrac{1}{2}p_0(w',w') + (\gamma+\tfrac{1}{2}\Lambda)(w,w) \\ &\geq \min(\tfrac{1}{2}p_0,\gamma+\tfrac{1}{2}\Lambda)\ \|w\|_1^2 = C_1\ \|w\|_1^2.\end{aligned}$$

On the other hand

$$(2.4)\qquad |(pw',w') + \gamma(w,w)| \leq C_2\ \|w\|_1^2,$$

$$C_2 = \max(\|p\|_\infty,\gamma),$$

which proves the lemma for $\gamma \geq 0$.

(ii) $\gamma < 0$: since $\gamma > -\Lambda$,

$$\gamma(w,w) > -(pw',w'), \qquad w \in H_0^1(I).$$

This implies that there exists an α, $0 < \alpha < 1$, such that

$$\gamma(w,w) \geq -\alpha(pw',w') > -(pw',w').$$

So

$$\begin{aligned}(pw',w') + \gamma(w,w) &= \tfrac{1}{2}(pw',w') + \tfrac{1}{2}\gamma(w,w) + \tfrac{1}{2}(pw',w') + \tfrac{1}{2}\gamma(w,w) \\ &\geq \tfrac{1}{2}(\Lambda+\gamma)\ \|w\|_0^2 + \tfrac{1}{2}(1-\alpha)(pw',w') \\ &\geq \tfrac{1}{2}(\Lambda+\gamma)\ \|w\|_0^2 + \tfrac{1}{2}(1-\alpha)p_0\ \|w'\|_0^2 \\ &\geq \tfrac{1}{2}\min(\Lambda+\gamma,(1-\alpha)p_0)\ \|w\|_1^2.\end{aligned}$$

The rest of the lemma is proved by application of (2.4). □

THEOREM 2.1 *Let* f(x,y) *be partially differentiable in* x *and* y *and satisfy*

$$\frac{f(x,y_1) - f(x,y_2)}{y_1 - y_2} \geq \gamma, \qquad y_1,y_2 \in H_0^1(I) \cap H^2(I), \tag{2.5}$$

with $\gamma > -\Lambda$, *where* Λ *is defined by* (2.2); *then the operator* N *defined by* (1.1) *acting on* $H_0^1(I) \cap H^2(I)$ *is strictly monotone.*

PROOF By means of partial integration, one obtains

$$\begin{aligned}(Ny_1-Ny_2,y_1-y_2) &= (p(y_1-y_2)',(y_1-y_2)') + (f(.,y_1)-f(.,y_2),y_1-y_2) \\ &\geq (p(y_1-y_2)',(y_1-y_2)') + \gamma(y_1-y_2,y_1-y_2),\end{aligned}$$

from which the theorem is proved by application of Lemma 2.1. □

We now obtain

THEOREM 2.2 *Let* (1.1) *admit a solution* y *and let* f *satisfy* (2.5); *then*

(i) y *is unique;*

(ii) y *strictly minimizes the functional*

$$I[w] = \int_a^b \{p(x)[w'(x)]^2 + 2\int_a^{w(x)} f(x,t)\, dt\}\, dx \tag{2.6}$$

over the space $H_0^1(I)$;

(iii) y *uniquely satisfies the weak Galerkin form*

$$(py',w') + (f(.,y),w) = 0, \qquad w \in H_0^1(I). \tag{2.7}$$

PROOF (See also CIARLET et al. [1967]).

(i) Suppose z is a second solution and $\varepsilon = y-z \neq 0$. Then after applying Theorem 2.1 we get

$$0 = (Ny-Nz,\varepsilon) \geq C\,\|\varepsilon\|_1^2 > 0$$

which is a contradiction.

(ii) Set $\varepsilon(x) = w(x)-y(x)$, $w \in H_0^1(I)$; then

$$I[w] - I[y] =$$

$$= \int_a^b \{p(x)[\varepsilon'(x)]^2 + 2p(x)\varepsilon'(x)y'(x) + 2\int_{y(x)}^{y(x)+\varepsilon(x)} f(x,t)dt\}\,dx$$

$$= \int_a^b \{p(x)[\varepsilon'(x)]^2 - 2f(x,y)\varepsilon(x) + 2\int_{y(x)}^{y(x)+\varepsilon(x)} f(x,t)dt\}\,dx$$

$$= \int_a^b \{p(x)[\varepsilon'(x)]^2 + 2\int_{y(x)}^{y(x)+\varepsilon(x)} [f(x,t)-f(x,y)]dt\}\,dx$$

$$\geq \int_a^b \{p(x)[\varepsilon'(x)]^2 + 2\int_{y(x)}^{y(x)+\varepsilon(x)} \gamma(t-y)dt\}\,dx$$

$$= \int_a^b \{p(x)[\varepsilon'(x)]^2 + \gamma[\varepsilon(x)]^2\}\,dx$$

$$\geq C\,\|\varepsilon\|_1^2 \geq 0.$$

So

$$I[y] \leq I[w], \qquad \forall w \in H_0^1(I).$$

If there is another $z \in H_0^1(I)$ which minimizes $I[w]$ then

$$0 = I[y] - I[z] \geq C\,\|y-z\|_1^2,$$

from which it follows that $y = z$.

(ii) This has already been proved by partial integration of (Ny,w). □

Now, since we have proved that y minimizes the functional $I[w]$ over $H_0^1(I)$ we may expect that y can be approximated by a function y_S which minimizes $I[w]$ over a finite dimensional subspace S of $H_0^1(I)$, just as is the case when N is a linear operator. We call this approximation method the Rayleigh-Ritz-Galerkin method.

THEOREM 2.3 *Let* S *be a finite dimensional subspace of* $H_0^1(I)$. *Then there is a unique* $y_S \in S$ *which strictly minimizes* $I[w]$ *over* S. *This* y_S *satisfies the weak Galerkin form*

$$(py_S', w_S') + (f(\cdot, y_S), w_S) = 0, \qquad w_S \in S.$$

PROOF See SCHULTZ [1973]. □

In the next section we apply Theorem 2.3 to the space S_h as defined in Section 1.

3. THE RAYLEIGH-RITZ-GALERKIN METHOD

In the previous section we have proved that the solution y of (1.1)-(1.2) can be approximated by a unique $y_h \in S_h$ which minimizes I[w] defined by (2.6) over S_h, provided that f(x,y) satisfies (2.5).

We first confine ourselves to the case that f(x,y) is linear in y, i.e.

$$f(x,y) = r(x)y - s(x).$$

In this case (1.1)-(1.2) becomes

$$(3.1) \qquad \begin{aligned} &-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + r(x)y = s(x), \qquad x \in I, \\ &y(a) = y(b) = 0. \end{aligned}$$

The weak Galerkin form (2.7) becomes

$$(3.2) \qquad (py', w') + (ry, w) = (s, w), \qquad w \in H_0^1(I),$$

and y strictly minimizes the functional

$$(3.3) \qquad J[w] = (pw', w') + (rw, w) - 2(s, w)$$

over $H_0^1(I)$.

THEOREM 3.1 *Let the space S_h of kth degree piecewise polynomials be defined by (1.5), and let $y \in H^{k+1}(I) \cap H_0^1(I)$ be the solution of (3.1); then there is a unique $y_h \in S_h$ which strictly minimizes the functional J[w], de-*

fined by (3.3), *over* S_h, *provided* $r(x) \geq \gamma > -\Lambda$. *This* y_h *is uniquely determined by the weak Galerkin form*

$$(3.4) \qquad (py_h', w_h') + (ry_h, w_h) = (s, w_h), \qquad w_h \in S_h,$$

and has the following error bounds:

$$(3.5) \qquad \|y-y_h\|_\ell \leq C\, h^{k+1-\ell} \|y\|_{k+1}, \qquad \ell = 0,1,$$

$$(3.6) \qquad |y(x_i)-y_h(x_i)| \leq C\, h^{2k} \|y\|_{k+1}, \qquad i = 0,\ldots,M,$$

where x_i *are the knots of the partition* π.

PROOF The existence of a y_h which minimizes $I[w]$ over S_h is proved by Theorem 2.3, which also states that y_h satisfies (3.4). The error bound (3.5) is proved in STRANG & FIX [1973]. The error bound (3.6) is proved by DOUGLAS & DUPONT [1974] for $r(x) \equiv 0$, but the proof can be extended to $r(x) \geq \gamma > -\Lambda$. □

We now return to our problem (1.1)-(1.2). We can rewrite it as follows:

$$(3.7) \qquad -(pu')' + \frac{\partial f}{\partial y} u = \frac{\partial f}{\partial y} y - f,$$

i.e., we put it in the form (3.2) with

$$(3.8) \qquad \begin{aligned} r(x) &= \frac{\partial f}{\partial y}(x,y), \\ s(x) &= y\frac{\partial f}{\partial y}(x,y) - f(x,y). \end{aligned}$$

The unique solution of (3.7) is of course $u(x) \equiv y(x)$.

We now derive the error bounds for the function $y_h \in S_h$ which minimizes $I[w]$ over S_h. To this end we study an auxiliary variational problem. This method has several analogies with a method used by RUSSELL [1974] to derive error bounds for the collocational solution of nonlinear boundary value problems.

LEMMA 3.1 *Let (1.1)-(1.2) have a solution* y *and let* $f(x,y)$ *satisfy (2.5). Let* $r(x)$ *and* $s(x)$ *be defined by (3.8). Let* S_h *be defined by (1.5). Then there is a unique* $u_h \in S_h$ *which strictly minimizes the functional* $J[w]$, *defined by (3.3), over* S_h. *This* u_h *is uniquely determined by the weak Galerkin form*

$$(3.9) \qquad (pu_h', w_h') + \left(\frac{\partial f}{\partial y}(\cdot,y)u_h, w_h\right) = \left(\frac{\partial f}{\partial y}(\cdot,y)y - f(\cdot,y), w_h\right), \qquad w_h \in S_h,$$

and has the following error bounds:

$$(3.10) \qquad \begin{aligned} \|y-u_h\|_\ell &\le C\, h^{k+1-\ell}\|y\|_{k+1}, & \ell &= 0,1, \\ |y(x_i)-u_h(x_i)| &\le C\, h^{2k}\|y\|_{k+1}, & i &= 0,1,\ldots,M. \end{aligned}$$

PROOF Direct application of Theorem 3.1 to problem (3.7). □

We now obtain

THEOREM 3.2 *Let (1.1)-(1.2) have a solution* y *and let* $f(x,y)$ *be twice partially differentiable in* x *and* y *and satisfy (2.5). Let* S_h *be defined by (1.5) and let* $y_h \in S_h$ *be the unique element which strictly minimizes the functional* $I[w]$ *defined by (2.6) over* S_h, *i.e. the solution of the weak Galerkin form*

$$(3.11) \qquad B_Q(y_h, w_h) \equiv (py_h', w_h') + (f(\cdot,y_h), w_h) = 0, \qquad w_h \in S_h.$$

Then y_h *has the following error bounds:*

$$(3.12) \qquad \begin{aligned} \|y_h-y\|_\ell &\le C\, h^{k+1-\ell}\|y\|_{k+1}, & \ell &= 0,1, \\ |y_h(x_i)-y(x_i)| &\le C\, h^{2k}\|y\|_{k+1}, & i &= 0,1,\ldots,M. \end{aligned}$$

PROOF We apply the quasibilinear operator B_Q defined by (3.11) to the solution u_h of (3.9). If we put $e_h = y-u_h$, then, for all $w_h \in S_h$, we get after application of (3.9)

$$(3.13) \qquad B_Q(u_h,w_h) = (pu_h',w_h') + (f(.,u_h),w_h)$$

$$= (e_h \frac{\partial f}{\partial y}(.,y) + f(.,u_h) - f(.,y),w_h)$$

$$= (e_h[\frac{\partial f}{\partial y}(.,y) - \frac{\partial f}{\partial y}(.,y+\theta e_h)],w_h)$$

$$= (-\theta e_h^2 \frac{\partial^2 f}{\partial y^2}(.,y+\theta' e_h),w_h).$$

But we also have

$$B_Q(u_h,w_h) = B_Q(u_h,w_h) - B_Q(y_h,w_h)$$

$$= (p(u_h'-y_h'),w_h') + (f(.,u_h)-f(.,y_h),w_h).$$

So, if we set $w_h = u_h - y_h = \delta_h$, we obtain after application of Lemma 2.1:

$$(3.14) \qquad B_Q(u_h,\delta_h) = (p\delta_h',\delta_h') + (\frac{\partial f}{\partial y}(.,y_h+\theta\delta_h)\delta_h,\delta_h)$$

$$\geq (p\delta_h',\delta_h') + \gamma(\delta_h,\delta_h) \geq C \|\delta_h\|_1^2.$$

If we combine (3.13) and (3.14), then we have:

$$(3.15) \qquad C \|\delta_h\|_1^2 \leq |B_Q(u_h,\delta_h)| \leq |(-\theta e_h^2 \frac{\partial^2 f}{\partial y^2}(.,y+\theta' e_h),\delta_h)|$$

$$\leq \|\frac{\partial^2 f}{\partial y^2}\|_\infty |(e_h^2,|\delta_h|)| \leq (b-a) M \|e_h\|_0^2 \|\delta_h\|_1 \text{ (Poincaré)}.$$

From (3.15) and (3.10) we get

$$(3.16a) \qquad \|\delta_h\|_1 \leq C \|e_h\|_0^2 \leq C h^{2k+2} \|y\|_{k+1}^2.$$

Poincaré's inequality gives

$$(3.16b) \qquad \begin{aligned} \|\delta_h\|_0 &\leq C h^{2k+2} \|y\|_{k+1}^2, \\ |\delta_h(x)| &\leq C h^{2k+2} \|y\|_{k+1}^2. \end{aligned}$$

From (3.16) we get

$$\|y-y_h\|_\ell \le \|y-u_h\|_\ell + \|u_h-y_h\|_\ell \le C_1 \|y\|_{k+1} h^{k+1-\ell} + C_2 \|y\|_{k+1}^2 h^{2k+2},$$

(3.17)

$$|y(x_i)-y_h(x_i)| \le |y(x_i)-u_h(x_i)| + |\delta_h(x_i)|$$
$$\le C_1 h^{2k} \|y\|_{k+1} + C_2 h^{2k+2} \|y\|_{k+1}^2,$$

which proves the theorem. □

4. QUADRATURE RULES

Let $y_h \in S_h$ be the solution of (3.11). If we represent y_h by

$$y_h(x) = \sum_{j=1}^{N} q_j \phi_j(x),$$

where $\{\phi_j\}_{j=1}^N$ is a basis of S_h, then the vector $(q_1,\ldots,q_N)^T$ is given by the nonlinear system

$$\sum_{j=1}^{N} (p\phi_i',\phi_j')q_j + (f(\cdot, \sum_{j=1}^{N} q_j\phi_j),\phi_i) = 0, \qquad i = 1,\ldots,N. \tag{4.1}$$

In order to solve (4.1) iteratively one has to evaluate the inner products $(p\phi_i',\phi_j')$ and $(f(\cdot, \sum_{j=1}^{N} q_j\phi_j),\phi_i)$. HERBOLD & VARGA [1970] suggest to evaluate $(p\phi_i',\phi_j')$ exactly and to use an interpolatory quadrature for $(f(\cdot, \sum_{j=1}^{N} q_j\phi_j),\phi_i)$ but they leave unsolved the problem how to evaluate $(p\phi_i',\phi_j')$. We, therefore, evaluate a method for approximating both inner products, which leaves the error bounds from §3 unchanged. This method was developed by DOUGLAS & DUPONT [1974].

4.1. Linear boundary value problems

We study the boundary value problem (3.1) where $r(x) \ge \gamma > -\Lambda$, Λ defined by (2.2). Let $\pi : a = x_0 < x_1 < \ldots < x_M = b$ be a partition of I with mesh width h. Let S_h be the space of k-th degree piecewise polynomials. We now introduce an approximation of (,) in the following way.

Let

$$(4.2) \qquad Q(f) = \sum_{\ell=1}^{s} w_\ell \, f(\eta_\ell), \qquad s \geq 1,$$

be an approximation of $\int_0^1 f(x)\,dx$ which is exact if f is a polynomial of degree less than 2k. Let $w_\ell > 0$, $\ell = 1,\dots,s$, and $0 \leq \eta_1 < \eta_2 < \dots < \eta_s \leq 1$. We define

$$\xi_{j,\ell} = x_{j-1} + h_j \eta_\ell, \qquad j = 1,\dots,M;\ \ell = 1,\dots,s;$$

$$(4.3) \qquad \langle\alpha,\beta\rangle_j = h_j \sum_{\ell=1}^{s} w_\ell \, \alpha(\xi_{j,\ell})\beta(\xi_{j,\ell}), \quad \alpha,\beta \in L^2(I_j), \quad j = 1,\dots,M;$$

$$\langle\alpha,\beta\rangle = \sum_{j=1}^{M} \langle\alpha,\beta\rangle_j .$$

Clearly, $\langle\alpha,\beta\rangle$ is equal to (α,β) if $\alpha*\beta \in P_{2k-1}(I_j)$, $j = 1,\dots,M$. If not, then the error $\langle\alpha\ \beta\rangle - (\alpha,\beta)$ is proportional to

$$\sum_{j=1}^{M} h_j^{2k}[D^{2k}(\alpha\beta)]_{x=\xi_j} \in \text{interior}(I_j) .$$

<u>THEOREM</u> 4.1 *Let* $\langle\cdot,\cdot\rangle$ *be defined by* (4.3) *and let* $p(x)$, $r(x)$ *and* $s(x)$ *sufficiently smooth,* $p(x) \geq p_0 > 0$, $r(x) \geq \gamma > -\Lambda$, Λ *defined by* (2.2); *then for sufficiently small* h *the modified weak Galerkin form*

$$(4.4) \qquad \langle pU_h', w_h'\rangle + \langle rU_h, w_h\rangle = \langle s, w_h\rangle, \qquad w_h \in S_h$$

has a unique solution U_h. *The following bounds exist for the difference between* U_h *and the solution of* (3.5)

$$(4.5) \qquad \begin{aligned} \|y-U_h\|_\ell &= O(h^{k+1-\ell}\|y\|_{2k}), \qquad \ell = 0,1; \\ |y(x_i)-U_h(x_i)| &= O(h^{2k}\ \|y\|_{2k}), \qquad i = 0,\dots,M. \end{aligned}$$

PROOF DOUGLAS & DUPONT [1974] gave a proof for $r(x) \equiv 0$, but it can also be proved for $r(x) \geq \gamma > -\Lambda$ after application of lemmas 2.1 and 4.1. □

4.2. The nonlinear boundary value problem

We will apply theorem 4.1 to obtain similar error bounds for the non-linear case. As in §3 we use the linearized boundary value problem (in Galerkin form)

$$(4.6) \qquad (pu',w') + (\frac{\partial f}{\partial y} u,w) = (y \frac{\partial f}{\partial y} - f(\cdot,y),w), \qquad w \in H_0^1(I),$$

as an auxiliary problem. In order to obtain the error bounds wanted we have to prove a few technical lemmas.

LEMMA 4.1 *Let* $\gamma > -\Lambda$, Λ *defined by* (2.2) *and let* $<\alpha,\beta>$ *be an approximation to* (α,β) *which is exact if* $\alpha\beta \in P_{2k-1}(I_j)$. *Then for sufficiently small* h

$$\|w_h\|_\gamma^* = \{<pw_h',w_h'> + \gamma<w_h,w_h>\}^{\frac{1}{2}}$$

is a norm on S_h *equivalent to* $\|w_h\|_1$.

PROOF We know by lemma 2.1 that $\|w_h\|_\gamma$ is a norm equivalent to $\|w_h\|_1$. We now compare $\|w_h\|_\gamma^2$ and $(\|w_h\|_\gamma^*)^2$ and, therefore, estimate the differences

$$(4.8) \qquad (pw_h',w_h') - <pw_h',w_h'> \text{ and } (w_h,w_h) - <w_h,w_h> ;$$

(a);
$$(pw_h',w_h') - <pw_h',w_h'> = \sum_{j=1}^{M} [(pw_h',w_h')_{I_j} - <pw_h,w_h'>_j] =$$

$$= \sum_{j=1}^{M} [((p-p_j)w_h',w_h')_{I_j} - <(p-p_j)w_h'.w_h'>_j] +$$

$$+ \sum_{j=1}^{M} p_j[(w_h',w_h')_{I_j} - <w_h',w_h'>_j] =$$

$$= \sum_{j=1}^{M} [((p-p_j)w_h',w_h')_{I_j} - <(p-p_j)w_h',w_h'>_j],$$

where p_j denotes the average value of $p(x)$ on I_j. Since on I_j

$$|p(x)-p_j| \le h_j |p'(\xi)| \le h_j \|p'\|_\infty ,$$

we get

$$|(pw_h',w_h') - \langle pw_h',w_h'\rangle| \le \sum_{j=1}^{M} h_j \|p'\|_\infty \{(w_h',w_h')_{I_j} + \langle w_h',w_h'\rangle_j\} =$$

$$= 2h\|p'\|_\infty \|w_h'\|_0^2 \le 2h\|p'\|_\infty \|w_h\|_1^2 .$$

(b); $$(w_h,w_h) - \langle w_h,w_h\rangle = O(h^{2k} \sum_{j=1}^{M} [D^{2k}(w_h^2)]_{x=\xi_j \in I_j}) =$$

$$= O(h^{2k} \sum_{j=1}^{M} [D^k w_h]^2_{x=\xi_j \in I_j}) = O(h^{2k-1} \sum_{j=1}^{M} \|D^k w_h\|^2_{I_j}) =$$

$$= O(h^{2k-1} \sum_{j=1}^{M} \|w_h\|^2_{H^k(I_j)} = O(h \sum_{j=1}^{M} \|w_h\|_{H^1(I_j)}) = O(h\|w_h\|_1^2).$$

So

$$|(pw_h',w_h') + \gamma(w_h,w_h) - \langle pw_h',w_h'\rangle - \gamma\langle w_h,w_h\rangle| \le$$

$$\le C_1\, h\|p'\|_\infty \|w_h\|_1^2 + C_2\, h\|w_h\|_1^2 \le C_3\, h\|w_h\|_1^2,$$

from which we can prove the lemma. □

<u>LEMMA</u> 4.2 *For any* $u,v \in H_0^1(I)$

(4.9) $$|\langle u,v\rangle| \le (b-a)\|u\|_\infty \|v\|_\infty \le (b-a)^2 \|u\|_1 \|v\|_1 .$$

<u>PROOF</u> $$|\langle u,v\rangle| = |\sum_{j=1}^{M} h_j \sum_{\ell=1}^{s} w_\ell\, u(\xi_{j,\ell})v(\xi_{j,\ell})| \le$$

$$\le \sum_{j=1}^{M} h_j \sum_{\ell=1}^{s} w_\ell |u(\xi_{j,\ell})||v(\xi_{j,\ell})| \le \sum_{j=1}^{M} h_j \sum_{\ell=1}^{s} w_\ell \|u\|_\infty \|v\|_\infty =$$

$$= \|u\|_\infty \|v\|_\infty \sum_{j=1}^{M} h_j \sum_{\ell=1}^{s} w_\ell = (b-a)\|u\|_\infty \|v\|_\infty \le (b-a)^2 \|u\|_1 \|v\|_1 . \quad \square$$

<u>LEMMA</u> 4.3 *Let* $p(x) \ge p_0 > 0$, $\partial f/\partial y \ge \gamma > -\Lambda$. *Then the weak modified Galerkin form*

(4.10) $$\langle pY_h',w_h'\rangle + \langle f_y Y_h,w_h\rangle = \langle yf_y - f(\cdot,y),w_h\rangle, \qquad w_h \in S_h,$$

has a unique solution y_h *with the error bounds*

(4.11)
$$\|y_h - y\|_\ell = O(h^{k+1-\ell}\|y\|_{2k}), \quad \ell = 0,1;$$
$$|y_h(x_i) - y(x_i)| = O(h^{2k}\|y\|_{2k}), \quad i = 0,1,\ldots,M.$$

PROOF Follows immediately from theorem 4.1 with $r(x) = \partial f/\partial y(x,y)$ and $S(x) = y\ \partial f/\partial y - f(x,y)$. □

THEOREM 4.2 *Let* $p(x) \ge p_0 > 0$, $f_y \ge \gamma > -\Lambda$. *Then, for sufficiently small* h, *the nonlinear modified Galerkin form*

(4.12)
$$<pz_h', w_h'> + <f(\cdot, z_h), w_h> = 0, \qquad w_h \in S_h,$$

has a unique solution $z_h \in S_h$ *which differs from the solution* y *of* (1.1)-(1.2) *by the following bounds*

(4.13)
$$\|y - z_h\|_\ell \le C\ h^{k+1-\ell}\|y\|_{2k}, \qquad \ell = 0,1;$$
$$|(y - z_h)(x_i)| \le C\ h^{2k}\|y\|_{2k}, \qquad i = 0,\ldots,M.$$

PROOF The same method is used by which theorem 3.2 was proved plus the technical lemmas from this section. Let y_h be the solution of (4.10) and put $\varepsilon_h = y - y_h$, $\delta_h = z_h - y_h$. Then analogue to (3.13) we get

$$|<py_h', w_h'> + <f(\cdot, y_h), w_h>| = |<-\theta\frac{\partial^2 f}{\partial y^2}(\cdot, y+\theta'\varepsilon_h)\varepsilon_h^2, w_h>| \le$$
$$\le \|\frac{\partial^2 f}{\partial y^2}\|_\infty\ |<\varepsilon_h^2, w_h>| \le C\ \|\varepsilon_h^2\|_\infty\ \|w_h\|_\infty \le C'\|\varepsilon_h\|_1^2\|w_h\|_1.$$

But also, if we put $w_h = \delta_h$, then after application of lemma 2.1, we obtain

$$<py_h', \delta_h'> + <f(\cdot, y_h), \delta_h> =$$

(4.15)
$$= <pw_h', \delta_h'> + <f(\cdot, y_h), \delta_h> - <pz_h', \delta_h> - <f(\cdot z_h), \delta_h> =$$

$$= \langle p\delta_h', \delta_h' \rangle + \langle f(\cdot, w_h) - f(\cdot z_h), \delta_h \rangle \geq$$

$$\geq \langle p\delta_h', \delta_h' \rangle + \gamma \langle \delta_h, \delta_h \rangle \geq C \| \delta_h \|_1^2 .$$

Combination of (4.14) and (4.15) gives after application of (4.11)

$$\text{(4.16a)} \qquad \begin{aligned} \| \delta_h \|_1^2 &\leq C \| \varepsilon_h \|_1^2 \| \delta_h \|_1 \; ; \\ \| \delta_h \|_1 &\leq C \| \varepsilon_h \|_1^2 \leq C \; h^{2k} \| y \|_{2k}^2 . \end{aligned}$$

From (4.16a) we get, applying Poincaré's inequality

$$\text{(4.16b)} \qquad \begin{aligned} \| \delta_h \|_0 &\leq C \; h^{2k} \| y \|_{2k}^2 ; \\ | \delta_h (x) | &\leq C \; h^{2k} \| y \|_{2k}^2 , \end{aligned}$$

whence we can prove (4.13).

The uniqueness can be proven from lemma 4.1. □

4.3. Lobatto quadrature

Now that we have proved that the use of a sufficiently accurate quadrature does not change the order of accuracy of Galerkin's method, let us give some examples of such quadratures. A well-known example is the k-point Gauss-Legendre quadrature which integrates polynomials of degree less than 2k exactly. We want, however, to spend special attention to another kind of quadrature with the same order of accuracy, namely k+1 - point *Lobatto* quadrature (see also HEMKER [1975] and ABRAMOWITZ [1964]). It is given by

$$Q_k(f) = \sum_{\ell=0}^{k} w_\ell \; f(\eta_\ell) \; ;$$

$$\eta_0 = 0, \quad \eta_k = 1, \quad \eta_\ell = \frac{1+t_\ell}{2}, \qquad \ell = 1, \ldots, k-1,$$

where t_ℓ are the zeros of $P_k'(t)$ on $(-1,+1)$, $P_k(t)$ being the k-th Legendre polynomial. The weights w_ℓ are uniquely determined by the requirement that

$$\sum_{\ell=0}^{k} w_\ell \, f(\eta_\ell) = \int_0^1 f(x)\, dx,$$

whenever f is a polynomial of degree less than 2k. We give Lobatto points and weights for k = 1,2,3.

<u>k = 1</u> (trapezoidal rule);

$$\eta_0 = 0, \quad \eta_1 = 1, \quad w_0 = w_1 = \tfrac{1}{2};$$

<u>k = 2</u> (Simpson's rule);

$$\eta_0 = 0, \quad \eta_1 = \tfrac{1}{2}, \quad \eta_2 = 1, \quad w_0 = w_2 = \frac{1}{6}, \quad w_2 = \frac{2}{3};$$

<u>k = 3</u>;

$$\eta_0 = 0, \quad \eta_1 = \frac{5-\sqrt{5}}{10}, \quad \eta_2 = \frac{5+\sqrt{5}}{10}, \quad \eta_3 = 1,$$

$$w_0 = w_3 = \frac{1}{12}, \quad w_1 = w_2 = \frac{5}{12}.$$

The great advantage of Lobatto quadrature is that we can let the points $\xi_{j,\ell}$ coincide with the *nodal* points of S_h (see also HEMKER [1975]): any member of S_h is entirely determined by the values at the points $\xi_{j,\ell}$. In the next section we will derive an efficient algorithm to solve (4.12) using Lobatto quadrature.

5. SOLUTION OF THE NONLINEAR SYSTEM

In this section we derive an algorithm to solve (4.12) using Lobatto quadrature.

5.1. <u>The nonlinear system</u>

Let π : $a=x_0 < x_1 < \ldots < x_M = b$ be a quasiuniform partition of [a,b].

We renumber the knots x_i as follows: $x_0, x_k, x_{2k} \ldots, x_{Mk}$. We now define the *interior mesh points* $x_{jk+\ell}$ as follows:

$$(5.1) \qquad x_{jk+\ell} = x_{jk} + h_{j+1}\,\eta_\ell \equiv \xi_{j+1,\ell}, \qquad j = 0,\ldots,M-1; \quad \ell = 1,\ldots,k-1.$$

We now define a basis $\{\phi_j(x)\}_{j=1}^{kM-1}$ of S_h by the requirement

$$(5.2) \qquad \phi_i(x_j) = \delta_{i,j}, \qquad 1 \le i,j \le kM-1.$$

Now, if we fill in ϕ_i in (4.12), we get

$$<pz_h', \phi_i'> + <f(\cdot, z_h), \phi_i> = 0, \qquad i = 1,\ldots,Mk-1.$$

But, since

$$<f(\cdot, z_h), \phi_i> = \sum_{j=1}^{M} h_j \sum_{\ell=0}^{k} w_\ell(\xi_{j,\ell}, z_\ell(\xi_{j,\ell}))\phi_i(\xi_{j,\ell}),$$

we get after application of (5.1) and (5.2)

$$<f(\cdot, z_h), \phi_i> = W_i\ f(x_i, z_h(x_i)),$$

where W_i is a constant weight determined by

$$W_{(j-1)k+\ell} = h_j\, w_\ell \quad , \qquad j = 1,\ldots,M; \quad \ell = 1,\ldots,k-1;$$

$$W_{jk} \quad = (h_j + h_{j-1})w_0, \qquad j = 1,\ldots,M-1.$$

Now, if we represent $z_h(x)$ by

$$z_h(x) = \sum_{j=1}^{kM-1} q_j\, \phi_j(x),$$

then $(q_1, q_2, \ldots, q_{kM-1})^T$ is determined by the nonlinear system

$$(5.3) \qquad A\vec{q} + \vec{F}(\vec{q}) = 0,$$

where

$$a_{ij} = \langle p\phi_i', \phi_j' \rangle, \qquad i,j = 1,\ldots,kM-1;$$

$$F_i = W_i\, f(x_i, q_i), \qquad i = 1,\ldots,kM-1.$$

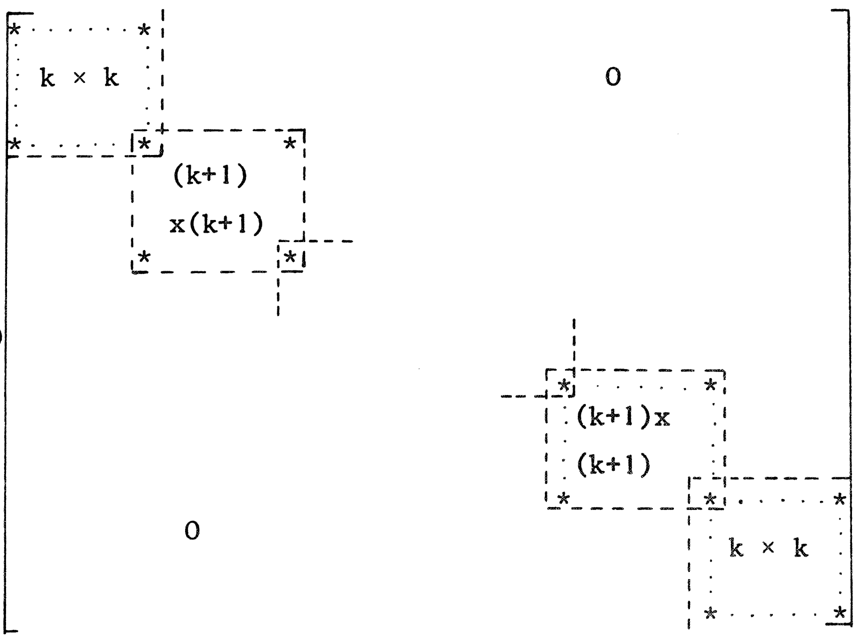
(5.3) is iteratively solved by means of the Newton-Raphson method (see e.g. ORTEGA & RHEINBOLD [1970]):

$$(5.4) \qquad \begin{aligned} J_n(\vec{q}^{(n)} - \vec{q}^{(n+1)}) &= A\vec{q}^{(n)} + \vec{F}(\vec{q}^{(n)}), \qquad n = 0,\ldots, \\ J_n &= (a_{ij} + W_i\, \delta_{ij} \frac{\partial f}{\partial y}(x_i, q_i^{(n)})). \end{aligned}$$

This method converges quadratically to the solution of (5.3) provided that $\|\vec{q}^{(0)} - \vec{q}\|$ is small enough.

Since $\phi_i(x)$ vanishes outside the segment to which x_i belongs, the matrix A has the following structure:

(5.5)



Since the Jacobian of (5.3) is only nonlinear in the main diagonal, updating of J can be done very easily. Iteration scheme (5.4) can be performed by subroutines using symmetric band matrices.

In the following sections we will discuss the questions how acurrately (5.3) should be solved and how to find initial guess for (5.4).

5.2. Solution strategy

Since the solution of (5.3) is itself an approximation of the solution of (1.1)-(1.2), which is of $O(h^{2k})$ if $i \equiv 0 (\mathrm{mod}\ k)$ and of $O(h^{k+1})$ otherwise it has no sense to solve (5.3) more accurately. At the other hand, since the approximation error itself is not known, it is hard to decide whether or not the iteration is to be stopped. We first prove the following

LEMMA 5.1 *Let* $Z_0 \in S_h$ *be an initial guess for the solution* z_h *of* (4.12); *then the sequence of functions* $\{Z_0, Z_1, \ldots, Z_n, \ldots\} \in S_h$ *defined by*

$$(5.6)\qquad \begin{aligned} &<pZ'_{n+1}, w'_h> + <\frac{\partial f}{\partial y}(.,Z_n)Z_{n+1}, v'_h> = \\ &= <\frac{\partial f}{\partial y}(.,Z_n)Z_n - f(.,Z_n), v_h> , \quad v_h \in S_h \end{aligned}$$

converges qaudratically to z_h, *provided* $\| z_h - Z_0 \|_0$ *is small enough.*

PROOF If one substitutes $\phi_i(x)$ in (5.6), one obtains scheme (5.4) which converges quadratically if $\| z_h - Z_0 \|_\theta$ is amall emough. □

We now outline the following strategy (see also Russell):

(i) Take an initial guess for Z_0;

(ii) Iteration scheme (5.6) is performed for *piecewise linear* functions, i.e. $k = 1$, until two subsequent iterates have a sufficiently small difference, say at $n = I$;

(iii) if $k = 1$, the process has been finished; if $k > 1$, then one can use Z_I as an initial guess for scheme (5.6) or (5.4) by interpolating at the interior knots; since $z_h - Z_I$ is of $O(h^2)$, iteration scheme (5.4) has to be performed once if $k = 2$ and twice if $k = 3$ or 4, in order to obtain the error bound (4.13).

5.4. Work estimate

In this section we briefly report how much work it costs to solve (5.3) by means of the Newton-Raphson method. We follow the strategy described in §5.2.

At the beginning the matrix $A = (p\phi_i', \phi_j')$ has to be evaluated for $k = 1$, which costs $M + 1$ evaluations of $p(x)$. Then iteration scheme (5.4) is performed I times. The cost of each iteration is

(i) updating of the Jacobian and the righthand side which costs $M + 1$ evaluations of $f(x,y)$ and $\frac{\partial f}{\partial y}(x,y)$.

(ii) solution of an $(M-1)$-dimensional linear system with a symmetric positive definite tridiagonal matrix.

After an initial guess has been obtained this way, the matrix $A = ((p\phi_i', \phi_j'))$ has to be reevaluated for $k > 1$. This costs $(k-1)M$ extra evaluations of $p(x)$. One now has to perform scheme (5.4) once if $k = 2$ and twice if $k = 3$ or 4. The cost of each iteration is

(i) $kM - 1$ evaluations of $f(x,y)$ and $kM - 1$ evaluations of $\frac{\partial f}{\partial y}(x,y)$;

(ii) the solution of a $(kM-1)$ dimensional linear system with a positive definite $(2k+1)$-diagonal matrix of the form (5.5).

If $k = 2$, the solution of the 5-diagonal system can be simplified by eliminating the components with odd index beforehand. This so called *static condensation* is made possible by the special structure of the Jacobian.

All together the amount of work needed for the Galerkin solution of (1.1) - (1.2) is

(i) $kM + 1$ evaluations of $p(x)$;

(ii) I times the solution of an $(M-1)$-dimensional tridiagonal linear system;

(iii) $I(M+1) + I_k(kM+1)$ evaluations of $f(x,y)$ and $\frac{\partial f}{\partial y}(x,y)$, with $I_1 = 0$, $I_2 = 1$, $I_3 = I_4 = 2$;

(iv) I_k times the solution of a $(kM-1)$ dimensional linear system with $(2k+1)$-diagonal matrix of the form (5.5).

6. GENERALIZATIONS

In this chapter a few generalizations of problem (1.1) - (1.2) are sketchily discussed.

6.1. The nonsymmetric case

In the previous sections the righthand side was expressed in x and y only. As a result the Jacobian matrix of the nonlinear system (5.6) was symmetric. We now study the problem

$$(6.1)\qquad \begin{aligned} &y''(x) = f(x,y(x),y'(x)), \qquad x \in I,\\ &y(a) = y(b) = 0. \end{aligned}$$

We suppose that f is sufficiently smooth in its three variables.

LEMMA 6.1 *The nonlinear operator* N *defined by* $Ny = -y'' + f(x,y,y')$ is strictly monotone, i.e. $(Ny-Nz,y-z) \geq c\|y-z\|_1^2$, $y,z \in H_0^1(I)$, if

$$(6.2)\qquad \frac{\partial f}{\partial y} - \tfrac{1}{2}\frac{d}{dx}\frac{\partial f}{\partial y'} \geq \gamma > -\Lambda, \qquad x \in I.$$

$$\Lambda = \inf_{w \in H_0^1(I)} \frac{\|Dw\|_0^2}{\|w\|_0^2} = +\left(\frac{\pi}{b-a}\right)^2.$$

PROOF We put $y - z = \delta$, y, $z \in H_0^1(I)$. Then, after partial integration, one obtains

$$\begin{aligned} &(Ny-Nz,\ y-z) =\\ &= \|\delta'\|_0^2 + (f(\cdot,y,y') - f(\cdot,z,z'),\delta) =\\ &= \|\delta'\|_0^2 + (\delta\frac{\partial f}{\partial y}(\cdot,y+\theta\delta,y'+\theta\delta'),\delta)\\ &\quad + (\delta'\frac{\partial f}{\partial y'}(\cdot,y+\theta\delta,y'+\theta\delta'),\delta) =\\ &= \|\delta'\|_0^2 + ([\frac{\partial f}{\partial y} - \tfrac{1}{2}\frac{d}{dx}\frac{\partial f}{\partial y'}](\cdot,y+\theta,\delta,y'+\phi,\delta')\delta,\delta)\\ &\geq \|\delta'\|_0^2 + \gamma\|\delta\|_0^2 \geq C\|\delta\|_1^2. \end{aligned}$$

The last inequality is proved by lemma 2.1. □

By the same techniques used in §§2-4 one can prove that, provided (6.2) holds, the respective Galerkin solutions of

(6.3) $$(y_h', w_h') + (f(\cdot, y_h, y_h'), w_h) = 0, \qquad w_h \in S_h;$$

and

(6.4) $$\langle z_h', w_h'\rangle + \langle f(\cdot, z_h, z_h'), w_h\rangle = 0, \qquad w_h \in S_h$$

are unique and have the error bounds (3.6) - (3.7) and (4.5) respectively. Furthermore, one can prove, using the techniques from §5 that the sequence $\{Z_0, Z_1, \dots\}$ generated by

(6.5)
$$\langle Z_{n+1}', w_h'\rangle + \langle \frac{\partial f}{\partial y}(\cdot, Z_n, Z_n')Z_{n+1} + \frac{\partial f}{\partial y'}(\cdot, Z_n, Z_n')Z_{n+1}', w_h\rangle =$$
$$= \frac{\partial f}{\partial y}(\cdot, Z_n, Z_n')Z_n + \frac{\partial f}{\partial y'}(\cdot, Z_n, Z_n')Z_n' - f(\cdot, Z_n, Z_n'), w_h\rangle, \qquad w_h \in S_h.$$

converges quadratically to the solution of (6.4).

6.2. Higher order problems

In this section we want to show that the results from §§3-5 can be extended to higher order self-adjoint boundary value problems with pure Dirichlet conditions. We therefore define the following 2m-th order self-adjoint boundary value problem

(6.7)
$$Ly \equiv \sum_{\ell=0}^{m} (-1)^{\ell+1} \frac{d^\ell}{dx^\ell}\left(p_\ell(x)\frac{d^\ell y}{dx^\ell}\right) = f(x,y), \qquad x \in [a,b] = I;$$
$$D^\ell y(a) = D^\ell y(b) = 0, \quad \ell = 0, \dots, m-1,$$

where $p_\ell(x)$ ($\ell=0,\dots,m$) are supposed to be sufficiently smooth and $p_m(x) \geq$ $\geq D^\ell > 0$.

The purpose of this section is to show that the properties of the Galerkin approximation of the solution to (1.1) - (1.2) can be generalized for m > 1. To this end we define

$$B(u,v) = \sum_{\ell=0}^{m} (p_\ell D^\ell u, D^\ell v), \qquad u,v \in H_0^m(I);$$

(6.8)
$$\Lambda = \inf_{u \in H_0^m(I)} \frac{B(u,u)}{\|u\|_0^2};$$

$$H_0^m(I) = \{u \mid u \in H^m(I)\; ; D^\ell u(a) = D^\ell u(b) = 0, \; \ell = 0,\dots, m-1\}$$

CIARLET et al. [1967] prove that the solution y of (6.7), if it exists, strictly minimines the functional

(6.9)
$$I[w] = B(w,w) + 2\int_a^b \Big[\int_a^{w(x)} f(x,t)dt\Big]dx$$

over $H_0^m(I)$ and satisfies the weak Galerkin from

(6.10)
$$B(y,w) + (f(\cdot,y),w) = 0, \qquad w \in H_0^m(I).$$

At a given partition $\pi : a = x_0 < x_1 < \dots < x_m = b$, we define the space $S_h^{k,m}$ as follows

(6.11)
$$S_h^{k,m} = \{w_h \mid w_h \in H_0^m(I);\; w_h \in P_k(I_j),\; j = 1,\dots,M\}.$$

One easily sees that $k \geq 2m - 1$. This space is a generalization of the space S_h in the §§2-5. In the sequel we denote this space by $S^{k,m}$. We now approximate y by minimizing $I[w]$ over $S^{k,m}$, which approximation is given by

(6.12)
$$B(y_h,w_h) + (f(\cdot,y_h),w_h) = 0, \qquad w_h \in S^{k,m}.$$

In order to get error bounds for y_h, we first confine ourselves to linear boundary value problems, i.e. to problems of the form

(6.13)
$$Lu = -s(x), \qquad x \in I\;;$$
$$D^\ell u(a) = D^\ell u(b) = 0, \qquad \ell = 0,\dots,m-1.$$

The Galerkin approximation $u_h \in S^{k,m}$ of u is given by the formula

$$(6.14) \qquad B(u_h,w_h) = (s,w_h)$$

<u>THEOREM</u> 6.1 *Let* $u \in H^{k+1}(I) \cap H_0^m(I)$ *be the solution of* (6.13) *and let* $u_h \in S^{k,m}$ *be the solution of* (6.14). *Let the symmetric bilinear operator* $B : H_0^m(I) \times H_0^m(I) \to \mathbb{R}$ *be bounded and strongly coercive, i.e.*

$$(6.15) \qquad \begin{aligned} |B(u,v)| &\le C_2\|u\|_m \|v\|_m; \ u,\ v \in H_0^m(I); \\ C_1\|w\|_m^2 \le B(w,w) &\le C_2\|w\|_m^2 \ ; \ w \in H_0^m(I); \end{aligned}$$

Then (6.14) has a unique solution and the error function $e_h(x) = (u-u_h)(x)$ has the following bounds

$$(6.16) \qquad \begin{aligned} \|e_h\|_\ell &\le Ch^{k+1-\ell} \|u\|_{k+1}; \quad \ell = 0,\ldots,m; \\ |D^\ell e_h(x_j)| &\le Ch^{2(k+1-m)} \|u\|_{k+1}; \quad \ell = 0,\ldots,m-1;\ j = 1,\ldots,M-1. \end{aligned}$$

<u>PROOF</u>: The uniqueness follows directly from the strong coercivity of B. The first error bound is proved in STRANG & FIX [1973]. In order to derive the second error bound, we introduce the Green's function of (6.13) (see also DOUGLAS & DUPONT [1974] and CODDINGTON & LEVINSON [1955]), i.e. the unique solution of

$$-L_\xi\, G(x,\xi) = 0, \qquad \xi \in I \setminus \{x\};$$

$$\frac{\partial^\ell}{\partial \xi^\ell} G(x,\xi) = 0, \qquad \xi = a,b;\ \ell = 0,\ldots,m-1.$$

$$(G(x,\cdot),Lu) = u(x),\ u \in H_0^m(I),\ x \in I.$$

This Green's function has the following properties

(i) $G(x,\cdot) \in H^{2m-1}(I) \cap H_0^m(I)$;

(ii) $G(x,\xi) = G(\xi,x)$, $(x,\xi) \in I \times I$;

(iii) $D_x^\ell G(x,\cdot) \in H^{k+1}[a,x] \cap H^{k+1}[x,b] \cap H_0^m(I)$, $\ell = 0,\dots,m-1$.

Now since $D_x^\ell G(x_j,\cdot) \in H^{k+1}[a,x_j] \cap H^{k+1}[x_j,b] \cap H_0^m(I)$, $j = 1,\dots, M$; $\ell = 0,\dots,m-1$, this function can be approximated by a $w_h \in S^{k,m}$ such that (see CIARLET & RAVIART [1972])

$$\|D_x^\ell G(x_j,\cdot) - w_h\|_m \le Ch^{k+1-m}\|D_x^\ell G(x_j,\cdot)\|_{\pi,k+1}$$

(6.17)

$$\|\cdot\|^2_{\pi,\ell} = \sum_{j=1}^m \|\cdot\|^2_{H^\ell(I_j)}, \quad \ell = 0,1,\dots .$$

Since for any $u \in H_0^m(I)$ one can write

$$B(u, G(x,\cdot)) = u(x),$$

$D^\ell e_h(x_j)$ is represented by

$$D^\ell e_h(x_j) = B(e_h, D_x^\ell G(x_j,\cdot)) =$$
$$= B(e_h,w_h) + B(e_h,D_x^\ell G(x_j,\cdot) - w_h), \quad w_h \in S^{k,m}.$$

It follows from (6.15) and (6.17) that

$$|D^\ell e_h(x_j)| \in C\|e_h\|_m \inf_{w_h\in S^{k,m}} \|D_x^\ell G(x_j,\cdot) - w_h\|_m \le C\, h^{k+1-m}\|u\|_{k+1} \,*$$

$$*\, C\, h^{k+1-m}\|D_x^\ell G(x_j,\cdot)\|_{\pi,k+1}$$

Since $\|D_x^\ell G(x_j,\cdot)\|_{\pi,k+1}$ is bounded, the second error bound has also been proved □.

Now that superconvergence at the knots has been established for higher order problems, it can be shown that the results of §§3-5 can be generalized in the following ways.

THEOREM 6.2 (Generalization of theorem 4.1). *Let* $\langle\alpha,\beta\rangle$, *defined by* (4.2) *and* (4.3) *be an approximation of* (α,β) *which is exact if* $\alpha\beta \in P_{2k-2m+1}(I_j)$, $j = 1,\ldots,M$. *Then, if* h *is small enough, the weak Galerkin form*

$$(6.18)\qquad \begin{aligned} B^*(z_h,w_h) &= \sum_{\ell=0}^{m} \langle p_\ell D^\ell z_h, D^\ell w_h\rangle = \\ &= \langle s,w_h\rangle,\ w_h \in S^{k,m}, \end{aligned}$$

has a unique solution z_h and for the error function $e_h = u - z_h$ (u is the solution of (6.13)) the error bounds (6.16) hold.

THEOREM 6.3 *Let* $\frac{\partial f}{\partial y}(x,y) \geq \gamma > -\Lambda$, Λ *defined by* (6.8). *Then, both* (6.10) *and* (6.12) *have a unique soltion* y *and* y_h, *respectively. For the error function* $e_h = y - y_h$ *the bounds* (6.16) *hold.*

THEOREM 6.4 *Let* $B^*(z_h,w_h)$ *be defined by* (6.18) *and let* $\frac{\partial f}{\partial y}(x,y) \geq \gamma > -\Lambda$. *Then, if* h *is small enough, the Galerkin form*

$$(6.19)\qquad B^*(z_h,w_h) + \langle f(\cdot,z_h),w_h\rangle = 0,\ w_h \in S^{k,m}$$

has a unique solution with error bounds (6.16).

THEOREM 6.5 *Let* $Z_0 \in S^{k,m}$ *be an initial guess of the solution of* (6.19) *and let the conditions of theorem* 6.5 *hold, then the sequence of functions* $\{Z_0,Z_1,\ldots\}$ *generated by*

$$(6.20)\qquad \begin{aligned} B^*(Z_{n+1},w_h) &+ \langle\frac{\partial f}{\partial y}(\cdot,Z_n)Z_{n+1},w_h\rangle = \\ &= \langle\frac{\partial f}{\partial y}(\cdot,Z_n)Z_n - f(\cdot,Z_n),w_h\rangle,\ w_h \in S^{k,m} \end{aligned}$$

converges quadratically to the solution z_h *of* (6.19), *provided that* $\|Z_0 - Z_h\|$ *is small enough.*

Since accurate proofs of these theorems would mainly consist of copying §§3-5, we just outline them.

PROOF of theorem 6.2 Analogue to DOUGLAS & DUPONT [1974] we prove for $\delta_h = u_h - z_h$ (the solution of (6.14) and (6.18), respectively)

$$B(\delta_h,\delta_h) = |B(u_h,\delta_h) - B(z_h,\delta_h)| \le$$
$$\le |(s,\delta_h) - \langle s,\delta_h\rangle| + |B^*(z_h,\delta_h) - B(z_h,\delta_h)|,$$

from which it can be proved that $\|\delta_h\|_m = O(h^{k+1-m})$ and hence it can be proved that $\|\delta_h\|_\ell = O(h^{k+1-\ell})$, for $\ell = 0,\ldots,m-1$. This is done by proving that $\|\delta_h\|_\ell \le Ch\|\delta_h\|_{\ell+1}$, $\ell = 0,\ldots,m-1$.

The errors $D^\ell\delta_h(x_j)$ at the knots x_j are given by

$$\begin{aligned}
|D^\ell\delta_h(x_j)| &= |B(\delta_h,D_x^\ell G(x_j,\cdot))| = \\
&= |B(\delta_h,D_x^\ell G(x_j,\cdot)-w_h) + B(\delta_h,w_h)| \le \\
&\le |B(\delta_h,D_x^\ell G(x_j,\cdot)-w_h)| \\
&+ |(s,w_h) - \langle s,w_h\rangle| \\
&+ |B^*(z_h,w_h) - B(z_h,w_h)|.
\end{aligned}$$

By taking w_h such that

$$\|D_x^\ell G(x_j,\cdot)-w_h\|_m \le C\|D_x^\ell G(x_j,\cdot)\|_{\pi,k+1}\, h^{k+1-m},$$

one obtains the error bounds (6.16), since one can prove that all three term are of order $h^{2(k+1-m)}$ (see also DOUGLAS & DUPONT [1974]). □

PROOF of theorem 6.3 One can prove that for any $u,v \in H_0^m(I)$ the inequality

$$B(u-v,u-v) + (f(\cdot,u) - f(\cdot,v),u-v) \ge C\|u-v\|_m^2$$

holds, which proves the uniqueness of the solutions of (6.10) and (6.12). The error bounds (6.16) are obtained by comparing y_h with the Galerkin solution $u_h \in S^{k,m}$ of

$$B(u_h,w_h) + (\frac{\partial f}{\partial y}(\cdot,y)u_h,w_h) =$$

$$= (\frac{\partial f}{\partial y}(\cdot,y)y - f(\cdot,y),w_h), \; w_h \in S^{k,m},$$

which has the same error bounds (6.16) □.

PROOF of theorem 6.4 One can prove that if h is small enough for any $u_h,v_h \in S^{k,m}$ the inequality (we set $\delta_h = u_h - v_h$)

$$B^*(\delta_h,\delta_h) + \langle f(\cdot,u_h) - f(\cdot,v_h),\delta_h\rangle \geq C\|\delta_h\|_m^2,$$

holds, which proves the uniqueness. The error bounds (6.16) are obtained by comparing z_h with the solution u_h of the form

$$B^*(u_h,w_h) + \langle\frac{\partial f}{\partial y}(\cdot,y)u_h,w_h\rangle =$$

$$= \langle\frac{\partial f}{\partial y}(\cdot,y)y - f(\cdot,y),w_h\rangle, \; w_h \in S^{k,m},$$

which has the error bounds (6.16) □.

PROOF of theorem 6.5 Let $\{\phi_i\}_{i=1}^N$ be a basis of $S^{k,m}$. If we set

$$z_h(x) = \sum_{i=1}^N q_i\phi_i(x),$$

and apply (6.19) for ϕ_i, $i = 1,\ldots,N$, we obtain the nonlinear system

$$A\vec{q} + \vec{F}(\vec{q}) = 0 \quad ,$$

$$A = (B^*(\phi_i,\phi_j)) \quad ;$$

$$\vec{F} = ((f(\cdot, \sum_{j=1}^N q_j\phi_j),\phi_i)).$$

This nonlinear system can be iteratively solved by the Newton-Raphson algorithm (see ORTEGA et al.)

$$(6.21)\qquad J_n(\vec{q}^{(n)}-\vec{q}^{(n+1)}) = A\vec{q}^{(n)} + \vec{F}(\vec{q}^{(n)}), \qquad n = 0,\ldots,$$

which converges quadratically to $\vec{q}$ provided the Euclidean norm $\|\vec{q}^{(0)}-\vec{q}\|$ is small enough. If one applies (6.20) for ϕ_i, $i = 1,\ldots,N$, one obtains (6.21). Since $\|Z_0-z_h\|_0$ is equivalent to the Euclidean vector norm $\|\vec{q}^{(0)}-\vec{q}\|$, iteration scheme (6.20) converges quadratically to z_h if $\|Z_0-z_h\|$ is small enough □.

7. NUMERICAL EXAMPLES

In this paragraph we give three examples of nonlinear two-point boundary value problems with Dirichlet boundary conditions. They were solved on a CDC CYBER 73/28 computer.

Example 1.

$$(7.1)\qquad \begin{aligned} &y'' = e^y, \quad x \in [0,1].\\ &y(0) = y(1) = 0. \end{aligned}$$

This classical example (see e.g. CIARLET et al. [1967], DE BOOR & SWARTZ [1973] and WEISS [1974]) has the analytic solution

$$(7.2)\qquad \begin{aligned} &y(x) = 2\ \ln \frac{c}{\cos \frac{c}{2}(x-\frac{1}{2})} + \ln 2,\\ &c = 1.336055695. \end{aligned}$$

In order to test the superconvergence at the knots we work as follows:

(i) $[0,1]$ is partitioned into 4, 8 and 16 segments I_j of equal length, respectively;

(ii) for $M = 4,8,16$ and $k = 1,2,3$ we define

$$e_{k,M} = \max_{i=1,\ldots,M-1} |y(x_i) - Z_{k,M}(x_i)|,$$

where $Z_{k,M} \in S_h$ is the solution of

(7.3) $\qquad \langle Z'_{k,M}, w'_h \rangle + \langle \exp(Z_{k,M}), w_h \rangle = 0, \qquad w_h \in S_h,$

$\langle \alpha,\beta \rangle$ defined by (4.1)-(4.3).

In the following table we list the quantities

(a) $e_{k,M}$, $\quad k = 1,2,3$; $\quad M = 4,8,16$;

(b) $r_{k,M} = \dfrac{\ln(e_{k,M} / e_{k,2M})}{\ln 2}$, $\quad k = 1,2,3$; $\quad M = 4,8$;

since $e_{k,M} \approx C(\frac{1}{M})^{2k}$, $r_{k,M}$ should have the approximate value 2k.

TABLE I; maximum errors and ratios for problem 1.

	k = 1	k = 2	k = 3
$e_{k,4}$	$5.03_{10}{-4}$	$2.87_{10}{-6}$	$2.49_{10}{-9}$
$r_{k,4}$	1.98	3.98	5.94
$e_{k,8}$	$1.27_{10}{-4}$	$1.82_{10}{-7}$	$4.04_{10}{-11}$
$r_{k,8}$	1.99	3.99	6.09
$e_{k,16}$	$3.19_{10}{-5}$	$1.14_{10}{-8}$	$5.91_{10}{-13}$

For each value of M three iterations of scheme (5.4) were needed to obtain the solution of (7.3) for k = 1.

This problem has been treated by several other authors.

(a) CIARLET et alii [1967] solved (7.1) by minimizimg the functional

$$I[w] = \int_0^1 ([w'(x)]^2 + 2e^{w(x)})dx$$

over the space $H_0^1(I) \cap P_N(I)$, i.e. the subspace of $H_0^1(I)$ consisting of polynomials of degree not greater than N ($N \geq 2$). This method gives for this example good results (e.g. a supremum error of $5.03_{10}{-8}$ for N = 6) but generally leads to ill-conditioned nonlinear systems.

(b) DE BOOR & SWARTZ [1973] solved (7.1) by collocation at sixth order Lobatto points, using twice differentiable Hermite quintics. They used a uniform grid $\pi : x_0 < x_1 < \ldots < x_M$. For M = 4 they found (with y_c the collocation solution):

$$\max_{i=1,2,3} |y(x_i)-y_c(x_i)| = 2.0_{10}{-9},$$

which is about the same as $e_{3,4}$ from table I.

(c) WEISS [1974] applied collocation at sixth order Lobatto points to solve the problem

$$\begin{cases} y' = z \\ z' = e^y, \quad x \in I, \quad y(0) = y(1) = 0. \end{cases}$$

He also used a uniform grid. For M = 3 he found

$$\max_{i=1,2} |y(x_i)-y_c(x_i)| = 2.66_{10}{-9},$$

which is also slightly greater than $e_{3,4}$ fron table I.

Example 2.

$$\frac{d}{dx}\left((1+e^x)\frac{dy}{dx}\right) = \exp(y+Ax+B), \qquad x \in [0,1] ;$$

$$\text{(7.4)} \qquad y(0) = y(1) = 0 ; \qquad A = \ln\frac{1+e}{2}, \; B = \ln 2.$$

This problem has the analytic solution

(7.5) $$y = \ell n(1+e^{x}) - Ax - B$$

Let $\pi: 0 = x_0 < x_1 < \ldots < x_m$ be a uniform partition of I for M = 4,8,16 and let $Z_{k,M} \in S_h$ be the solution of

(7.6) $$\begin{aligned} &\langle pZ'_{k,M}, w'_h\rangle + \langle f(\cdot, Z_{k,M}), w_h\rangle = 0, \quad w_h \in S_h\ ; \\ &p(x) = 1 + e^{x}; \\ &f(x,y) = \exp(y+Ax+B), \end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is defined by (4.1) - (4.3).

As with problem 1, we list the quantities $e_{k,M}$ and $r_{k,M}$.

TABLE II; maximum errors and ratios from problem 2.

	k = 1	k = 2	k = 3
$e_{k,4}$	$1.70_{10}{-4}$	$4.29_{10}{-7}$	$7.64_{10}{-11}$
$r_{k,4}$	1.99	4.01	6.00
$e_{k,8}$	$4.28_{10}{-5}$	$2.68_{10}{-8}$	$1.19_{10}{-12}$
$r_{k,8}$	2.00	4.00	6.25
$e_{k,16}$	$1.07_{10}{-5}$	$1.67_{10}{-9}$	$1.57_{10}{-14}$

Two iterations of scheme (5.4) were needed to solve (7.6) for k = 1, when we took $Z^{(0)}_{k,M} \equiv 0$.

Since no numerical results were known from the literature for this problem, no comparison with other problems was made.

Example 3

$$(7.7)\quad \begin{aligned} & y^{iv} - 2y'' + (y)^3 = -12\cos 2x + \sin^6 x, \quad x \in [0,\pi] \ ; \\ & y(0) = y'(0) = y(\pi) = y'(\pi) = 0. \end{aligned}$$

The analytic solution is $y = \sin^2 x$. A uniform grid $0 = x_0 < \ldots < x_M = \pi$ was made for $M = 4,8,16$. The solution of (7.7) was approximated in the space $S^{k,2}$ by the solution $Z_{k,M}$ of

$$(7.8)\quad \begin{aligned} & \langle Z''_{k,M}, w''_h \rangle + 2\langle Z'_{k,M}, w'_h \rangle + \langle Z^3_{k,M}, w_h \rangle = \\ & = \langle \sin^6 x - 12\cos 2x, w_h \rangle, \quad w_h \in S^{k,2} \end{aligned}$$

where $\langle \alpha,\beta \rangle$ is an approximation of (α,β) which is exact if $\alpha\beta \in P_{2k-3}(I_j)$, $j = 1,\ldots,M$. To that end we use k-point Lobatto quadrature.
Besides $e_{k,M}$ and $r_{k,M}$ we define for $k = 3,4,5$ and $M = 4,8,16$.

$$e'_{k,M} = \max_{i=1,\ldots,M-1} |y'(x_i) - Z'_{k,M}(x_i)|, \quad k = 3,4,5 \ ; \quad M = 4,8,16 \ ;$$

$$r'_{k,M} = \frac{\ell n(e'_{k} \ / e'_{k,2})}{\ell n 2} \ ; \quad k = 3,4,5 \ ; \quad M = 4,8 \ .$$

The results are given in table III.

TABLE III; maximum errors and ratios of problem 3

	k = 3	k = 4	k = 5
$e_{k,4}$	$4.82_{10}{-3}$	$2.92_{10}{-4}$	$9.11_{10}{-6}$
$r_{k,4}$	4.98	6.17	8.27
$e'_{k,4}$	$1.94_{10}{-2}$	$1.14_{10}{-4}$	$9.37_{10}{-6}$
$r'_{k,4}$	4.13	5.91	8.26
$e_{k,8}$	$1.53_{10}{-4}$	$4.04_{10}{-6}$	$2.96_{10}{-8}$
$r_{k,8}$	4.33	6.04	9.17
$e'_{k,8}$	$1.11_{10}{-3}$	$1.90_{10}{-6}$	$3.06_{10}{-8}$
$r'_{k,8}$	4.03	5.99	9.02
$e_{k,16}$	$7.58_{10}{-6}$	$6.15_{10}{-8}$	$5.12_{10}{-11}$
$e'_{k,16}$	$6.76_{10}{-5}$	$2.99_{10}{-8}$	$5.90_{10}{-11}$

Starting with $Z^{(0)}_{k,M} \equiv 0$ as an initial guess for the solution of (7.8) it took four Newton-Raphson iteration steps to solve (7.8) for k = 3. Since $\|Z_{3,M}-y\|_0 = O(h^4)$, only one further iteration step was needed to solve (7.8) for k = 4,5, using $Z_{3,M}$ as an initial guess. For k = 3 we used piecewise Hermite cubics, as a basis for $S^{k,2}$. Each iteration step involved the solution of a (2M-2)-dimensional linear system with positive definite pentadiagonal matrix (see STRANG & FIX [1973]).

Since no numerical results were known from the literature for this problem, no comparisons were made.

REFERENCES

ABRAMOWITZ, M. & I.A. STEGUN, (eds.), *Handbook of Mathematical functions*, Dover Publications, Inc., New York, 1965.

CODDINGTON, E.A. & N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York - Toronto - London, 1955.

DE BOOR, C. & B. SWARTZ, *Collocation at Gaussian points*, SIAM J. Num. An. 10, p. 582-607 (1973).

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

CIARLET, P.G. & M.H. SCHULTZ & R.S. VARGA, *Numerical Methods of High-Order Accuracy for Nonlinear Boundary Value Problems*, I One Dimensional Problem, Num. Math. 9, p. 394-430 (1967); V Monotone Operators Theory, Num. Math. 13, p. 51-77 (1969).

DOUGLAS, J.jr. & T. DUPONT, *Some superconvergence Results for Galerkin Methods for the Approximate Solution of Two-Point Boundary Problems*, from: J.J.H. MILLER (ed.), *Proceedings of the Royal Irish Academy Conference on Numerical Analysis*, 1972, Academic Press, London and New York, 1973.

——— & ———, *Galerkin Approximation for the Two Points Boundary Problem using Continuous, Piecewise Polynomical Spaces*, Num. Math. 22, p. 99-109 (1974).

———, Private Communication, 1975.

HEMKER, P.W., *Galerkin's method and Lobatto points*, MC report NW 24/75, Mathematisch Centrum, 1975.

HERBOLD, R.J. & M.H. SCHULTZ & R.S. VARGA, *The Effect of Quadrature Errors in the Numerical Solution of Boundary Value Problems by Variational Techniques*, Reg. Math. 3, p. 247-270 (1969).

ORTEGA, J.M. & W.C. RHEINBOLDT, *Iterative solution of nonlinear equations in several variables*, Academic Press, London and New York, 1970.

PROTTER, M.H. & H.F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

RUSSELL, R.D., *Collocation for Systems of Boundary Value Problems*, Num. Math. 23, p. 109-133 (1974).

———, *A Comparison of Collocation and Finite Differences for Two-Point Boundary Value Problems*, to appear.

STRANG, G. & G.J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1973.

SCHULTZ, M.H., *Spline Analysis*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1973.