ASPECTS OF THE FINITE ELEMENT METHOD

MIENTE BAKKER

UvA
STELLINGEN

bij het proefschrift

ASPECTS OF THE FINITE ELEMENT METHOD

van

M. BAKKER

3 november 1982

BIBLIOTHEEK MATHEMATISCH CENTRUM
AMSTERDAM
De $C^0$-Collocatie-Galerkin-methode kan gezien worden als een $C^0$-Galerkin-methode met gebruik van Lobatto-kwadratuur.


II

Zij gegeven het Stefan-probleem

\[ \begin{align*}
\frac{\partial T}{\partial z} - \frac{1}{h} \left( \kappa(T) \frac{\partial T}{\partial t} \right) &= 0, & z > 0, & \kappa(T) > 0, \\
\frac{\partial T}{\partial z} &= 0, & z = 0, & T(z,0) = T_0 (= 300^\circ \text{K}), \\
\frac{\partial T}{\partial z} &= 0, & z = 0, & T_\infty = 0, \\
\kappa(T) &> 0, & \kappa(T) &> 0.
\end{align*} \]

\[ \begin{align*}
x(t) &= \int_0^t \frac{T_\infty - T(T_0, t)}{T_\infty - T(t)} \, dt, & \text{als } T_\infty = T(t), & T_0 = 1685^\circ \text{K}, \\
x(t) &= \frac{T_\infty - T_0}{T_\infty - T(t)} \frac{1}{T_0} \int_0^t \kappa(T) \frac{\partial T}{\partial t} \, dt, & \text{als } T_\infty > T(t).
\end{align*} \]

\[ x(t) = \frac{(1-x)x(t) + x_0}{1-x}, \quad z \in (0,1), \quad \text{als } z \in (s(t), x_0); \]

waarna de gekoppelde stelsels begin-randwaardeproblemen opgelost kunnen worden door een combinatie van de "Backward Euler" methode en de eindige differentiële- of eindige elementen methode [1].

De bewering van M.F. Wheeler c.s. dat $C^2$-Galerkin-methoden voor cirkel- en bollsymmetrische gebieden superconvergent zijn in de roosterpunten is onjuist wat het middelpunt van de bol of de cirkel betreft.


Het spectrum van de Boltzmann botsingsoperator van een gas met harde bollen bestaat uit een continu en een discreet gedeelte. De bewering van Résibois en De Leener dat het spectrum discreet is, is derhalve niet juist.


In [2] verklaart Wood dat bij laser-anneling de grootte van de absorbtiecoefficient vanaf $2,0 \times 10^4 \text{ cm}^{-1}$ niet meer van invloed is op de smeltdiepte en dat dit wordt bevestigd door de resultaten van het computerprogramma HEATINGS. Deze bewering is in tegenspraak met de resultaten van het computerprogramma WANG [1].


\[ \int_0^\infty J_{2\nu}(2zr)z^\nu \cos \theta \, dz = I_{\nu}^2(z), \quad z \in \mathbb{C}. \]
Laat Turkin's functie $T_m(z,a)$ gedefinieerd zijn door

\[(1) \quad T_m(z,a) = \sum_{n=-\infty}^{\infty} \frac{J_n(z)J_{n-m}(z)}{n-a}; \quad n \in \mathbb{Z}, a \notin \mathbb{Z}, z \in \mathbb{C};\]

waarin $J_m(z)$ de bekende Besselfunctie voorsteelt. Voor $T_m(z,a)$ geldt de inhomogene recursie

\[(2) \quad T_{m-1}(z,a) + T_{m+1}(z,a) = \frac{2(a-m)}{z} T_m(z,a) + \frac{2}{z} \delta_{m,0}; \quad m \in \mathbb{Z}\]

waarbij $\delta_{m,0}$ het symbool van Kronecker is. Op grond van (1) en (2) kan de bewering van Neuberger dat

\[(3) \quad T_m(z,a) = (-1)^{m+1} \frac{\nu}{\sin \nu} J_{m-n}(z)J_n(z)\]

niet correct zijn voor alle $m \in \mathbb{Z}$. 


VIII

Als de voorstellen van de tweede Commissie-Wagner om de ontslag-procedures te vereenvoudigen, worden uitgevoerd, zal het verschil in rechtspositie tussen overheidsdienaren en andere werknemers nog groter worden.

NRC/Handelsblad, 30 juni 1982.

IX

Het in wetsartikelen gericht tegen discriminatie met name noemen van één of meer speciale categorieën, is discriminerend in positieve zin ten opzichte van elke niet genoemde categorie.

X

De nieuwe wet op het basisonderwijs, welke beoogt het onderwijspeil te verhogen, zal in de praktijk zoveel extra werk voor het onderwijsend personeel met zich meebrengen dat te vrezen valt dat het onderwijspeil door deze wet zal dalen.
ASPECTS OF THE FINITE ELEMENT METHOD
ASPECTS OF THE FINITE ELEMENT METHOD

ACADEMISCH PROefsCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR IN DE WISKUNDE EN NATUURWETENSCHAPPEN AAN DE UNIVERSITEIT VAN AMSTERDAM, OP GEZAG VAN DE RICTOR MAGNIFICUS DR. R.W. BRESTERS, HOOGLEERAAR IN DE FACULTEIT DER WISKUNDE EN NATUURWETENSCHAPPEN, IN HET OPENBAAR TE VERLEDIGEN IN DE AULA DER UNIVERSITEIT (TIJDELIJK IN DE LUTHERSE KERK, INGANG SINGEL 411, HOEK SPU1) OP WOENSDAG 3 NOVEMBER 1982 TE 15.00 UUR

DOOR

MIENTE BAKKER

geboren te den Haag

1982

MATHEMATISCH CENTRUM, AMSTERDAM
PROMOTOR : PROF.DR. P.J. VAN DER HOUWEN
COPROMOTOR : DR. P.W. HEMKER
COREFERENT : PROF.DR. M. VAN VELDHUIZEN
aan tineke
aan mijn familie
CONTENTS

INTRODUCTORY PART

1. INTRODUCTION i
2. THE FINITE ELEMENT METHOD i
   2.1 The essence of the method iii
   2.2 Properties of the method iv
   2.3 One-dimensional problems v
   2.4 Time-dependent problems v

REFERENCES vi

THE FIVE PAPERS

On the Numerical Solution of Parabolic Equations in a Single Space Variable by Means of the Continuous Time Galerkin Method 1

A Note on C\(^0\) Galerkin Methods for Two-Point Boundary Problems 17

Galerkin Methods for Even-Order Parabolic Equations in One Space Variable 24

A Program to Solve Rotating Plasma Problems 41

A Program to Solve a Solute Diffusion Problem with Segregation at a Moving Interface 52

SAMENVATTING 65
ACKNOWLEDGEMENTS

The author wishes to thank the board of directors of "Mathematisch Centrum" for giving him the opportunity to carry out the research described in this thesis. He also wants to thank them for publishing this thesis.

He wishes to thank prof.dr. P.J. van der Houwen for being his promotor.
He wishes to thank dr. P.W. Hemker for being his copromotor, for initiating him into the Finite Element Method and for the many discussions on that subject.
He wishes to thank prof.dr. M. van Veldhuizen for his willingness to be coreferent and for the numerous discussions on the Finite Element Method.
He wishes to thank dr. M.S. van den Berg and dr. D. Hoonhout, the co-authors of the papers [D] and [E], respectively, for their cooperation and for giving him such nice material.
He wishes to thank drs. E. Slagt for allowing him to do scientific research besides his other duties.
He wishes to thank mr. A.C. IJsselstein for his assistance with the typesetting of the introductory part of this thesis.
Finally, he wishes to thank mr. D. Zwarst and his team for the printing of this thesis and mr. T. Baanders for the design of the cover and the figures.
INTRODUCTORY PART
1. INTRODUCTION

This thesis consists of five scientific papers [A-E]. They all relate to the numerical solution of parabolic or elliptic partial differential equations by means of the Finite Element Method.

The papers [A-C] deal with the Finite Element solution of (initial) two-point boundary value problems. Paper [D] describes a computer program to solve a two-dimensional elliptic problem arising from plasmaphysics. Paper [E] describes a computer program to solve a one-dimensional moving boundary problem arising from laser annealing. Papers [D-E] were written in cooperation with the physicists M.S. van den Berg and D. Hoomhout, respectively, who in their turn used the programs to produce numerical results for their theses [2,19].

In order to outline the framework into which the five papers fit, they are preceded by a short sketch of the Finite Element Method with a brief history of its origins.

2. THE FINITE ELEMENT METHOD

The Finite Element Method is a now-a-days highly popular method to solve partial differential equations by computer. Especially in the field of elliptic problems (potential theory, elasticity theory) and parabolic problems (heat and mass transport), it has become a powerful competitor of the Finite Difference Method. It was not, however, before the 1970s, that the Finite Element Method gained wide popularity among applied mathematicians, while it already existed in the 1950s (or in the 1940s, if Courant's early paper [6] is taken into account). This was largely due to the fact that (see also [23]) the Finite Element Method was originally developed by a group of mathematical outsiders: the engineers.

In the 1950s, with the advance of the electronic computers, aeronautic engineers developed numerical methods to solve complicated problems in structural mechanics. These Finite Element Methods, as they were called, were effective but were not yet based on (solid) mathematical foundation. It lasted until the late 1960s, before it gradually became clear that the engineers were practicing a modern version of an old mathematical method: the Galerkin method, also called Ritz-Galerkin or Rayleigh-Ritz-Galerkin method. In this field of approximation theory, extensive literature was already available (see [24,28] for references). Once the mathematical basis became clear, the Finite Element Method was felt mathematically acceptable and became increasingly popular, witness the rising number of publications on this subject since the early 1970s. For that break-through, much of the credit must be given to the engineers Argyris, Clough and Zienkiewicz.

The Finite Difference Method, on the other hand [7,14], was already known to applied mathematicians since the 1920s and had proved to be an efficient and accurate method to solve many kinds of differential equations. For this reason and because the two methods showed many similarities, it was rather remarkable that the Finite Element Method gained so much ground in the field of elliptic and parabolic problems. Reasons were a.o.

- problems with complex domains and complex meshes could readily be solved;
- matrix and right hand side vector of Finite Element schemes were constructed in a systematic and efficient way [13,28];
- the error analysis had a broad foundation in approximation theory and functional analysis [3,5,24,28].

Still, there are fields where the Finite Element Method is performing less well, such as first order wave equations with discontinuous initial conditions. Therefore, it seems necessary for a numerical analyst who wants to solve partial differential equations, to be familiar with both methods.
Figure 1.
(a) profile of $U(x)$; (b) profile of $\phi_i(x)$
2.1. The essence of the method.

Many partial differential equations arising from engineering problems can be formulated in variational form: find a function \( u(x) \) that minimizes some energy expression \( E(v) \), i.e., find \( u(x) \) such that \( E(u) \leq E(v) \) for all "admissible" functions \( v \). For example, the solution \( u \) of the two-point boundary value problem

\[
\frac{d^2u}{dx^2} = f(x), \quad u(0) = u(1) = 0
\]

minimizes the energy functional \( E(v) \) defined by

\[
E(v) = \int_0^1 (v'^2 - 2f v) \, dx
\]

over the space

\[
V = \{ v \mid v \in H^1(0,1); \quad v(0) = v(1) = 0 \}.
\]

In 1915, Galerkin [15] initiated the idea of approximating \( u \) by a \( U(x) = \sum_{i=1}^N c_i \phi_i(x) \) that minimizes \( E(\sum_{i=1}^N c_i \phi_i(x)) \) for all possible real numbers \( c_1, \ldots, c_N \). Here the set \( \{ \phi_i \}_{i=1}^N \) is a basis of an a priori selected subspace of admissible functions. Examples of such basis functions are: trigonometric functions, Bessel functions and Legendre functions. In this way, the problem can be solved by minimizing a function of several variables. In its classical form, the Galerkin method is successful as long as \( \{ \phi_i \}_{i=1}^N \) are eigenfunctions of the differential operator and are hence orthogonal with respect to some weight factor. In that case, the Hessian matrix \( \sum_{i,j=1}^N \phi_i \phi_j \frac{d^2}{dx^2} \) is diagonal and computation of \( c_1, \ldots, c_N \) is rather simple. If it is not possible to construct a set of eigenfunctions, application of the classical Galerkin method may become difficult, if not impossible.

The modern Galerkin method, on the other hand, uses a basis of "near-orthogonal" functions: the Hessian matrix is sparse. This sparseness is due to the simple fact that the basis functions have small support, i.e., each of them is only non-identically zero on a small part of the domain of the function. In this form, the Galerkin method is also known as the Finite Element Method.

In the simplest form of the Finite Element Method, a function of one variable is approximated by a piecewise linear function (fig. 1a). In that case, the functions \( \phi_i \) of the representation

\[
U(x) = \sum_{i=1}^N c_i \phi_i(x)
\]

are the hat or chapeau functions (fig. 1b). The weights \( c_i \) have direct physical meaning: they correspond with the values of \( U(x) \) at \( x_i \). The profile of \( U(x) \) is a polygon with vertices on the lines \( x = x_i \).

There are numerous ways to generalize the example of the hat functions. For instance, the \( \phi_i \) can be piecewise quadratics or piecewise cubics. In the case of two-dimensional problems, the \( \phi_i \) are pyramidal functions, etc. [22,24,28]. But, no matter how complicated they are, all these spline functions are generally constructed in basically the same way:

- the domain of the function to be approximated is divided into small pieces, also called segments or finite elements. In \( \mathbb{R}^1 \), these elements are plain intervals, in \( \mathbb{R}^2 \), they are triangles, quadrangles, etc., in \( \mathbb{R}^3 \), they are tetrahedra, blocks, etc.;
- the basis functions are piecewise polynomial, i.e., they are of polynomial form on every segment of the mesh (fig. 1b);
- only on a few adjacent elements, these functions are non-identically zero (fig. 1b); they are said to have small support;
at specific points, the nodes or nodal points, the $\phi_i$ or their partial derivatives have prescribed values, usually 0 or 1 (fig. 1b); this property gives the coefficients $c_i$ of $\sum_{i=1}^{N} c_i \phi_i(x)$ their direct physical meaning.

- depending on the energy function to be minimized, the $\phi_i$ have to satisfy some smoothness conditions, such as continuity, differentiability, etc., with discontinuities of the first or second derivatives at the interfaces of the segments.

This way of approximating a function is an essential part of most versions of the Finite Element Method. Of course, there exist more sophisticated versions, such as exponentially fitted methods for boundary layer problems [17] or isoparametric methods for domains with curved boundaries [28], but they share the properties of basis functions with small support and prescribed function (and derivative) values at the nodes.

2.2. Properties of the method.

In this section, some properties of the Finite Element Method are mentioned that may explain why it has become so popular.

Robustness.

In practice, it is almost always possible to generate a mesh and to construct a basis of spline functions to approximate the solution.

Sparse-matrix systems.

For many problems, application of the Finite Element Method leads to a (linear or nonlinear) system of the form

$$F_i(c_1, \cdots, c_N) = 0, \quad i = 0, \cdots, N;$$

where the Jacobian matrix $(\partial F_i / \partial c_j)$ is sparse and of banded structure. E.g., it is easily verified that minimization of $E(v)$ defined by (2), over a space of piecewise linear functions leads to a tridiagonal system of linear equations. In more dimensions, the Jacobian has a more complex but nonetheless clear structure and is still sparse. By its sparseness, (3) can be solved by relatively cheap and fast methods, such as iterative methods (SOR, GS [32], CG [27], MG [18], ICCG [21]) or direct methods (LU decomposition [31], nested dissection [16, 25]).

Error estimates.

Most Finite Element Methods have the property that the rate by which the Galerkin approximation converges to the true solution in some specific sense, is a priori known. For instance, let $U$ be the piecewise linear function that minimizes the energy functional $E(v)$ defined by (2). Then it is known that [4,17,24]

$$||u - U||_{L^2(\Omega)} \leq C h^2;$$

where $h$ is the mesh-size: the maximum of the elements' diameters. $C$ is independent of $h$. This property makes it possible to estimate the $L^2$ error or to improve the Galerkin approximation by some refinement method, e.g. Richardson extrapolation. Error bounds like (4) are very commonplace in Finite Element theory. They are usually of the form $Ch^{k+1}$; $k$ is called the degree of the Galerkin approximation because for piecewise $k$-th degree polynomials $||u - U||_{L^2(\Omega)}$ vanishes.

For the proofs of (4) and similar bounds, the theory of Sobolev spaces is amply used, because the space of admissible functions for the minimization of (2) is a Sobolev space $W^{m,p}_0(\Omega)$, i.e. a Banach space of functions that, together with their first $m$ distributional partial derivatives, are $L^p$ integrable over the domain $\Omega$ and which satisfy certain boundary conditions. For $V$ defined by (2b), $p = 2$, $m = 1$, $\Omega = (0,1)$. It is especially the approximation of Sobolev spaces by spaces of spline functions that gives the Finite Element Method its broad theoretical support [3,5,24,28].
Quadrature rules.

In practice, the energy functional $E(v)$ cannot be evaluated exactly, hence some quadrature rule has to be applied, e.g., the extended trapezoidal rule if $E(v)$ is minimized over a space of piecewise linear functions. If proper quadrature rules are used, the order of convergence is preserved [A-C,11,24,26,28]. In some cases, application of numerical quadrature can even lead to systems of equations that are substantially sparser than if $E(v)$ would be evaluated exactly. In paper [D] of this thesis, e.g., a matrix with 11 diagonals instead of 27 diagonals is constructed in this way.

2.3. One-dimensional problems.

One-dimensional Finite Element Methods, although not the most important ones, have some convergence properties that appeal to many numerical analysts. In the early 1970s, it was detected that the error function and sometimes also some of its derivatives were essentially smaller at the mesh-points than elsewhere [1,8,10,11] on the domain. This phenomenon was called superconvergence. Pioneers were Douglas, Dupont, DeBoor and Swartz. Later, superconvergence was also found at other points on the interiors of the segments [B,C,4,30]. In papers [B,C] it is proved that this kind of superconvergence occurs at the zeros of special Jacobi polynomials shifted to the segments.

2.4. Time-dependent problems.

A special class of problems that are solved by the Finite Element Method are the transient problems, such as heat conduction and diffusion problems. The essence of the Finite Element Method is here that the temperature or mass density is represented by

$$U(x,t) = \sum_{i=1}^{N} c_i(t) \phi_i(x).$$

One important method of solving a transient problem is the continuous time Galerkin method or Faedo-Galerkin method [9,24,28,29]. The partial differential equation is discretized in its space variables and the resulting ordinary differential equation is integrated by some adaptive time-integrator [20]. Application of this method generally leads to implicit initial value problems of the form

$$G \frac{dc}{dt} + Ac = b, \ t \geq 0; \ c(0) = c_0;$$

where $c(0)$ is selected such that $\sum_{i=1}^{N} c_i(0) \phi_i(x)$ is a proper approximation of $u(x,0)$. In [A,C,9,24,28,29], it is proved that $U(x,0)$ can be defined such that

$$||u(t) - U(t)||_{L^2(\Omega)} \leq C(t,\mu) h^{k+\epsilon}, \ t \geq 0;$$

where $k$ is the degree of the Finite Element space. A frequent choice of $c_i(0)$ is the interpolator of $u$ and its derivatives at the nodal points. In [A,C] of this thesis, it is proved that for parabolic equations with one space variable, interpolation at Jacobi points is a good initial approximation of $u(x,0)$.

In (6), the mass matrix $G$ and the stiffness matrix $A$ are both banded and usually of the same structure. However, there exist some versions of the Faedo-Galerkin method [A,26] where $G$ is diagonal, which makes (6) purely explicit.

In the case of one space variable, the phenomenon of superconvergence also occurs, provided that $U(x,0)$ is defined properly [A,C,12]. This superconvergence, however, is not uniform in time as paper [C] shows with a simple counter-example.
REFERENCES.

THE FIVE PAPERS.

If they are cataloged in Mathematical Review or Zentralblatt für Mathematik, the catalog numbers are given.


OTHER REFERENCES.


THE FIVE PAPERS
ON THE NUMERICAL SOLUTION OF PARABOLIC EQUATIONS IN A SINGLE SPACE VARIABLE BY THE CONTINUOUS TIME GALERKIN METHOD

MIENTE BAKKER

Abstract. We consider the Galerkin method to solve a parabolic initial boundary value problem in one space variable, using piecewise polynomial functions and give an alternative proof of superconvergence. Then by means of Lobatto quadrature, we obtain purely explicit vector initial value problems without loss in the order of accuracy, global or pointwise.

1. Introduction. We consider the linear initial boundary value problem

$$
\frac{\partial u}{\partial t} = -Lu = \frac{\partial}{\partial x} \left[ p(x) \frac{\partial u}{\partial x} \right] - q(x) u;
$$

(1.1)

$$
x \in [0, 1] = I; \quad t \in [0, \infty) = J;
$$

$$
u(t, 0) = u(t, 1) = 0; \quad u(0, x) = u_0(x),
$$

where the functions $p(x), q(x)$ and $u_0(x)$ are supposed to be sufficiently smooth, which will be specified later, and where $u_0(0) = u_0(1) = 0$.

In §2 we introduce some notations we need throughout this paper and give a summary of the theory on the continuous time Galerkin method (also called Paedo-Galerkin method) using continuous piecewise polynomials.

Douglas, Dupont and Wheeler [8], [9], [11] have proved superconvergence at the knots for this type of problem. In §3 we obtain a similar, albeit nonuniform, result in an alternative way, for doing this we use Laplace transform as Cerrutti and Parter [4] have done for collocation methods.

In §4, we use Lobatto quadrature formulas; this yields purely explicit vector initial value problems with sparse Jacobian, where very easily computable initial data are to be provided in order to preserve accuracy. Finally, in §5, we give a simple numerical example.

Throughout this paper $C, C_1, C_2, \cdots$ will denote positive constants not necessarily the same.

2. The continuous time Galerkin method.

2.1. Notations. For any interval $E \subset I$, we introduce the Sobolev space $H^m(E)$, $m \geq 0$, by

$$
H^m(E) = \{ v | D^j v \in L^2(E), j = 0, \cdots, m \},
$$

(2.1)

where $D^j$ denotes $d^j/dx^j$. $H^m(E)$ has the usual Sobolev inner product and norm

$$
(\alpha, \beta)_{L^2(E)} = \int_E \alpha(x) \beta(x) \, dx,
$$

(2.2b)

where

$$
\| \alpha \|_{H^m(E)} = \sum_{j=0}^{m} (\alpha, D^j \alpha)_{L^2(E)}^{1/2}.
$$

(2.2a)

* Received by the editors August 22, 1978, and in revised form April 16, 1979.
† Mathematisch Centrum, Amsterdam, The Netherlands.
PARABOLIC EQUATIONS IN A SINGLE SPACE VARIABLE

In the sequel, we use \((u, v)\) instead of \((u, v)_{1/2}\), and \(\|u\|_m\) instead of \(\|u\|_{M^{-1}}\).

The subspace \(H^1_0(I)\) of \(H^1(I)\) is defined by

\[
H^1_0(I) = \{u \in H^1(I) : u(0) = u(1) = 0\}.
\]

Next we define the bilinear functional \(B : H^1_0(I) \times H^1_0(I) \to \mathbb{C}\) associated with the operator \(L, L\) defined by (1.1), as follows:

\[
B(u, v) = (pu, v) + (qu, v); \quad u, v \in H^1_0(I).
\]

We assume that \(p(x)\) and \(q(x)\) are such that \(B\) is strongly coercive, i.e., there is a \(C\) such that

\[
B(u, u) \geq C\|u\|_1^2, \quad u \in H^1_0(I).
\]

A sufficient condition is \(p(x) \geq p_0 > 0, q(x) \geq q_0 > -p_0 \alpha^2\) (see [3]).

2.2. Galerkin's method. Let \(u : J \to H^1_0(I) \cap H^2(I)\) be the solution of (1.1). Then, as is well known (see e.g. [18], [19]), \(u\) can be approximated by a \(U : J \to S\), where \(S\) is some suitable finite-dimensional subspace of \(H^1_0(I)\). This \(U\) is given by the equation

\[
\begin{align*}
&\frac{\partial}{\partial t} U(t, \cdot) + B(U(t, \cdot), V) = 0, \quad V \in S, \ t \geq 0, \\
&U(0) = U_0 \in S,
\end{align*}
\]

where \(U_0\) is some suitable approximation to \(u_0\).

For \(S\) we select the following subspace. Let \(\Delta : 0 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1\) be a uniform partition of \(I\), i.e. \(x_j = jh = jN^{-1}, j = 0, \ldots, N\). By \(I_j\), we denote the segment \([x_{j-1}, x_j]\). We define the space \(M^0_\lambda(\Delta)\) (\(\lambda\) a constant positive integer) by

\[
M^0_\lambda(\Delta) = \{V \in H^1_0(I) : V \in P^l(I_j), j = 1, \ldots, N\},
\]

where for any interval \(E \subset IP(E)\) denotes the space of polynomials of degree \(d \leq l\) restricted to \(E\).

We also define the partition norm with respect to \(\Delta\) and \(H^\kappa(\Delta)\) by

\[
\|u\|_{\Delta}^2 = \sum_{j=1}^N \|u_j\|^2_{p^j}.
\]

\[
H^\kappa(\Delta) = \{u \in H^\kappa(I), \; j = 1, \ldots, N\}.
\]

Lemma 1. Let \(U_0 \in M^0_\lambda(\Delta)\) be an approximation to \(u_0\) satisfying

\[
\|u_0 - U_0\|_0 = O(h^{-\lambda})|u_0|_{\lambda+1}
\]

and let \(U : J \to M^0_\lambda(\Delta)\) satisfy (2.6) for all \(t\) with \(U_0\) as initial function. Then the error function \(e(t) = u(t) - U(t)\) has the \(L^1\) bound

\[
\|e(t)\|_1 = e^{-1}\|e(0)\|_1 + C\delta^{-1}\|u(t)\|_{\lambda+1} + \int_0^t e^{\lambda(\tau-t)}\|u_\tau\|_{\lambda+1} d\tau,
\]

where \(\lambda\) is the smallest (positive) eigenvalue of the operator \(L\).

Proof. For the proof see [16].

3. Superconvergence at the knots. As in the case of the two-point boundary value problems (see [7], [10]), the order of convergence at the knots is much higher than the global order of convergence, namely \(O(h^\lambda)\) vs. \(O(h^\kappa)\). Douglas et al. ([8], [9], [11]) gave proofs for several continuous time Galerkin methods. In this section, we intend to give a proof based on the use of the Laplace transform combined with the superconvergence results on two-point boundary value problems (see also [4]).
For any \( v: J \rightarrow V \), \( V \) a set of functions defined on \( I \), we define the Laplace transform \( \mathcal{L} = \hat{\mathcal{L}} \) by

\[
\hat{v}(s, x) = \int_{t_0}^{\infty} e^{-st}v(t, x) \, dt, \quad s \in \mathbb{C}, \quad x \in I.
\]

(3.1)

According to the semi-group theory on stationary differential operators (see e.g. [12]), we can apply \( \mathcal{L} \) to problem (1.1) to obtain for \( \hat{u}(s) = \mathcal{L}u \) the two-point boundary value problem

\[
\begin{align*}
L \hat{u}(s) + s \hat{u}(s) &= u_0, \quad x \in I; \\
\hat{u}(s, 0) - \hat{u}(s, 1) &= 0.
\end{align*}
\]

(3.2a)

We recall that the definition (2.2) of inner product also extends to complex-valued functions, and therewith the definitions of \( B(u, v) \) and \( (u, v)_{H^1(I)} \). We write (3.2a) in its Galerkin form to obtain

\[
B(\hat{u}(s), v) + s(\hat{u}(s), v) = (u_0, v), \quad v \in H^1_0(I).
\]

(3.2b)

The solution of (3.2b) can be approximated in \( M_0^r(\Delta) \) by the solution \( \hat{U}(s) \) of

\[
B(\hat{U}(s), V) + s(\hat{U}(s), V) = (u_0, V), \quad V \in M_0^r(\Delta).
\]

(3.3)

At the other hand, the solution \( U(t) \in \mathcal{L}^r(\Delta) \) of (2.6) has the Laplace transform \( \hat{U}_1(s) \) which is the solution of the boundary value problem

\[
B(\hat{U}_1(s), V) + s(\hat{U}_1(s), V) = (U_0, V), \quad V \in M_0^r(\Delta).
\]

(3.3a)

One easily verifies that if \( U_0 \in \mathcal{L}^r(\Delta) \) is defined by

\[
(u_0 - U_0, V) = 0, \quad V \in M_0^r(\Delta),
\]

then \( \hat{U}_1(s) = \hat{U}(s) \), which is illustrated in diagram 1 (see next page). Also, one easily verifies that \( U_0 \) satisfies (2.9), since it minimizes \( \|u_0 - V\| \) over \( M_0^r(\Delta) \).

It is known that (see e.g. [16], [18])

\[
\begin{align*}
u(t, x) &= \sum_{n=1}^{\infty} a_n e^{-\lambda_n t} \phi_n(x); \\
\hat{u}(s, x) &= \sum_{n=1}^{\infty} a_n s + \lambda_n^{-1} \phi_n(x); \\
a_n &= (u_0, \phi_n), \quad n = 1, 2, \cdots; \\
U(t, x) &= \sum_{n=1}^{M} A_n e^{-\lambda_n t} \phi_n(x); \\
\hat{U}(s, x) &= \sum_{n=1}^{M} A_n s + \lambda_n^{-1} \phi_n(x); \\
A_n &= (U_0, \phi_n), \quad n = 1, 2, \cdots, M = rN - 1.
\end{align*}
\]

(3.5)

where \( \lambda_1 \leq \lambda_2 \leq \cdots \) are the (positive) eigenvalues of \( L \) with orthonormal eigenfunctions \( \phi_1, \phi_2, \cdots \) and where \( \Lambda_1 \leq \Lambda_2 \leq \cdots \leq \Lambda_M \) are the eigenvalues of the eigenvalue problem

\[
B(\Phi_n, V) = \Lambda_n(\Phi_n, V), \quad V \in M_0^r(\Delta), \quad n = 1, \cdots, M.
\]

(3.6)
with orthonormal eigenfunctions $\Phi_1, \Phi_2, \ldots$. Note that $\lambda_1 > \lambda_1$ since

$$\lambda_1 = \inf_{V \in \mathcal{D}(G)} \frac{B(V, V)}{(V, V)} \geq \inf_{(V, V) > 0} \frac{B(V, V)}{(V, V)} = \lambda_1.$$ (3.7)

Hence, the error function $\tilde{\varepsilon}(s, x) = \tilde{u}(s, x) - \tilde{U}(s, x)$ is meromorphic in the complex plane with the set $\{-\lambda_n\}_{n=1}^\infty U \{-\lambda_n\}_{n=1}^\infty$ as only poles.

From Laplace theory (see e.g. [6]), we know that the inverse $\mathcal{L}^{-1}$ of $\mathcal{L}$ is given by

$$\varphi(t, x) = \mathcal{L}^{-1}\tilde{\varepsilon}(s, x) = \frac{1}{2\pi i} \int C e^{st} \tilde{\varepsilon}(s, x) \, ds,$$ (3.8)

where $C = \{s | s = \sigma + i\tau, -\infty < \tau < +\infty\}$ is a contour in the convergence half-plane of $\tilde{\varepsilon}(s).$ For $\tilde{\varepsilon}(s) = \tilde{\varepsilon}(s),$ the convergence half-plane is given by $\Re s > -\lambda_1.$ One intuitively feels now that one can use local convergence results for $\tilde{\varepsilon}(s, x)$ to establish similar results for $\varphi(t, x)$ by applying (3.8). There are, however, two obstacles:

(a) The convergence of (3.8) is not absolute;

(b) The standard convergence results for two-point boundary problems do not automatically apply to (3.2), since the problem becomes singularly perturbed as $|x| \to \infty.$

What we have to do is the following:

(1) Replace (3.8) by an absolute convergent integral with another integration path;

(2) Prove that on the new integration path the convergence results for two-point boundary problems hold.

For $\Gamma$ we take the imaginary axis ($\sigma = 0$). We now define the contours $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$ and $\Gamma_5$ by (see Fig. 1)

$$\Gamma_1 = \{s | s = \tau + i\tau; -R \leq \tau \leq R\};$$

$$\Gamma_2 = \{s | s = \alpha + iR; -R \leq \alpha \leq 0\};$$

$$\Gamma_3 = \{s | s = \alpha(-1 + i); 0 \leq \alpha \leq R\};$$

$$\Gamma_4 = \{s | s = -\alpha(1 + i); 0 \leq \alpha \leq R\};$$

$$\Gamma_5 = \{s | s = \alpha - iR; -R \leq \alpha \leq 0\};$$

where $R$ is an arbitrary positive number. Since $\tilde{\varepsilon}(s)$ has no poles outside the negative real axis, we can apply the main theorem of Cauchy on complex analysis to obtain the important relation

$$\int_{\Gamma_{\epsilon=1,P_2,P_3,P_4}} \tilde{\varepsilon}(s, x) e^{st} \, ds = 0, \quad x \in I, \quad t \in I,$$ (3.10)

where by $P_1P_2 \cdots P_n$ we mean the polygonal line starting in $P_1$ connecting $P_i$ with $P_{i+1}$.
etc. and ending at $P^*$. We now consider $F_1F_2 = \Gamma_2$. If we apply (3.2) for $v = \hat{u}(s)$, $s = -\alpha + iR$, $0 \leq \alpha \leq R$, we get

\begin{equation}
B(\hat{u}(s), \hat{u}(s)) + (iR - \alpha)(\hat{u}(s), \hat{u}(s)) = (\hat{u}(s), u_0).
\end{equation}

If we take the squared moduli of both sides, we get

\begin{equation}
|B(\hat{u}(s), \hat{u}(s)) - \alpha(\hat{u}(s), \hat{u}(s))|^2 + R^2|\hat{u}(s), \hat{u}(s)|^2 = |(\hat{u}(s), u_0)|^2, \quad s = -\alpha + iR.
\end{equation}

It is easily seen from (3.12) that

\begin{equation}
R|\hat{u}(s), \hat{u}(s)| \leq \|\hat{u}(s), u_0\|; \quad \frac{R}{\sqrt{\alpha^2 + R^2}}|B(\hat{u}(s), \hat{u}(s))| \leq \|\hat{u}(s), u_0\|;
\end{equation}

hence

\[ \|\hat{u}(s)\|_b \leq R^{-1}\|u_0\| \]

and

\begin{align}
\|\hat{u}(s)\|_b^2 & \leq CB(\hat{u}(s), \hat{u}(s)) \leq C\|\hat{u}(s)\|_b\|u_0\| \leq CR^{-1}\|u_0\|; \\
\|\hat{u}(s)\|_b & \leq CR^{-1/2}\|u_0\|; \quad \|\hat{u}(s, x)\| \leq \|\hat{u}(s)\|_b \leq CR^{-1/2}\|u_0\|.
\end{align}
The next-to-last inequality is obtained by applying Poincaré's inequality. In a similar way one proves that
\begin{equation}
\tilde{U}(s, x) \preceq CR^{-1} |U_0|_h,
\end{equation}
hence
\begin{equation}
\hat{\epsilon}(s, x) \preceq CR^{-1/2} |u_0|_h + |U_0|_h,
\end{equation}
on \Gamma_2.

After this, we obtain
\[ \int_{\mathcal{P}_2} |\hat{\epsilon}(s, x) e^{\nu} ds| \preceq \int_{\mathcal{P}_2} |\hat{\epsilon}(s, x)| |e^{\nu}| dx \]
\[ \preceq CR^{-1/2} (|u_0|_h + |U_0||_h) \int_{\mathcal{P}_2} e^{-\nu} dx \]
\[ \preceq CR^{-1/2} (|u_0|_h + |U_0||_h), \]
which implies that
\begin{equation}
\lim_{R \to \infty} \int_{\mathcal{P}_2} \hat{\epsilon}(s, x) e^{\nu} ds = 0, \quad t > 0,
\end{equation}
In a similar way one proves that
\begin{equation}
\lim_{R \to \infty} \int_{\mathcal{P}_2} \hat{\epsilon}(s, x) e^{\nu} ds = 0, \quad t > 0,
\end{equation}
so we have proved that
\begin{equation}
e(t, x) = \frac{1}{2\pi} \lim_{R \to \infty} \int_{\mathcal{P}_2} \hat{\epsilon}(s, x) e^{\nu} ds
\end{equation}
\[ = \frac{1}{2\pi} \int_{\mathcal{P}_2} \left[ \hat{\epsilon}(s - \alpha, x) e^{-\nu} + \hat{\epsilon}(s + \alpha, x) e^{\nu} \right] e^{-\nu} dx, \]
which is an absolutely convergent integral, since \( \hat{\epsilon}(s, x) \) is bounded in \( \alpha \). We now have to obtain local error bounds for \( \hat{\epsilon}(s, x) \) on the lines \( s = \alpha (1 \pm i), \alpha \approx 0 \) only, if we want to derive local error bounds for \( e(t, x) \).

We return to problems (3.2) and (3.4), but only for \( s = -\alpha (1 + i) \). For simplicity, we confine ourselves to \( s = -\alpha (1 + i) \), since for \( s = \alpha (1 - i) \) the results are obtained in a similar way. Also, we will sometimes omit the argument \( s \).

**Lemma 2.** Let \( s = -\alpha (1 + i) \). Then, for sufficiently small \( h \), we have
\begin{equation}
\|\hat{e}(s)\|_l \preceq C (1 + |h|^{1/2}) \|\hat{u}(s)\|_{l+1} h^{-l-1}, \quad l = 0, 1.
\end{equation}

**Proof.** For \( s = 0 \), (3.19) is standard (see e.g. [16]). For \( s \neq 0 \), we apply the same tricks as we did to prove (3.13). We see from (3.2b) and (3.3) that
\begin{equation}
B(\hat{e}(s), V) + s(\hat{e}(s), V) = 0, \quad V \in M_0^l(\Delta).
\end{equation}
Next, we introduce the elliptic projection \( \hat{U}(s) \) of \( \hat{u}(s) \) defined by
\begin{equation}
B(\hat{u}(s) - \hat{U}(s), V) = 0, \quad V \in M_0^l(\Delta).
\end{equation}
We know (see e.g. [16]) that
\begin{equation}
\|\hat{u}(s) - \hat{U}(s)\|_l \preceq C h^{l+1} \|\hat{u}(s)\|_{l+1}, \quad l = 0, 1.
\end{equation}
If we set \( \delta(s) = \hat{U}(s) - \hat{U}(s) \) and subtract (3.21) and (3.20), we obtain after putting...
\( V = \delta(s) \)

(3.23) 
\[ B(\delta, \delta) + s(\delta, \delta) = s(\delta - \delta_0, \delta). \]

If we fill in \( s = -\alpha(1 + i) \) and square the moduli of both sides of (3.23), we obtain

(3.24) 
\[ |B(\delta, \delta) - \alpha(\delta, \delta)|^2 + |s(\delta, \delta)|^2 = 2\alpha^2 |(\delta - \delta_0, \delta)|^2. \]

From (3.24), we easily get

(3.25) 
\[ \alpha^2 |\delta|_0^2 \leq 2\alpha^2 |\delta|_0^2 |\delta - \delta_0|_0^2; \quad |\delta|_0 \leq \sqrt{2} |\delta - \delta_0|_0 \leq C |\delta|_{L^1} \leq h^{-1}, \]
\[ |s|_0 \leq |\delta - \delta_0|_0 \leq C |\delta|_{L^1} \leq h^{-1}. \]

In a similar way we prove

\[ |B(\delta, \delta)|^2 \leq 2\alpha^2 |\delta|_0^2 |\delta - \delta_2|_0^2 \leq \alpha^2 Ch^{4r+s} |\delta|_0^2, \]

hence

(3.26) 
\[ |\bar{\delta}(s)|^- \leq C\sqrt{\alpha} h^{-1}|\bar{\delta}(s)|_0^{-1}; \quad \|\bar{\delta}(s)\|_1 \leq C(1 + \sqrt{\alpha} h^{-1}) |\bar{\delta}(s)|_0^{-1}, \]
\[ \|s(s)\|_1 \leq |\delta(s)|_1 + \|\delta(s) - \delta_0(s)|_1 \leq C(1 + \sqrt{\alpha} h^{-1}) |\delta(s)|_0^{-1}, \]

which proves the lemma. \( \square \)

We now introduce the Green's function \( \hat{G}(s, x, \cdot) \in H^{1}_0(I) \cap H^{1,1}(0, \cdot) \cap H^{1,1}(x, 1) \) associated to problem (3.2) defined by

(3.27) 
\[ B(v, \hat{G}(s, x, \cdot)) + s(v, \hat{G}(s, x, \cdot)) = v(x), \quad v \in H^{1}_0(I), \quad x \in I. \]

For \( j = 1, \ldots, N-1 \), we define \( \hat{G}_j(s) \in H^{1}_0(I) \cap H^{1,1}(\Delta) \) by

(3.28) 
\[ \hat{G}_j(s) = G(s, x_j, \cdot), \quad \xi \in (0, x_j) \cup (x_j, 1). \]

Application of (3.27) for \( x = x_j \) and \( v = \hat{G}(s) \) gives

(3.29) 
\[ \hat{G}(s, x_j) = B(\hat{G}(s), \hat{G}_j(s)) + s(\hat{G}(s), \hat{G}_j(s)) \]
\[ = B(\hat{G}(s), \hat{G}_j(s) - V) + s(\hat{G}(s), \hat{G}_j(s) - V), \quad V \in M^\ast_0(\Delta), \]

the last equality given by (3.20). We now take a \( V \) such that

(3.30) 
\[ |\hat{G}_j(s) - V|_{L^1} \leq C h^{1-s/2-i} |\hat{G}_j(s)|_{L^1} \leq C h^{1-s/2-i} |\hat{G}_j(s)|_{L^1, \Delta}, \quad i = 0, 1, \ldots, r; \]
\[ \|V\|_{L^1} \leq C \|\hat{G}_j(s)\|_{L^1, \Delta}. \]

It is now clear that, if \( s = -\alpha(1 + i) \),

(3.31) 
\[ |\hat{\delta}(s, x_j)| \leq |B(\hat{\delta}(s, x_j), \hat{G}_j(s) - V)| + |s(\hat{\delta}(s, x_j), \hat{G}_j(s) - V)| \]
\[ \leq C |\hat{\delta}(s)|_0 \|\hat{G}_j(s) - V\|_0 + 2\alpha |\hat{\delta}(s)|_0 \|\hat{G}_j(s) - V\|_0 \]
\[ \leq C h^{-1}(1 + h^{1-s/2-i}) |\hat{\delta}(s)|_0 \|\hat{G}_j(s)-V\|_0 \|\hat{G}_j(s)|_{L^1, \Delta}. \]

Superconvergence seems to be within grasp, but we need yet explicit dependence on \( u_0 \).

From (3.2a), we find that

\[ B(\hat{u}, \hat{u}) + s(\hat{u}, \hat{u}) = (u_0, \hat{u}); \]
\[ (L\hat{u}, L\hat{u}) + sB(\hat{u}, \hat{u}) = B(u_0, \hat{u}); \]
\[ B(L\hat{u}, L\hat{u}) + sL(\hat{u}, L\hat{u}) = (L\hat{u}_0, L\hat{u}). \]
from which we easily obtain that
\[ \| \hat{u}_m \| \leq C \alpha^{-m}, \quad m = 0, 1, 2, \]
if \( s \) is large enough. After this, we obtain by induction
\[ \| \hat{u}_m \| \leq C_{\alpha^{R-m}} \| u_m \|, \quad m = 0, 1, \ldots, r + 1; \]
\[ R(m) = \lfloor m - 3/3 \rfloor. \]
(3.32)

Concerning \( \hat{\Omega}(x, s) \), a direct computation shows that
\[ \| \hat{\Omega} \|_{L_m} \leq C \alpha^{(m-1)/2}, \quad m \geq 0, \]
if \( s \) is large enough. Combining (3.32) and (3.33), we obtain
\[ \| e(t, x) \| \leq C \| u_0 \| \left( \int_{0}^{\tau} e^{-\mu t} \right)^{1/m}, \quad j = 1, \ldots, N - 1; \quad t > 0, \]
(3.34)
where \( F_j(\alpha) \) is of at most polynomial growth in \( \alpha \).

Formula (3.34) is less good than earlier results by Douglas and Dupont [8], because the superconvergence is not uniform on \( J \), but only valid on any internal \( [\tau, \infty) \) with \( \tau > 0 \). However, we will use (3.34) in the next section to establish superconvergence uniformly on \( J \) for another and more practicable variant of the continuous time Galerkin method.

For the moment, we have proved

**Theorem 1.** Let \( u : J \rightarrow H^{s+1}(I) \cap H^1(I) \) be the solution of (1.1), with \( u_0 \in H^{s+1}(I) \cap H^1(I) \) and let \( U : J \rightarrow M_0^s(\Delta) \) be the solution of (2.6), with \( U_0 \) defined by (3.4). Then the error function \( e(t, x) \) has the global bound (2.10) for \( t \geq 0 \) and the local bound
\[ (3.34) \]
for any \( t > 0 \). \( \Box \)

**Remark.** One can improve (3.33) by taking the contour \( |s| = -\alpha - \mu \pm ia, \alpha \geq 0 \) as new integration path, where \( \mu \) is a positive number between \( 0 \) and \( \lambda_1 \). In that case, one obtains the error bound
\[ \| e(t, x) \| \leq C e^{-\alpha t} \left( \int_{0}^{\tau} e^{-\mu t} \right)^{1/m}, \quad t > 0. \]

4. **Quadrature rules.** In practice, one is forced to evaluate \( B(U, V) \) by some quadrature rule (e.g. Newton-Cotes, see [17]). In this section we advocate the use of Lobatto quadrature rule because it has the following advantages:

(1) there is no loss in the order of accuracy, neither global nor pointwise;

(2) a purely explicit initial value problem can be obtained by selecting a proper basis of \( M_0^s(\Delta) \), with preservation of sparseness.

4.1. **Lobatto quadrature.** It is known (cf. e.g. [5]) that if \( f \in H^2(I) \), the integral
\[ \int_{0}^{1} f(\sigma) d\sigma \]
can be approximated by the \((r + 1)\)-point Lobatto quadrature
\[ Q(f) = \sum_{j=0}^{r} w_j f(\sigma_j), \]
(4.1)
where \( \sigma_0 = 0, \sigma_1 = 1 \) and \( \sigma_i \) are distinct points inside \( J \). The weights \( w_i \) are positive. Examples are the trapezoidal rule \( r = 1 \) and Simpson's rule \( r = 2 \). The approximation (4.1) is exact when \( f \in P_{2r-1}(J) \), otherwise the error is of \( O(h^{2r} f(\xi)), \xi \in (0,1) \), or of \( O(||D f||_{\mathcal{B}}) \). Next, we define \( (\alpha, \beta)_h \), by
\[
(\alpha, \beta)_h = \sum_{j=1}^{N} \langle \alpha, \beta \rangle_{x_j}^*;
\]
(4.3)
\[
(\alpha, \beta)^* = h \sum_{i=0}^{\infty} w_i \alpha(x_i) \beta(x_i), \quad j = 1, \ldots, N;
\]
\[
\xi_i = x_{i-1} + h \xi_i; \quad j = 1, \ldots, N; \quad l = 0, \ldots, r.
\]
Note that \( \xi_{i+1} = \xi_i \). Now \( (\alpha, \beta)_h = (\alpha, \beta) \) if \( \alpha \beta \in P_{2r-1}(J) \), \( j = 1, \ldots, N \); otherwise if \( \alpha \beta \in H^{r}(J) \),
\[
(\alpha, \beta)_h - (\alpha, \beta) = O(h^{2r}) \sum_{j=1}^{N} ||\alpha \beta||_{H^r(J)}.
\]
(4.4)
Inequality (4.4) can be proved by direct application of the lemma of Bramble and Hilbert [1, p. 114].

**Lemma 3.** For any \( U \in M_c^r(\Delta) \),
\[
||U||_h = ((U, U)_h)^{1/2}
\]
is a norm equivalent to \( ||U||_h \).

Next, we define the bilinear functional \( B_h : M_c^r(\Delta) \times M_c^r(\Delta) \rightarrow \mathbb{C} \) by
\[
B_h(U, V) = (p U, V)_h + (q U, V)_h, \quad U, V \in M_c^r(\Delta).
\]

**Lemma 4.** For sufficiently small \( h \), the following inequalities hold:
\[
h^{m} ||U||_{H^m} \leq C h^{m} ||U||_{H^{m,\Delta}}, \quad 0 \leq m \leq l \leq r;
\]
(4.7)
\[
||B_h(U, V) - B(U, V)||_{H^m(\Delta)} \leq C h^{m} ||U||_{H^m(\Delta)} ||V||_{H^m(\Delta)};
\]
(4.8)
\[
||U, V)_h - (U, V)||_{H^m(\Delta)} \leq C h^{m} ||U||_{H^m(\Delta)} ||V||_{H^m(\Delta)};
\]
(4.9)
\[
||f, V)_h - (f, V)||_{H^m(\Delta)} \leq C h^{m} ||f||_{L^2(\Delta)} ||V||_{H^m(\Delta)};
\]
(4.10)
\[
U, V \in M_c^r(\Delta), \quad 0 \leq l, \ m \leq r.
\]

**Proof:** Let us represent \( D^m U(x), x \in I \), by
\[
D^m U(x) = \sum_{i=0}^{n-m} a_i L_i^m(x),
\]
(4.11)
where \( L_i^m(x) \) are the orthonormal Legendre polynomials shifted to \( I \). Then
\[
||D^m U||_{L^2(I)}^2 = \left( \sum_{i=0}^{n-m} a_i^2 \right)^{1/2}.
\]

\( D^m U(x) \) is then represented by
\[
D^m U(x) = \sum_{i=0}^{n-m} a_i D^m L_i^m(x) = \sum_{i=0}^{n-m} a_i P_i^{m-m}(x) h^{m-m},
\]
where \( P_i^{m-m} \) are the Jacobi polynomials \( P_i^{m-m}(x) \) shifted to the interval \( I \). We see
that
\[ \|D' U\|_{2}^2 = h^{2(l+1-m)} \sum_{k=1}^{l+m} \| a_k P_{l+k-m}^* \|_{2}^2 \]
\[ = h^{2(l+1-m)} \sum_{k=1}^{l+m} a_k a_k (P_{l+k-m}^*, P_{l+k-m})_{L^2(I)} \]
\[ \leq h^{2(l+1-m)} \sum_{k=1}^{l+m} a_k a_k \sum_{k=0}^{l} \| (P_k^*, P_k^*)_{L^2(I)} \| \]
\[ \leq C h^{2(l+1-m)} \sum_{k=1}^{l+m} a_k a_k \sum_{k=0}^{l} \| P_{l+k-m}^* \|_{L^2(I)}^2 \]
\[ \leq C h^{2(l+1-m)} \|D^m U\|_2^2 \sum_{i=0}^{l} \| P_{l+i}^* \|_{L^2(I)}^2. \]

It remains to be proved that \( \sum_{i=0}^{l} \| P_{l+i}^* \|_{L^2(I)} \) does not depend on \( h \). This is, however clear since for any \( i, P_{l+i}^* \) is represented by
\[ P_{l+i}^* (x) = \sum_{n=0}^{i} c_n L_n^* (x), \]
where \( C_n \) are constants not depending on \( h \), hence
\[ \| P_{l+i}^* \|_{L^2(I)} = \sum_{n=0}^{i} C_n^2, \]
which proves that
\[ h^{l} \|D' U\|_{L^2(I)} \leq C h^{l} \|D^m U\|_{L^2(I)}; \]
\[ h^{l} \|D U\|_{L^{0,\infty}} \leq C h^{l} \|D^m U\|_{L^{0,\infty}}. \]

After this, the further proof of (4.7) is trivial and will be omitted. Concerning the inequalities (4.8)-(4.10), Hemker [13] proved that
\[ |B_k(U, V) - B(U, V)| \leq C h^{2} \|U\|_{L^2(I)} \|V\|_{L^2(I)}; \]
\[ |(U, V)_h - (U, V)| \leq C h^{2} \|U\|_{L^2(I)} \|V\|_{L^2(I)}; \]
\[ |(f, V)_h - (f, V)| \leq C h^{2} \|f\|_{L^{0,\infty}} \|V\|_{L^2(I)}, \]
after which the inequalities are easily proved by applying (4.7). \( \square \)

**Corollary 1.** If \( h \) is sufficiently small, then \( B_h \) is strongly coercive on \( M_0^0 (\Delta) \times M_0^0 (\Delta) \).

**Proof.** From (4.8), we know that
\[ |B(U, U) - B_h(U, U)| \leq C \|U\|_2^2 h^3, \]
hence
\[ B_h(U, U) \geq (C - C h^3) \|U\|_2^2 \geq C \|U\|_2^2, \]
if \( h \) is sufficiently small. \( \square \)

After these technical lemmas, we arrive at:

**Theorem 2.** Let \( p, q \in H^2 (\Delta) \), let \( u \in H^1 (I) \cap H^{r+1} (I) \cap H^2 (\Delta) \) and let \( Y_0 \in M_0^0 (\Delta) \) be an approximation to \( u_0 \), defined by
\[ (Y_0, V)_h = (u_0, V)_h, \quad V \in M_0^0 (\Delta). \]
Then the solution $Y:[0, \infty) \to M_0'(\Delta)$ of
\begin{equation}
B_h(Y, V) + (Y, V)_h = 0, \quad t \geq 0, \quad V \in M_0'(\Delta); \quad Y(0) = Y_0,
\end{equation}
has the following error bounds
\begin{equation}
\begin{split}
\|u - Y\|_h & \leq \|e(t)\|_h + Ch^{-1}\|u\|_{h+1} \sqrt{t}, \quad t \geq 0; \\
\|u - Y(t, x)\| & \leq C(h) t^{\frac{1}{2}}\|u\|_{2, h}, \quad j = 1, \ldots, N - 1, \quad t \geq 0,
\end{split}
\end{equation}
where $\|e(t)\|_h$ is bounded according to Lemma 1.

Note that (4.12) implies $Y(t, \xi, j) = u_0(\xi, j), j = 1, \ldots, N; t = 0, \ldots, r$.

Proof. Let $U, Z: J \to M_0'(\Delta)$ be the solutions of (2.6) with initial functions $U_0$ and $Y_0$, defined by (3.4) and (4.12) respectively and let $\eta$ and $e$ be defined by
\begin{equation}
\begin{split}
e(t, x) & = U(t, x) - Y(t, x); \\
\eta(t, x) & = Z(t, x) - Y(t, x); \quad x \in I, \quad t \in J.
\end{split}
\end{equation}
We note that $u - Z$ satisfies the conditions of Lemma 1.

If we subtract (4.13) from (2.6) and substitute $V = \eta(t, x)$ we get after application of Lemma 4
\begin{equation}
\begin{split}
\left( \frac{\partial \eta}{\partial t}, \eta \right)_h & + B_h(\eta(t, x)) = B_h(Z, \eta) - B(Z, \eta) + (Z, \eta)_h - (Z, \eta) \\
& \leq C h^{-1}\|\eta\|_{h, 2} + \|Z\|_{2, h}. \end{split}
\end{equation}
Let $\Pi u \in M_0'(\Delta)$ be a projection of $u$ defined by
\begin{equation}
(u - \Pi u, V) = 0, \quad V \in M_0'(\Delta).
\end{equation}
Then from Ciarlet and Raviart, we know that
\begin{equation}
\|u - \Pi u\|_{h+1} \leq Ch\|u\|_{h+1}, \quad l = 0, \ldots, r.
\end{equation}
Hence
\begin{equation}
\begin{split}
\|Z\|_{h+1} & \leq \|Z - \Pi u\|_{h+1} + \|u - \Pi u\|_{h+1} + \|u\|_{h+1} \\
& \leq C h^{-1}\|Z - \Pi u\| + Ch\|u\|_{h+1} + \|u\|_{h+1} \\
& \leq C h\|u\|_{h+1} + \|u\|_{h+1} + C\|u\|_{h+1}.
\end{split}
\end{equation}
In a similar way, we prove that
\begin{equation}
\|Z\|_{h+1} \leq C\|u\|_{h+1} \leq C\|u\|_{h+1}
\end{equation}
hence after applying Corollary 1, we get
\begin{equation}
\frac{d}{dt}\|\eta\|_{h+1} + C\|\eta\|_{h+1} \leq C_2\|\eta\|_{h+1} + Ch\|u\|_{h+1}.
\end{equation}
After application of Gronwall's inequality, we get
\begin{equation}
\begin{split}
\frac{d}{dt}\|\eta\|_{h+1} & \leq Ch\|u\|_{h+1}, \quad t \geq 0; \\
\|\eta\|_{h+1} & = 0, \quad t = 0.
\end{split}
\end{equation}
This differential inequality has the solution
\begin{equation}
\|\eta\|_{h+1} \leq Ch\|u\|_{h+1}.
\end{equation}
which implies that
\[ |u - Y|_h \leq |e(t)|_0 + C h^{r+1} \sqrt{t} |e(t)|_{r+3}, \]
from which one easily proves the first part of (4.14). In order to prove superconvergence at the knots, we prove that
\[ |e(s, x)| \leq C(t) h^{2j}, \quad j = 1, \cdots, N - 1. \]
To that end, we introduce the Laplace transform \( \hat{Y}(s, x) \) of \( Y(t, x) \). It is plain that \( \hat{Y}(s) \) is the solution of
\[ B_h(\hat{Y}(s), V) + s(\hat{Y}(s), V) = (Y_\infty, V)_h, \quad V \in M_0(\Delta). \]
As in Theorem 1, one easily proves that (see Fig. 1)
\[ e(s, x) = \frac{1}{2\pi i} \lim_{R \to \infty} \int_{C_R} \hat{e}(s, x) e^{-st} ds, \]
where \( \hat{e}(s, x) = \hat{Y}(s, x) - \hat{U}(s, x) \). To that end, it suffices to show that
\[ |\hat{Y}(-\alpha \pm iR)| \leq C \alpha^{-1/2}, \quad -R \leq \alpha \leq 0, \]
if \( h \) is sufficiently small. If we apply (4.21) for \( s = -\alpha \pm iR \), we get
\[ B_h(\hat{Y}, \hat{Y}) - \alpha(\hat{Y}, \hat{Y})_h \pm iR(\hat{Y}, \hat{Y})_h = (Y_\infty, \hat{Y})_h. \]
One immediately sees that
\[ |\hat{Y}|_h \leq CR^{-1/2}(\hat{Y}_{\infty}, \hat{Y})_h; \quad B_h(\hat{Y}, \hat{Y}) \leq |(Y_\infty, \hat{Y})_h| \]
and hence after application of Lemma 3,
\[ \|\hat{Y}\|_h \leq C R^{-1/2} \|Y_\infty\|_h; \quad B_h(\hat{Y}, \hat{Y}) \leq C(\alpha + R) R^{-2} \|Y_\infty\|_h. \]
After application of Lemma 4 and Poincaré's inequality, we get
\[ |\hat{Y}(s, x)| \leq C \|\hat{Y}\|_h \leq CR^{-1/2} \|Y_\infty\|_h, \]
which proves (4.23).
From now on, we are only interested in the behavior of \( e(s, x) \) on the lines \( \{s = -\alpha (1 \pm i), \alpha \geq 0\} \). If we subtract (4.21) from (3.3), put \( V = \hat{e} \) and apply (3.4) and (4.12), we get
\[ B_h(e, e) - \alpha(1 \pm i)(e, e)_h - (\hat{U}, e)_h (\hat{U}, \hat{e}) = (u_0, e)_h - (u_0, \hat{e}) + B(\hat{U}, \hat{e}) - B_h(\hat{U}, \hat{e}). \]
One easily verifies that, if \( h \) is sufficiently small,
\[ |B_h(e, e) - \alpha(1 \pm i)(e, e)_h| \leq \frac{1}{2} \sqrt{2} B_h(e, e) \equiv C \|e\|. \]
Furthermore, application of Lemma 4 proves that
\[ [(u_0, e)_h - (u_0, \hat{e})] \leq C \|u_0\|_{L_2} \|\hat{e}\|_{h^{-1}}; \]
\[ |B(\hat{U}, \hat{e}) - B_h(\hat{U}, \hat{e})| \leq C \|\hat{U}\|_{h^{-1}} \|\hat{e}\|_{h^{-1}}; \]
\[ |(\hat{U}, \hat{e}) - (\hat{U}, \hat{e})_h| \leq C \|\hat{U}\|_{h^{-1}} \|\hat{e}\|_{h^{-1}}; \]
so we easily get, after application of (3.32),
\[ \|e\|_h \leq C(1 + \alpha) \|u_0\|_{L_2} \|\hat{e}\|_{h^{-1}}; \]
\[ \|e\|_h \leq C(1 + \alpha) \|\hat{U}\|_{h^{-1}} \|\hat{e}\|_{h^{-1}}; \]
\[ \|\hat{e}\|_h \leq C(1 + \alpha) \|u_0\|_{L_2} \|\hat{e}\|_{h^{-1}}; \]
\[ \|\hat{e}\|_h \leq C(1 + \alpha) \|\hat{U}\|_{h^{-1}} \|\hat{e}\|_{h^{-1}}. \]
For the estimation of \( \xi(s, x_i) \), we again use the Green's function:
\[
|\xi(s, x_i)| = |B(\xi, \partial_i) + s(\xi, \partial_i)|
\leq |B(\xi, \partial_i - V)| + |s(\xi, G_i - V)| + |B(\xi, V)| - s(\xi, V)|
\leq C_h ||\partial_i - V||_h + |s(\xi)| ||\partial_i - V||_h + ||(u_0, \partial_i V) - (u_0, V)||_h
+ |B_h(\xi, V)| - |B_h(\xi, V)| - s(\xi, V)|
\leq C_h ||\partial_i - V||_h + |s(\xi)| ||\partial_i - V||_h + C_h ||(u_0, \partial_i V)||_h + C_h ||(u_0, V)||_h
+ C_h ||\partial_i V||_h + C_h ||(u_0, V)||_h
\tag{4.28}
\]
If we take \( V \) such that (3.30) holds, then
\[
||\partial_i - V||_h \leq C_h ||\partial_i||_{h+1, h} \Rightarrow ||V||_{h, h} \leq C ||\partial_i||_{h+1, h}.
\tag{4.29}
\]
Furthermore
\[
||\partial_i - V||_h \leq C_h ||\partial_i||_{h+1} + ||\partial_i||_h
\leq C_h \left( ||\partial_i||_{h+1} + ||\partial_i||_h \right)
\leq C_h ||(u_0, \partial_i V)||_h + C_h ||(u_0, V)||_h
\leq C_h \left( ||(u_0, \partial_i V)||_h + ||(u_0, V)||_h \right)
\leq C_h \left( ||(u_0, \partial_i V)||_h + ||(u_0, V)||_h \right)
\tag{4.30}
\]
So we finally have
\[
\xi(s, x_i) \leq F_i(\alpha) h^3 ||(u_0)||_{h+1, h}
\tag{4.31}
\]
where \( F_i(\alpha) \) is of at most polynomial growth in \( \alpha \). From this, we easily derive
\[
||u - Y||_h(t, x_j) \leq ||e(t, x_j)|| + ||e(t, x_j)||
\leq C_i(t) h^3 ||(u_0)||_{h+1, h}, \quad j = 1, \ldots, N - 1, \quad t > 0.
\]
Again, \( C_i(t) \) may tend to infinity as \( t \to 0 \). However, since
\[
||u - Y||_h(0, x_j) = 0, \quad j = 1, \ldots, N - 1,
\]
the error bound can be extended to the closed interval \( J \). Hence, we have
\[
||u - Y||_h(t, x_j) \leq C_i(t) h^3 ||(u_0)||_{h+1, h}, \quad t > 0.
\tag{4.32}
\]
where \( C_i \) is bounded on \( J \), \( C_i(0) = 0 \) and where \( C_i \) is of \( O(t^{-1}) \), as \( t \to \infty \). This completes the proof. \( \square \)

**Remark.** By exactly the same methods, but now applied to the contour \( \{ s \mid s = -\mu = \alpha \} \), this result can be improved to
\[
||u - Y||_h(t, x_j) \leq C_i(t, \mu) e^{-\mu t} h^3 ||(u_0)||_{h+1, h},
\]
provided that
\[
\mu = \inf_{v \varepsilon M_h} \frac{B(V, V)}{|V|^{2v}} - \Lambda^*,
\]
where \( |\Lambda^* - \lambda_i| \) is known to be of \( O(h^2) \).

**4.2. Preservation of sparseness: purely explicit initial value problems.** Another well-known feature of the continuous time Galerkin method is that it results in a vector initial value problem of sparse structure.
We define the points \( z_0, z_1, \cdots, z_N \) by
\[
\begin{align*}
z_j &= x_{j}, & j &= 0, \cdots, N; \\
z_{j+1} &= z_{j+1} + \Delta z, & j &= 0, \cdots, N - 1; \quad l = 1, \cdots, r - 1.
\end{align*}
\] (4.33)

Then one can define a basis \( \{ \phi_i \}_{i=1}^{M} \) for \( M_c(\Delta) \) by
\[
\phi_i(z_j) = \delta_{ij}, \quad 1 \leq i, j \leq M.
\] (4.34)

If one represents \( Y(t, x) \) by
\[
Y(t, x) = \sum_{i=1}^{M} a_i(t) \phi_i(x),
\]
then \( a \) satisfies the initial value problem:
\[
\begin{align*}
K_1 \frac{da}{dt} + K_2 a &= 0, & t &\geq 0; \\
a_i(0) &= u_0(z_i), & r &= 0; \quad i = 1, \cdots, M.
\end{align*}
\] (4.35)

As is easily checked, the \( M \times M \) matrix \( K_2 \) is \((2r+1)\)-diagonal of the structure
\[
\begin{pmatrix}
\ddots & \ddots & & & \\
& \ddots & \ddots & & \\
& & \ddots & \ddots & \\
& & & \ddots & \ddots \\
& & & & \ddots
\end{pmatrix}
\] (4.36)

However, what is more important,
\[
\begin{align*}
(\phi_i, \phi_j) &= \delta_{ij} \| \phi_i \|_2^2,
\end{align*}
\] (4.37)

which means that \( K_1 \) is a diagonal matrix. One sees that by use of Lobatto quadrature the matrix \( (\phi_i, \phi_j) \) has been lumped (physicists' term) to a diagonal matrix. So while for the solution of (2.6) one would have to integrate the implicit differential equation
\[
\begin{align*}
K_1 \frac{db}{dt} + K_2 b &= 0,
\end{align*}
\] (4.38)

where \( K_2 \) and \( K_1 \) are both of the structure (4.36), problem (4.35) is of purely explicit form, which is another advantage of Lobatto quadrature.
5. Numerical example. In order to illustrate the superconvergence at the knots when Lobatto quadrature is applied, we integrated the following simple problem

\[
\frac{\partial^2 u}{\partial t^2} = 2 \frac{\partial^3 u}{\partial x^3} + (x^{10} + 180x^2 - x) e^{-t}, \quad t \geq 0;
\]

\[
u = 0, \quad x = 0, 1; \quad u = x - x^{10}, \quad t = 0.
\]

The exact solution is \((x - x^{10}) e^{-t}\). It was partitioned into \(N\) segments of equal length for \(N = 4, 8, 16\) respectively. For \(r = 1, 2, 3\) this problem was integrated from 0 to 1 by an adaptive Runge–Kutta method. Below, we list the errors in the points 0.25, 0.50 and 0.75 for \(r = 1, 2, 3\) and \(N = 4, 8, 16\). One easily verifies that the errors decrease by about \(2^{-2r}\) when \(N\) is doubled, which confirms the superconvergence at the knots.

\[
\begin{array}{cccc}
\hline
r = 1 & 4 & 8 & 16 \\
0.25 & 3.90(-2) & 1.11(-2) & 2.87(-3) \\
0.50 & 7.65(-2) & 2.17(-2) & 5.61(-3) \\
0.75 & 9.96(-2) & 2.80(-2) & 7.26(-3) \\
\hline
r = 2 & 4 & 8 & 16 \\
0.25 & 1.87(-3) & 1.25(-4) & 7.97(-6) \\
0.50 & 3.61(-3) & 2.40(-4) & 1.53(-5) \\
0.75 & 4.25(-3) & 2.80(-4) & 1.71(-5) \\
\hline
r = 3 & 4 & 8 & 16 \\
0.25 & 1.15(-5) & 1.83(-7) & 2.78(-9) \\
0.50 & 2.04(-5) & 3.23(-7) & 4.91(-9) \\
0.75 & 2.01(-5) & 3.17(-7) & 4.79(-9) \\
\hline
\end{array}
\]

REFERENCES


A Note on $C^0$ Galerkin Methods for Two-Point Boundary Problems

Miente Bakker
Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam

Summary. As is known [4], the $C^0$ Galerkin solution of a two-point boundary problem using piecewise polynomial functions, has $O(h^k)$ convergence at the knots, where $k$ is the degree of the finite element space. Also, it can be proved [5] that at specific interior points, the Gauss-Legendre points the gradient has $O(h^{k+1})$ convergence, instead of $O(h^k)$. In this note, it is proved that on any segment there are $k-1$ interior points where the Galerkin solution is of $O(h^{k+1})$, one order better than the global order of convergence. These points are the Lobatto points.

Subject Classifications: AMS (MOS) 65 N 30; CR: 5.17.

1. Introduction

We consider the two-point boundary problem

\[ Lu = -(p(x)u')' + q(x)u = f(x), \quad x \in [0, 1] = I; \]
\[ u(0) = u(1) = 0. \]  \hspace{1cm} (1)

We suppose that $p$, $q$ and $f$ are such that (1) has a unique and sufficiently smooth solution.

Let, for a constant integer $N$, $0 = x_0 < x_1 < \ldots < x_N = 1$ be a partition of $I$ with

\[ h = N^{-1}; \quad x_j = jh; \quad I_j = [x_{j-1}, x_j] \]

and let for a constant integer $k \geq 2$ and for any interval $E \subseteq I$, $P(E)$ be the class of polynomials of degree at most $k$ restricted to $E$.

We define for $m \geq 0$ and $s \geq 1$

0029-599X/82/0038/0447/S01.40
\[ W^m(E) = \{ u \in H^m(E) : u|_{E_j} = 0, \ldots, m_j \} \]
\[ H^m(E) = W^m(E) \]
\[ H^1_0(E) = \{ u \in H^1(E) : u(\partial E) = 0 \} \]
\[ M^d_{k, l}(\Delta) = \{ u \in H^1_0(I), u \in L^2(I), j = 1, \ldots, N \} \]
\[
\begin{align*}
|u|_{m, k} &= \left( \sum_{j=0}^{N} |D^j u|_{m_j, k_j}^2 \right)^{1/2} \\
|u|_{m, k^*} &= \left( \sum_{j=0}^{N} (D^j u, D^j u)_{m_j, k_j} \right)^{1/2}
\end{align*}
\]

where \( D^j \) denotes \( \cdot \, d^j \). If \( E = I \), we write \( \langle x, \beta \rangle \) instead of \( \langle x, \beta \rangle_{E_j} \), \( j \), and \( |\cdot|_{m, k} \) instead of \( |\cdot|_{m, k, j} \).

Let \( u \in M^d_{k, l}(\Delta) \) be the unique solution of
\[ B(U, V) = (f, V), \quad V \in M^d_{k, l}(\Delta), \tag{3} \]
where \( B : H^1_0(I) \times H^1_0(I) \to \mathbb{R} \) is defined by
\[ B(u, v) = (\beta u, v) + (\alpha u, v), \quad u, v \in H^1_0(I). \tag{4} \]
We assume that \( B \) is strongly coercive, i.e., there exists a \( C > 0 \) such that
\[ B(v, v) \geq C |v|_1^2, \quad v \in H^1_0(I). \tag{5} \]
In the sequel, \( C, C_1 \) are generic positive constants not necessarily the same.

**Lemma 1.** Let \( u \in H^1_0(I) \cap H^{r+1}(I) \) be the solution of (1) and let \( U \in M^d_{k, l}(\Delta) \) be the solution of (3). Then the error function \( e(x) = u(x) - U(x) \) has the bounds
\[
\begin{align*}
|e|_{l, k} &\leq C h^{k+1} \|u\|_{l, k}, \quad l = 0, 1; \\
|e|_{l, k} &\leq C h^k \|u\|_{l, k}, \quad j = 1, \ldots, N - 1; \\
|e|_{l, k} &\leq C h^{k+1} \|u\|_{l, k}, \quad l = 0, \ldots, k.
\end{align*}
\]

**Proof.** See [6], [4] and [7]. \( \square \)

In the next \( \S \), we prove that the local order of convergence improves slightly at specific points interior to \( I_j \), if \( u \) satisfies stricter smoothness requirements on the interior of \( I_j \).

### 2. Order of Convergence at Lobatto Points

On the segment \([-1, +1] \), we define the Lobatto points \( \sigma_0, \ldots, \sigma_k \) by
\[
(1 - \sigma^2)^{d} \frac{d}{d\sigma} P_l(\sigma) - 0, \quad l = 0, \ldots, k \tag{7}
\]
where \( P_l(\sigma) \) is the \( l \)-th degree Legendre polynomial. Associated to this polynomial is the quadrature formula (see [1, formula 25.4.32]).
\[
\int f(\sigma) d\sigma = \sum_{l=0}^{k-1} w_{l} f(\sigma_{l}) \frac{(k+1)k/2^{k-1}[(k-1)!]^{2}}{(2k+1)[(2k)!]} \int f^{(2k)}(\sigma), \quad \sigma \in \{-1, +1\}
\]

\[
w_{l} = \frac{2}{k(k+1)[R(\sigma_{l})]}^{2}, \quad l = 0, \ldots, k.
\]

From (7) and (8), we define

\[
\zeta_{j} = x_{j-1} + \frac{h}{2} [1 + \sigma_{j}], \quad j = 0, \ldots, k;
\]

\[
(z, \beta)^{T} = \frac{h}{2} \sum_{l=0}^{k} w_{l} z(\xi_{l}) \beta(\xi_{l}), \quad z, \beta \in W^{2k-1}(I), \quad j = 1, \ldots, N;
\]

\[
(z, \beta)_{h} = \sum_{j=1}^{N} (z, \beta)^{T}.
\]

We return to problems (1) and (3). It is known that

\[
B(e, V) = 0, \quad V \in M^{k}_{2}(d).
\]

For any \(I_{j}\), we define

\[
M^{k}_{2}(I_{j}) = \{ V \in M^{k}_{2}(d), \text{supp}(V) = I_{j} \}.
\]

We temporarily drop the subscript \(j\) from the numbers \(\zeta_{ij}\). We define a natural basis \(\{\phi_{i}\}_{i=1}^{2k-1}\) for \(M^{k}_{2}(I_{j})\) by

\[
\phi_{i}(\xi_{j}) = \delta_{ij}, \quad 1 \leq i \leq k - 1.
\]

where \(\delta_{ij}\) is the Kronecker symbol. If we elaborate (10) for \(V = \phi_{i}, i = 1, \ldots, k - 1\), we get

\[
(e, L\phi_{i}) = \int_{I_{j}} [\eta(x) e(x) \phi_{i}(x)] \xi_{i}^{2k-1}, \quad i = 1, \ldots, k - 1.
\]

Approximation of \((e, L\phi_{i})\) by Lobatto quadrature yields

\[
\sum_{i=1}^{k-1} w_{i} L\phi_{i}(\xi_{j}) e(\xi_{i}) = 2h^{-1} \int_{I_{j}} [\eta(x) e(x) \phi_{i}(x)] d\xi_{i} + w_{0} e(\xi_{0}) L\phi_{0}(\xi_{0}) - w_{k} e(\xi_{k}) L\phi_{k}(\xi_{k}) + C h^{2k} D^{2k} (e L\phi_{i}) (\xi)_{I_{j}}, \quad i = 1, \ldots, k - 1.
\]

This is a linear system for \((w_{i} L\phi_{i}(\xi_{j}))\) and \((e(\xi_{i}))\). We have to prove the non-singularity of \((w_{i} L\phi_{i}(\xi_{j}))\) and compute the order of the solution. We know that

\[
h B(\phi_{i}, \phi_{j}) = h (L\phi_{i}, \phi_{j})
\]

\[
= h^{2} \sum_{i=1}^{k-1} w_{i} L\phi_{i}(\xi_{j}) \phi_{i}(\xi_{j}) + C h^{2k+1} D^{2k} (L\phi_{i}(\xi) \phi_{i}(\xi)), \xi \in I_{j}
\]

\[
= h^{2} w_{i} L\phi_{i}(\xi_{j}) + C h^{2k+1} D^{2k} (L\phi_{i}(\xi) \phi_{i}(\xi)), \xi \in I_{j}.
\]
Hence we have

$$|hB(\phi_i, \phi_i) - h^2 w_i L\phi_i(\xi_i)| \leq C h^2. \quad (15)$$

This means that $M_i = (h^2 w_i L\phi_i(\xi_i))$ is nearly equal to a symmetric positive definite matrix whose entries and positive eigenvalues are of $O(1)$ and consequently has an inverse with the same properties. If we represent $(hB(\phi_i, \phi_i))$ by $M_i$, we find that

$$M_i = M_1 + h^2 M_2 = M_1 (I + h^2 M_1^{-1} M_2),$$

where all $M_i$ have entries of $O(1)$. Since the spectral radius of the perturbation matrix is $O(h^2)$, it is evident by power series expansion that

$$M_i^{-1} = M_1^{-1} + h^2 M_2,$$

where the entries of $M_i$ are of $O(1)$. This proves that $M_i^{-1}$ has entries of $O(1)$ and so we have that $(w_i L\phi_i(\xi_i))^{-1}$ has entries of $O(h^2)$.

We turn to the second part of our problem. The first three terms of the right hand side of (14) are of $O(h^{2k-2} ||w||_{L^\infty})$. For the last term, we prove that

$$||D^{2k}(\varepsilon L\phi_i)||_{L^\infty} \leq C \varepsilon_{w_{2k-1}}||L\phi_i||_{W^{2k-1}}, \quad (16)$$

From [3], it can be proved that

$$||D^l \varepsilon||_{L^\infty} \leq C h^{k+l-1} ||w||_{k+l-1}, \quad l \leq k; \quad ||D^l \varepsilon||_{L^\infty} \leq ||D^l ||_{L^\infty}, \quad l > k. \quad (17)$$

Furthermore,

$$||L\phi_i||_{W^{2k-1}} \leq C h^{-k}, \quad (18)$$

hence we summarily have

$$\sum_{i=1}^{k-1} w_i L\phi_i(\xi_i) \varepsilon(\xi_i) \leq C h^k [||w||_{k+1}, h^{k-2} + ||w||_{W^{2k-1}},]. \quad (19)$$

This was the last step in the proof of

**Theorem 1.** Let $w \in H^k(\Omega) \cap W^{k+1}(\Omega) \cap \bigcap_{j=1}^{N} W^{2k-1}(f_j)$ be the solution of (1) and let $U \in M_2^2(\Delta)$ be the solution of (3). Then the error function has the local error bound.

$$||\varepsilon_i|| \leq C h^{k+2} [||w||_{k+1}, h^{k-2} + ||w||_{W^{2k-1}}], \quad j=1, \ldots, N; \quad l=1, \ldots, k-1. \quad (20)$$
3. Lobatto Quadrature

Usually, \( B(\cdot, \cdot) \) and \( (\cdot, \cdot) \) are to be evaluated by numerical quadrature. We will show that Lobatto quadrature leaves the order of convergence at the Lobatto points invariant.

We define

\[
B_h(\alpha, \beta) = (p\alpha', \beta')_h + (q\alpha, \beta)_h; \quad \alpha, \beta \in \bigoplus_{j=1}^{N} W^{2k-1}(I_j),
\]

where \( (\cdot, \cdot)_h \) is defined by (9).

**Lemma 2.** Let \( Y \in M_0^h(I) \) be the solution of

\[
B_h(Y, V) = (f, V)_h, \quad V \in M_0^h(I)
\]

and let \( u \in H^1(I) \cap H^{k+1}(I) \cap \bigoplus_{j=1}^{N} W^{2k-1}(I_j) \) be the solution of (1). Then the error function \( \eta = u - Y \) has the bounds

\[
|\eta(x)| \leq C h^{2k-1} \|f\|_{2k-1}, \quad j = 1, \ldots, N - 1,
\]

if \( h \) is small enough, with

\[
\|f\|_{2k-1} = \left( \sum_{j=1}^{N} \|f\|_{W^{2k-1}(I_j)}^2 \right)^{1/2}.
\]

**Proof.** See [4]. □

We now consider \( \varepsilon(x) = U(x) - Y(x) \), where \( U \) is the solution of (3). From (3) and (22), we obtain for every \( I_j \)

\[
|B(\varepsilon, V)| \leq \|f, V\|_h - (f, V)_h + |B_h(Y, V) - B(Y, V)|
\]

\[
\leq C h^{2k+1} \|V\|_{W^{2k+1}(I_j)} + \|f\|_{W^{2k+1}(I_j)} + \|Y\|_{W^{2k+1}(I_j)}, \quad V \in M_0^h(I_j).
\]

If we take for \( V \) any of the basis functions \( \phi_i \) of \( M_0^h(I_j) \), as defined by (12), we have

\[
|B(\varepsilon, \phi_i)| \leq C h^{2k+1} \|f\|_{W^{2k+1}(I_j)} + \|Y\|_{W^{2k+1}(I_j)}, \quad i = 1, \ldots, k - 1.
\]

Since

\[
\sum_{i=1}^{k-1} w_i \varepsilon(\zeta_i) L \phi_i(\zeta_i) - 2 h^{-1} B(\varepsilon, \phi_i)
\]

\[
- w_i \varepsilon(\zeta_i) L \phi_i(\zeta_i) - w_i \varepsilon(\zeta_i) L \phi_i(\zeta_i)
\]

\[
- \frac{2}{h} \left( \frac{p(x) \phi_i(x)}{p(x)} \right)^{\frac{1}{2}} + C h^{2k} D^{2k}(eL \phi_i)(\zeta \in I_j)
\]

and

\[
\|D^{2k}(eL \phi_i)\|_{L^\infty(I_j)} \leq C \|e\|_{W^{2k-1}(I_j)} \|\phi_i\|_{W^{2k-1}(I_j)}
\]

\[
\leq C h^{-2k} \|\phi_i\|_{L^\infty(I_j)} \leq C h^{-2k+1} \|f\|_{2k-1},
\]

where \( L \phi_i(\zeta_i) = \frac{\phi_i(\zeta_i)}{p(x)} \).
we have
\[
\sum_{j=1}^{k-1} w_i \varphi_j(z_j) L \phi_i(z_j) \leq C_i h^k [ \| f \|_{H^{k+1}(\Omega_j)} + \| g \|_{H^{k+1}(\Omega_j)} ] + C_j h^{k-1} \| f \|_{2k, \Delta}.
\]
(28)

The nonsingularity of \((w_i L \phi_i(z_j))\) has already been proved, its inverse is of \(O(h^2)\), hence we have
\[
\| e(z_j) \| \leq C h^{k+2} [ \| f \|_{H^{k+1}(\Omega_j)} + \| g \|_{H^{k+1}(\Omega_j)} ] + C h^{k+3} \| f \|_{2k, \Delta}.
\]
(29)

Since (see [3]),
\[
\| Y \|_{H^s(\Omega_j)} \leq \| \eta \|_{H^s(\Omega_j)} + \| u \|_{H^s(\Omega_j)} \leq C h \| u \|_{k+1} + \| u \|_{H^s(\Omega_j)} \leq C \| u \|_{k+1}.
\]
(30)

we can prove by combination of (20), (29) and (30)

**Theorem 2.** Let \( u \in H^k(I) \cap H^{k+1}(I) \cap \bigcap_{j=1}^{N} H^{2k+1}(I_j) \)

be the solution of (1) and let \( Y \in H^k(I) \)

be the solution of (22). Then the error function \( \eta \) has the bounds
\[
\| \eta \|_{H^s(\Omega_j)} \leq C h^{k+2} [ \| f \|_{H^{k+1}(\Omega_j)} + \| u \|_{k+1} ] + C h^{k+3} \| f \|_{2k, \Delta};
\]
\[
j = 1, \ldots, N; \quad l = 1, \ldots, k-1.
\]

**4. Conclusions**

We have found a weaker form of superconvergence at other points than the knots. The findings of this paper stress the important part that Lobatto points play in the \( C^0 \) Galerkin method for two-point boundary problems. This is especially true for \( k=2 \), since in that case the error is of \( O(h^3) \) at all Lobatto points.

The results of this paper can be easily applied to the case of two-point initial boundary problems (see [2]) and probably to other cases, such as nonlinear boundary problems.

**References**

C^s Galerkin Methods for Two-Point Boundary Problems


Received December 12, 1979; Revised April 12, 1981
GALERKIN METHODS FOR EVEN-ORDER PARABOLIC EQUATIONS IN ONE SPACE VARIABLE*

MIENTE BAKKER

Abstract. For parabolic equations in one space variable with a strongly coercive self-adjoint 2mth order spatial operator, a 4th degree Fdaro-Galerkin method is developed which has local convergence of order \( 2k + 1 - m \) at the knots for the first \( m - 1 \) spatial derivatives and, if \( k \geq 2m \), convergence of order \( k + 2 \) at specific interior nodal points. These nodal points are the zeros of the Jacobi polynomial \( P_n^{(a,b)}(x) \) \((n = k + 1 - 2m)\) shifted to the segments of the partition. All these convergence properties are preserved if suitable quadrature rules are used.

1. Introduction. We consider the 2mth order initial boundary problem

\[
\frac{\partial u}{\partial t}(t,x) + Lu(t,x) = 0, \quad x \in [-1,1], \quad t \in [0,\infty) = J,
\]

\[
Lu = \sum_{j=0}^{m} (-1)^j \frac{\partial}{\partial x^j} \left[ p_j(x) \frac{\partial^j u}{\partial x^j} \right],
\]

\[
\frac{\partial^u}{\partial x^j} = 0, \quad x = \pm 1, \quad l = 0, \ldots, m - 1, \quad t \in J,
\]

\[
u(0,x) = u_0(x).
\]

We suppose that \( p_0, \ldots, p_m \) and \( u_0 \) are such that \( u(t) \) is sufficiently smooth for every \( t \in J \).

1.1. Notation. For any interval \( E \subset I \), we define the Sobolev spaces \( W^l(E) \) and \( H^l(E) \), \( l \geq 0 \), and their norms by

\[
W^l(E) = \{ v | D^j v \in L^2(E), j = 0, \ldots, l \},
\]

\[
H^l(E) = \{ v | D^j v \in L^2(E), j = 0, \ldots, l \},
\]

\[
\|v\|_{W^l(E)} = \max_{l = 0, \ldots, m} \|D^l v\|_{L^2(E)},
\]

\[
\|v\|_{H^l(E)} = \left( \sum_{j=0}^{l} \|D^j v\|_{L^2(E)} \right)^{1/2},
\]

where \( D^l \) denotes \( d^l/dx^l \) or \( \partial^l/\partial x^l \) and the complex-valued inner product \( (\cdot, \cdot)_E \) is defined by

\[
(\alpha, \beta)_E = \int_E \overline{\alpha(x)} \beta(x) \, dx, \quad \alpha, \beta \in L^2(E).
\]

For convenience, since we use them frequently, we make the following replacements:

\[
|\alpha|_2 = \|\alpha\|_{H^{1/2}}, \quad (\alpha, \beta) = (\alpha, \beta)_E.
\]

Furthermore, we define \( H^\infty_0(I) \) and the bilinear functional \( B : H^\infty_0(I) \times H^\infty_0(I) \to \mathbb{C} \) by

\[
H^\infty_0(I) = \{ v \mid v \in H^\infty(I), \nabla^l v(x) = 0, l = 0, \ldots, m - 1 \},
\]

\[
B(u, v) = (Lu, v) = (u, Lv) = \sum_{l=0}^{m} (p_l D^l u, D^l v), \quad u, v \in H^\infty_0(I).
\]

* Received by the editors April 22, 1980, and in revised form April 10, 1981.
* Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands.
We assume that $p_0, \ldots, p_m$ are such that $B$ is strongly coercive; i.e., that there exist positive constants $C_1$ and $C_2$ depending on $p_0, \ldots, p_m$, only, such that
\begin{equation}
|B(u, v)| \leq C_1|u||v|, \quad u, v \in H^m_0(I),
\end{equation}
\begin{equation}
B(u, v) \geq C_2|u||v|, \quad v \in H^m_0(I).
\end{equation}
Note that this implies that $p_m(x) > 0, x \in I$.

In the sequel, $C, C_1, C_2$, etc. will be positive generic constants, not necessarily the same.

1.2. The Fused–Galerkin method. Let $N \geq 2$ be a constant integer and define the partition $\Delta = \{x_i\}_{i=0}^N$ of $I$ by
\begin{equation}
h = \frac{2}{N},
\end{equation}
\begin{equation}
x_i = -1 + ih, \quad j = 0, \ldots, N, \quad I_j = [x_{j-1}, x_j], \quad j = 1, \ldots, N.
\end{equation}
Let $k \geq 2m - 1$ be a constant integer. Then we define the finite element space $S(\Delta) = H^m_0(I)$ by
\begin{equation}
S(\Delta) = \{ V \in H^m_0(I) : V \in P_k(I_j), j = 1, \ldots, N \},
\end{equation}
where for any $l \geq 0 P_l(E)$ denotes the class of polynomials of degree at most $l$ defined on the interval $E$.

In the sequel, we will use the following constant integers associated to $k, m$ and $N$:
\begin{equation}
r = k + 1 - m, \quad n = k + 1 - 2m, \quad M = rN - m.
\end{equation}
In (1.9) $n$ is the number of interior nodal points of $S(\Delta)$ on $I$, and $M$ is the dimension of $S(\Delta)$.

In connection with $\Delta$, we define the partition spaces $W^l(\Delta)$ and $H^l(\Delta)$ together with their norms by
\begin{equation}
W^l(\Delta) = \{ v \in W^l(I_j) : j = 1, \ldots, N \},
\end{equation}
\begin{equation}
\|v\|_{W^l(\Delta)} = \max_{j=1, \ldots, N} \|v\|_{W^l(I_j)},
\end{equation}
\begin{equation}
H^l(\Delta) = \{ v \in H^l(I_j) : j = 1, \ldots, N \},
\end{equation}
\begin{equation}
\|v\|_{H^l(\Delta)} = \left[ \sum_{j=1}^N \|v\|_{H^l(I_j)}^2 \right]^{1/2}.
\end{equation}

After these preliminary definitions, we can define a finite element solution of (1.1). Let $U : \tilde{S}(\Delta) \rightarrow S(\Delta)$ be the solution of the initial boundary problem
\begin{equation}
\frac{\partial U}{\partial t} + B(U, V) = 0, \quad V \in S(\Delta), \quad t \geq 0,
\end{equation}
\begin{equation}
U(0, x) = U_0(x),
\end{equation}
where $U_0 \in S(\Delta)$ is an approximation of $u_0$ satisfying
\begin{equation}
\|u_0 - U_0\| \leq Ck^{m+1/2}\|u_0\|_{m+1}, \quad l = 0, \ldots, m.
\end{equation}
**Lemma 1.** Let \( u : J \to H_0^1(I) \cap H^{s-1}(I) \) be the solution of (1.1) and let \( U : J \to S(\Delta) \) be the solution of (1.11) with condition (1.12). Then \( e(t) = u(t) - U(t) \), has the \( L^2 \) error bound

\[
\|e(t)\|_0 \leq C \lambda_1^{1/2 - s} \left\{ \|u(t)\|_{H^1} + \|u(t)\|_{H^{m-1}} + \int_0^t e^{\lambda_1 \tau} \|Lu(t)\|_{H^{m-1}} \, d\tau \right\},
\]

where \( \lambda_1 \) is the smallest eigenvalue of \( L \).

**Proof.** See [11]. \( \square \)

**1.3. Summary of results in this paper.** In § 2 the occurrence of superconvergence at the knots is investigated. It appears that this depends crucially on a proper choice of \( U_0 \). A surprisingly simple choice of \( U_0 \) is made with the only additional requirement that \( u(t) \in H_0^m(I) \cap H^{s-1}(I) \cap W_0^{m-1}(\Delta) \), \( t \in J \). In that case \( D_0^e(t, x) \) is \( O(h^m) \), \( t \neq 0 \). Furthermore, if \( m \geq 1 \), there are on each \( I, \) specific interior points, where \( e(t) \) is of \( O(h^{m-1}) \), one order better than the optimal order of convergence.

In § 3, it is shown that all the results from § 2 remain valid if \( B(\cdot, \cdot) \) is approximated by a proper quadrature rule.

**2. Superconvergence phenomena.** For \( m = 1 \) and \( k \geq 2 \), J. Douglas, Jr. et al. [7], [8], [9], [10] have proved that the order of convergence at the knots is \( 2k \), while the optimal order is \( k + 1 \). We generalize their results for \( m > 1 \). Also, we establish a minor superconvergence at interior points. For these purposes, the Laplace transforms of \( u(t) \) and \( U(t) \) are used, because they transform initial boundary problems into boundary problems which are simpler to handle.

**2.1. The Laplace transform.** Let \( V \) be a class of functions defined on \( I \). Then for any continuous mapping \( v : J \to V \), we define the Laplace transform \( \mathcal{L} \) by

\[
\mathcal{L}v(s, x) = \hat{v}(s, x) = \int_0^\infty e^{-st} v(t, x) \, dt,
\]

where \( s \) lies in the convergence half-plane of \( v(t) \).

For the general properties of \( \mathcal{L} \) and for the convergence criteria for (2.1), we refer to [6]. If we apply \( \mathcal{L} \) to the problems (1.1) and (1.11), we get for \( \hat{u} \) the two-point boundary problem (in classical and weak Galerkin form)

\[
\begin{align*}
(2.2a) & \quad \hat{L} \hat{u} + s \hat{u} = \hat{u}_0, \quad x \in I, \\
(2.2b) & \quad B(\hat{u}, v) + s(\hat{u}, v) = (u_0, v), \quad v \in H_0^m(I),
\end{align*}
\]

and for \( \hat{U} \) the weak Galerkin form

\[
(2.3) \quad B(\hat{U}, V) + s(\hat{U}, V) = (U_0, V), \quad V \in S(\Delta).
\]

Note that (2.3) is not the standard finite element solution of (2.2). Since the dependence on \( s \) appears from the roof-sign, we usually omit the argument \( s \).

We first formulate a technical lemma which we will use a couple of times.

**Lemma 2.** Let \( x_1 \) and \( x_2 \) be nonnegative numbers, let \( \mu, \gamma \) and \( D \) be positive parameters, let \( s \) be a complex number and let the following inequalities hold:

\[
\begin{align*}
|x_1 + ix_2| & \leq D \sqrt{x_1}, \quad x_1 \geq x_2, \quad s = -\alpha + i\beta, \\
\mu & \leq \alpha \leq |\beta| + \mu, \quad 0 \leq \mu < \gamma.
\end{align*}
\]
Then $x_1$ and $x_2$ have the bounds

$$x_1 \equiv \begin{cases} \frac{\gamma D^2}{(\gamma-\alpha)^2 + \beta^2}, & \text{if } \alpha^2 + \beta^2 \geq \gamma^2, \\ \frac{D^2}{2\beta^2}[\alpha + \sqrt{\alpha^2 + \beta^2}], & \text{if } \alpha^2 + \beta^2 \geq \gamma^2, \end{cases}$$

$$(2.5)$$

$$x_2 \equiv \begin{cases} \frac{D^2}{(\gamma-\alpha)^2 + \beta^2}, & \text{if } \alpha \leq \gamma, \\ \frac{D^2}{\beta^2}, & \text{if } \alpha \geq \gamma. \end{cases}$$

Proof. We substitute

$$x_1 = y_1 + \alpha y_2, \quad x_2 = y_2.$$  

Then, for $y_1$ and $y_2$, we have the inequalities

$$y_1^2 + \beta^2 y_2^2 \leq D^2 y_2, \quad y_1 \geq (\gamma-\alpha)y_2, \quad y_2 \geq 0,$$  

$$(2.7)$$

$$\mu \geq \alpha \geq |\beta| + \mu, \quad \mu < \gamma,$$

so $x_1$ and $x_2$ are linear functions of $y_1$ and $y_2$ with constraints (2.7). Elaboration for all possible values of $\beta$ delivers (2.5). □

We turn to the problems (2.2) and (2.3). Let $\mu$ be a positive number with $\mu < \lambda_1$ and define $P_1, P_2, \cdots, P_5$ in the complex plane (see Fig. 1) by

$$(2.8) \quad P_1 = -\mu, \quad P_{2,5} = -\mu \pm iR, \quad P_{3,4} = -(\mu + R) \pm iR, \quad R > 0.$$  

By $P_1, \cdots, P_n$ we denote the broken straight line starting in $P_1$, going to $P_2$, etc. and ending in $P_n$.

---

![Diagram](https://via.placeholder.com/150)  

**Fig. 1**
Lemma 3. Let \( \psi(t) = u(t) - U(t) \) and \( \hat{\psi} = \hat{u} - \hat{U} \), where \( u(t) \), \( U(t) \), \( \hat{u} \) and \( \hat{U} \) are the solutions of (1.1), (1.11), (2.2) and (2.3), respectively. Then for \( t > 0 \) and \( h \) sufficiently small, we have

\[
D^l \psi(t, x) = \lim_{R \to \infty} \frac{1}{2\pi i} \int_{\gamma_R} D^l \hat{\psi}(s, x) \exp(st) \, ds
\]

\[
eq \frac{e^{-\mu t}}{\pi} \int_0^\infty e^{-\mu s} \text{Im} \left[ (1 - i) e^{-is} D^l \hat{\psi}(-\alpha - \mu - i\alpha, x) \right] \, da,
\]

\[l = 0, \ldots, m - 1.
\]

Proof. It is known [11] that

\[
\hat{u}(s, x) = \sum_{i=1}^N (U_0, \Phi_i) \frac{\Phi_i(x)}{s + \lambda_i}, \quad \hat{U}(s, x) = \sum_{i=1}^M (U_0, \Phi_i) \frac{\Phi_i(x)}{s + \lambda_i},
\]

where \( \lambda_1, \lambda_2, \ldots \) are the positive eigenvalues of \( L \) in nondecreasing order, with orthonormal eigenfunctions \( \Phi_1, \Phi_2, \ldots \), and where \( \Lambda_1, \Lambda_2, \ldots, \Lambda_M \) (in nondecreasing order) and \( \Phi_1, \Phi_2, \ldots, \Phi_M \) are the positive eigenvalues and eigenfunctions of the problem

\[
B(\Phi, \Phi') = \Lambda_i (\Phi, \Phi'), \quad V \in S(\Lambda), \quad i = 1, \ldots, M.
\]

Note that

\[
\Lambda_1 = \inf_{V \in S(\Lambda)} \inf_{V \leq \lambda_1} \frac{B(V, V)}{\langle V, V \rangle} = \lambda_1.
\]

From (2.10), we see that \( D^l \hat{\psi} \) is meromorphic in \( s \) with the set \( \{-\lambda_i\}_{i=1}^\infty \cup \{-\Lambda_i\}_{i=1}^M \) as the only possible poles. Since these singularities lie outside the contours \( \gamma_1, \gamma_2, \gamma_3, \gamma_4 \) and \( \gamma_1, \gamma_2, \gamma_3, \gamma_4 \), we have by Cauchy's theorem

\[
\int_{\gamma_1, \gamma_2, \gamma_3, \gamma_4} D^l \hat{\psi}(s, x) \exp(st) \, ds = 0.
\]

Furthermore, since \( \gamma_1, \gamma_2, \gamma_3, \gamma_4 \) lies in the convergence half-plane of \( \hat{\psi} \), we can apply the complex inversion formula [6] to obtain

\[
D^l \hat{\psi}(s, x) = \frac{1}{2\pi i} \int_{\gamma_1, \gamma_2, \gamma_3, \gamma_4} D^l \hat{\psi}(s, x) \exp(st) \, ds.
\]

Hence we see immediately from (2.12) and (2.13) that we only have to prove that

\[
\lim_{R \to \infty} \int_{\gamma_1, \gamma_2, \gamma_3, \gamma_4} D^l \hat{\psi}(s, x) \exp(st) \, ds = 0.
\]

From (2.2), we can derive that

\[
|B(\hat{u}, \hat{u}) + s(\hat{u}, \hat{u})| = |(U_0, \hat{u})| \leq ||u||_0 ||\hat{u}||_0.
\]

Application of Lemma 2 for \( s = -\mu - \alpha \pm iR \) yields

\[
|B(\hat{u}, \hat{u})| \leq \frac{1}{2} ||u||_0^2 \left[ \alpha + \sqrt{\alpha^2 + R^2} \right], \quad |D^l \hat{u}(x)| \leq C ||\hat{u}||_\infty \leq CR^{-1/2} ||u||_0,
\]

if \( R \to \infty \). The last inequality was proved by Sobolev's embedding theorems [11] and by the strong coercivity of \( B \). In a similar way, we can prove from (2.3) that

\[
|D^l \hat{U}(s, x)| \leq CR^{-1/2} ||u||_0, \quad l = 0, \ldots, m - 1,
\]
if $R \to \infty$ and $s = \pm i (R - \alpha - \mu)$. From (2.16) one easily proves (2.14) and therewith the lemma. 

As in [2], we can exploit (2.9) to transfer local convergence properties of $\hat{\varepsilon}$ immediately to $e(\ell)$. Since these properties are not standard if $|x| \to \infty$, we have to prove them here explicitly, of course only for $s = -\alpha - \mu \pm i \alpha$. In the sequel $C(\alpha)$, $C_{1}(\alpha)$, etc. are positive functions of $\alpha$ which are polynomially bounded on $[0, \infty)$, not necessarily the same ones.

**Lemma 4.** Let $u_{0} \in S(\Delta)$ be any approximation of $u_{0}$ satisfying (1.12). Then $\hat{\varepsilon} = \hat{u} - \hat{U}$ has the bound 

\[(2.17) \quad \|\hat{\varepsilon}_{l}\| \leq C(\alpha) h^{k+1-l} \|u_{0}\|_{k+1}, \quad l = 0, \ldots, m.\]

**Proof.** From (2.2b) and (2.3), we find that

\[(2.18) \quad B(\hat{u} - \hat{U}, V) + s(\hat{u} - \hat{U}, V) = (u_{0} - U^{0}, V), \quad V \in S(\Delta).\]

Next, we introduce the elliptic projection $\hat{U}_{2} \in S(\Delta)$ of $\hat{u}$ by

\[(2.19) \quad B(\hat{u} - \hat{U}_{2}, V) = 0, \quad V \in S(\Delta).\]

It is standard [11] that $\|\hat{u} - \hat{U}_{2}\| \leq C h^{k+1} \|\hat{u}\|_{k+1}, \quad l = 0, \ldots, m$. If we put $v = \hat{\varepsilon} = \hat{U}_{2} - \hat{U}$ and subtract (2.19) from (2.18), we find

\[(2.20) \quad B(\hat{\varepsilon}, \hat{\varepsilon}) + s(\hat{\varepsilon}, \hat{\varepsilon}) = (u_{0} - U^{0} - s(\hat{u} - \hat{U}_{2}), \hat{\varepsilon}) \leq C \|\hat{\varepsilon}\|_{h}^{k+1} \|u_{0}\|_{k+1} + |x| \|\hat{u}\|_{k+1}.\]

Application of Lemma 2 to (2.20) yields

\[(2.21) \quad \|\hat{\varepsilon}\|_{h} \leq C(\alpha) h^{2(k+1)} \|u_{0}\|_{k+1} + |x| \|\hat{u}\|_{k+1}.\]

We now have

\[(2.22) \quad \|\hat{\varepsilon}\|_{l} \leq \|\hat{u} - \hat{U}_{2}\|_{l} + \|\hat{\varepsilon}\|_{l} \leq Ch^{k+1} \|\hat{u}\|_{k+1} + C(\alpha) h^{l} \|u_{0}\|_{k+1} + |x| \|\hat{u}\|_{k+1}, \quad l = 0, \ldots, m.\]

We still need an estimation of $\|\hat{u}\|_{k+1}$. From (2.2), we can derive that, since $L \hat{u} = \hat{L}^{h}(\hat{l}),$

\[(2.23) \quad B(L \hat{u}, L \hat{u}) + s(L \hat{u}, L \hat{u}) = (L u_{0}, L \hat{u}) \leq \|L u_{0}\| \|L \hat{u}\|_{0}.\]

Application of Lemma 2 yields

\[(2.24) \quad \|\hat{u}\|_{k+1} \leq C(\alpha) \|u_{0}\|_{k+1}.\]

From (2.22) and (2.24), we get (2.17), which proves the lemma. 

**Remark.** Although $C(\alpha)$ in (2.17) is polynomially bounded, it tends to be of $O(\lambda_{1}^{-\mu^{-1}})$ near $\alpha = 0$, as $\mu^{1/\lambda_{1}}$.

Now that we have established convergence of $\hat{\varepsilon}$ on the contour $\hat{P}_{4} P_{2} P_{3}$, we can investigate the superconvergence at the knots.
For any \( x \in (-1, +1) \) and \( l \in \{0, 1, \cdots, m-1\} \), we define the generalized Green's function \( \hat{G}_l(x, \xi) \in H_0^m(I) \cap H^{l+1}(0, x) \cap H^{l+1}(x, 1) \) associated to \( L \), by
\[
\begin{align*}
L \hat{G}_l(x, \xi) + \bar{S} \hat{G}_l(x, \xi) &= 0, \quad \xi \in I(x), \\
B(x) \hat{G}_l(x) + s(x, \hat{G}_l(x)) &= D^l \chi(x), \quad \nu \in H_0^m(I),
\end{align*}
\]
where the subscript \( \xi \) of \( L \) denotes partial differentiation with respect to \( \xi \). If we denote
\[
\hat{G}_l(\xi) = \hat{G}_l(x, \xi), \quad j = 1, \cdots, N-1, \quad l = 0, \cdots, m-1,
\]
we find for \( D^l \chi(x) \) the bound
\[
\begin{align*}
|D^l \chi(x)| &\leq |B(\xi, \hat{G}_l) + s(\xi, \hat{G}_l)| \\
&\leq |B(\xi, \hat{G}_l - V)| + |B(\xi, V)| + |s(\xi, \hat{G}_l - V)| + |s(\xi, V)| \\
&\leq C(\alpha) \|\hat{G}_l\|_m + \|\hat{G}_l - V\|_m + \|\hat{G}_l - V\|_m + \|\hat{G}_l - V\|_m,
\end{align*}
\]

Since \( \hat{G}_l \in H_0^m(I) \cap H^{l+1}(\Delta) \), we can take \( V \) such that
\[
\|\hat{G}_l - V\|_m \leq C \|\hat{G}_l\|_m, \quad \|\hat{G}_l - V\|_{m+1, \Delta} \leq C \|\hat{G}_l\|_{m+1, \Delta},
\]
Then it is easily proved from (2.17) and (2.27) that
\[
\|\hat{G}_l - V\|_m \leq C(\alpha) h^m \|\hat{G}_l\|_{m+1, \Delta} + \|\hat{G}_l - V\|_m + \|\hat{G}_l - V\|_m,
\]

We have yet to estimate \( \|\hat{G}_l - V\|_m \) and \( \|\hat{G}_l\|_{m+1, \Delta} \). We have to choose \( V \), such that \( \|\hat{G}_l - V\|_m \) is of \( O(h^m) \). A seductive choice of \( U_0 \) would be the \( L^2 \) projection of \( u_0 \), annihilating \( (u_0 - U_0) v \). A drawback, however, is that it is not easily computed.

In the next sections, we will construct a \( U_0 \) which also has superconvergence of \( D^l \chi(x) \), is simpler and which imposes rather mild extra conditions to \( u_0 \) and \( u(t) \): they also have to be in \( W^{2, \Delta} \). Although we chose \( \Delta \) uniform, for reasons of convenience, it can, of course, also be chosen quasiuniform, if this helps to meet the extra conditions.

2.2. Choice of nodal points, Jacobi polynomials. In order to construct a proper approximation \( U_0 \) of \( u_0 \), we first define the \( \nu \)th degree Jacobi polynomial \( P_{\nu, \alpha, \beta}(x) \) by [1], [13]
\[
\begin{align*}
P_{\nu, \alpha, \beta}(x) &= (w(x))^{-\frac{1}{2}} D^\nu [((1-x)^{\alpha}) w(x)], \quad \nu \geq 0, \\
w(x) &= (1-x)^{\alpha}(1+x)^{\beta}, \quad x \in (-1, +1), \quad \alpha, \beta \geq 1.
\end{align*}
\]

These polynomials have the properties [1], [13]
\[
\begin{align*}
(w P_{\nu, \alpha, \beta}, P_{\mu, \alpha, \beta}) &= \delta_{\mu \nu} (w P_{\nu, \alpha, \beta}, P_{\nu, \alpha, \beta}), \quad \mu, \nu \geq 0, \\
P_{\nu, \alpha, \beta}(x_{\nu}) &= 0, \quad -1 < x_1 < x_2 < \cdots < x_{\nu} < 1,
\end{align*}
\]
where \( \delta_{\mu \nu} \) is the Kronecker symbol.

Within the context of this paper, we are only interested in the case \( \alpha = \beta = m \).

We recall that \( r = k + 1 - m \) and \( n = k + 1 - 2m \). Let \( \sigma_1, \cdots, \sigma_n \) be the zeros of \( P_{\nu, \alpha, \beta}(\sigma) \); i.e.,
\[
P_{\nu, \alpha, \beta}(\sigma_l) = 0, \quad l = 1, \cdots, n.
\]
Of course, (2.32) only makes sense if \( n \equiv 1 \). In the sequel, it is tacitly assumed that the formulae which make no sense if \( n = 0 \) are to be omitted.

Given a partition \( \Delta \) of \( I \), we define the points \( \xi_l \) by

\[
(2.33) \quad \xi_l = x_{j-1} + \frac{h}{2}(1 + \sigma), \quad j = 1, \cdots, N, \quad l = 1, \cdots, n.
\]

Next, we introduce the linear interpolation \( \Pi : H^2_0(I) \cap W^{2r}(\Delta) \rightarrow S(\Delta) \) by

\[
(2.34) \quad D^r \Pi f(x_l) = D^r f(x_l), \quad l = 0, \cdots, m - 1, \quad j = 1, \cdots, N - 1,
\]

\[
\Pi f(x_l) = f(\xi_l), \quad l = 1, \cdots, n, \quad j = 1, \cdots, N.
\]

**Lemma 5.** For any \( V \in S(\Delta) \) and \( f \in H^2_0(I) \cap W^{2r}(\Delta) \)

\[
(2.35) \quad |(f - \Pi f, V)| \leq C h^{2r} ||f||_{w^{2r}(\Delta)} ||V||_{w^{2r}(\Delta)}.
\]

**Proof.** For \( n = 0 \), (2.35) is trivial [11]. For \( n \equiv 1 \), we consider an arbitrary segment \( I_l \). If we substitute \( x = \frac{1}{2}(x_{j-1} + x_j + h\sigma) \), \( \sigma \in I \), we find that

\[
(f - \Pi f, V)_{h_l} = \frac{1}{2} h \int_{I_l} [(f - \Pi f) V] \left( \frac{1}{2}(x_{j-1} + x_j + h\sigma) \right) d\sigma
\]

\[
= \frac{1}{2} h \int_{I_l} \left[ 1 - (1 - 2^2)^m \right] ||P_{m,m}(\sigma) ||_0 ||gV|| \left( \frac{1}{2}(x_{j-1} + x_j + h\sigma) \right) d\sigma,
\]

where \( g \) is bounded on \( I \). From (2.31), we conclude that \( (f - \Pi f, V)_{h_l} = 0 \) if \( g \in P_{m,m}(I_l) \) or \( fV \in P_{2r-1}(I_l) \). Application of Bramble and Hilbert's Lemma [3] yields

\[
(2.36) \quad |(f - \Pi f, V)| \leq C h^{2r} ||D^r fV||_{w^{2r}(\Delta)} \quad j = 1, \cdots, N.
\]

Elaboration of (2.36) and summation over all \( I_l \) results in (2.35) and proves the lemma.

Note that by (2.34) we have defined all the nodal points of \( S(\Delta) \).

### 2.3. Order of convergence at the knots

We return to (2.29) recalling that

\[
|D^r \hat{e}(x_l)| \leq C(\alpha) h^{2r} ||u_0||_{w^{2r}(\Delta)} ||\hat{G}_l||_{w^{2r+1}(\Delta)} + ||u_0 - U_0||_{w^{2r}(\Delta)} V(l), \quad l = 1, \cdots, m - 1,
\]

where \( V \) is an approximation of \( \hat{G}_l \) satisfying (2.28). If we take \( U_0 = \Pi u_0 \), \( \Pi \) defined by (2.34), then application of (2.28) and Lemma 5 gives

\[
|D^r \hat{e}(x_l)| \leq C(\alpha) h^{2r} ||u_0||_{w^{2r}(\Delta)} ||\hat{G}_l||_{w^{2r+1}(\Delta)} + ||u_0||_{w^{2r}(\Delta)} ||V||_{w^{2r}(\Delta)} \quad l = 1, \cdots, m - 1.
\]

It is easily proved that \( ||\hat{G}_l||_{w^{2r+1}(\Delta)} \) is polynomially bounded, hence we can prove by combination of (2.37) and Lemma 3 that

\[
(2.38) \quad |D^r \hat{e}(t, x_l)| \leq h^{2r} e^{-\gamma t} ||u_0||_{w^{2r}(\Delta)} \int_0^t e^{-\gamma \tau} C(\alpha) d\tau, \quad t > 0.
\]

We see that (2.38) does not hold down to \( t = 0 \). It is tempting to conclude from the equalities

\[
D^r \hat{e}(0, x_l) = 0, \quad l = 0, \cdots, m - 1, \quad j = 1, \cdots, N - 1,
\]

that (2.38) can be extended to the entire interval [2].
But this conclusion is erroneous as a not too far-fetched counter-example in the appendix shows. That \( U_0 = \Pi u_0 \) satisfies (1.12) is trivial since \( \Pi \) leaves all members of \( S(\Delta) \) invariant. This concludes the proof of the following theorem.

**Theorem 1.** Let \( u : I \rightarrow H^m \cap H^{m+1}(I) \cap W^2(\Delta) \) be the solution of (1.1) and let \( U : I \rightarrow S(\Delta) \) be the solution of (1.11) with \( U_0 \) defined by (2.34). Then the error function \( e = u - U \) has the global bound (1.13), if \( h \) is small enough. Furthermore, for any \( \tau > 0 \) and \( \mu \in (0, \lambda) \) there exists a positive \( F(\tau, \mu) \) depending on \( \tau \) and \( \mu \) only, such that

\[
|D^j e(x, t)| \leq F(\tau, \mu) e^{-\tau h^{2j}} \|u_0\|_{W^{2j+2}}
n = 0, \ldots, m - 1, \quad j = 1, \ldots, N - 1, \quad t \geq \tau.
\]

(2.39)

**2.4. Order of convergence at Jacobi points.** In this section, we will prove that the order of convergence at the points \( \xi_i \) defined by (2.33) is of \( O(h^{k+2} e^{-\alpha h}) \). Since these points only exist if \( n \equiv 1 \), we confine our attention to the case \( k \equiv 2m \).

For any \( I \in \Delta \), we define

\[
S(I) = \{ V \mid V \in S(\Delta); \supp(V) = I \}.
\]

It is evident that \( S(I) \) has dimension \( n \) and that

\[
D^j V(x) = 0, \quad x \in \partial I, \quad V \in S(I), \quad j = 0, \ldots, m - 1.
\]

(2.41)

We define a basis \( \{ \phi_i \}_{i=1}^{\infty} \) of \( S(I) \) by

\[
\phi_i(x) = \delta_{ih}, \quad i \leq i, \quad l \leq n.
\]

(2.42)

If we apply (2.18) for \( \phi_1, \ldots, \phi_m \) we find after partial integration that

\[
(\xi, L \phi_i + \delta \phi_i)_{\eta} = (u_0 - U_0, \phi_i)_{\eta} + \sum_{i=1}^{m} \sum_{j=0}^{l-1} \int_I \{(-1)^j D^j (p D^j D^i - \xi i_i) i_i\} \, dx.
\]

(2.43)

In order to approximate the inner product \( (\cdot, \cdot) \) by a quadrature rule involving the function values at \( \xi_i \) which is accurate enough, we define for \( f \in W^2(I) \) the approximation

\[
\int_{-1}^{1} f(\sigma) \, d\sigma \approx \Pi f(\sigma) \, d\sigma,
\]

(2.44)

where \( \Pi : W^2(I) \rightarrow P_{n-1}(I) \) is defined by (2.34) shifted from \( I \) to \( I \). Note that in the case \( m = 1 \), we obtain Lobatto's quadrature rule [1].

**Lemma 6.** Quadrature rule (2.44) is exact if \( f \in P_{2r-1}(I) \).

**Proof.** Since

\[
f(\sigma) - \Pi f(\sigma) = (1 - \sigma^2)^m P_n^m(\sigma) g(\sigma),
\]

where \( g(\sigma) \) is bounded, it is evident that (2.44) is exact if \( g \in P_{n-1}(I) \), i.e., if \( f \in P_{2r-1}(I) \). \( \square \)

Elaboration of (2.44) yields

\[
\int_{-1}^{1} \Pi f(\sigma) \, d\sigma = \sum_{i=1}^{n} \theta_i f(-1) + \sum_{i=1}^{n} \theta_i f(+1).
\]

(2.45)

where \( \sigma_1, \ldots, \sigma_n \) are the zeros of \( P_n^m(\sigma) \) and \( \theta_1, \theta_2, \) and \( \omega_i \) are constant weights.
By applying (2.44) to \( f(\sigma) = (1 - \sigma^2)^m P_{\infty + m}(\sigma)/(1 - \sigma^l) \), \( l = 1, \ldots, n \), one can prove that [13, Chapt. XV]

\[ \omega_l = \mu_l (1 - \sigma^l)^{m_l} \]

where \( \mu_1, \ldots, \mu_n \) are the positive Gauss-Chebyshev numbers for the \( n \)-point Gauss-Jacobi quadrature formula with weight function \((1 - \sigma^l)^m\). This proves that \( \omega_l > 0 \), \( l = 1, \ldots, n \).

Next, we define for \( \alpha, \beta \in W^{2\alpha}(I) \)

\[ (\alpha, \beta)_l = \frac{1}{2} \sum_{i=1}^{n} \omega_l \beta_i \beta_i = \frac{1}{2} \sum_{i=1}^{n} \left( \frac{1}{2} \right) \left[ \beta_i D^l(\alpha \beta_i)(x_i, \sigma) + \beta_i D^l(\alpha \beta_i)(x_i, -\sigma) \right] \]

This quadrature rule has the error bound [3]

\[ |\alpha, \beta|_l - (\alpha, \beta)_l| \leq C \delta^{2\alpha} \|D^{2\alpha}(\alpha \beta)_l\|_{L^\infty(I)} \]

If we apply (2.46) to (2.43) and multiply by \( 2h^{2m-1} \), we obtain

\[ \left| \sum_{l=0}^{n} \frac{h^{2m}}{2} \omega_l [L \phi_l(\xi_i) + s \delta_l] \hat{e}_l(\xi_i) \right| \]

\[ = 2h^{2m-1} \left| \sum_{l=0}^{n} \frac{1}{2} \sum_{i=1}^{n} \beta_i D^l(\alpha \beta_i)(x_i, \sigma) \right| + (\beta, L \phi_l + s \delta_l) L^l - (\beta, L \phi_l + s \delta_l) + (u_0 - U_0) \phi_l \]

\[ + \sum_{l=1}^{n} \sum_{i=0}^{n} \sum_{j=0}^{n} \left| (-1)^j D^{j}(p D^i \phi_l) D^{i-1} \hat{e}_l \right| \]

\[ \leq C(\alpha, \beta) h^{2m} \|\phi_l\|_{L^{\infty}(I)} \sum_{l=0}^{n} \sum_{i=1}^{n} \|D^l(\alpha \beta_i)(x_i, \sigma)\|_{L^{\infty}(I)} \]

\[ + C_2 h^{2m+2} \|D^{2\alpha}(\alpha \beta_i)(x_i, \sigma)\|_{L^{\infty}(I)} + C_3 h^{2m+2} \|D^{2\alpha}(u_0 \phi_l)\|_{L^{\infty}(I)} \]

\[ + C_4 h^{2m+2} \|\phi_l\|_{L^{\infty}(I)} \sum_{l=0}^{n} \sum_{i=0}^{n} \|D^{l}(\alpha \beta_i)(x_i, \sigma)\|_{L^{\infty}(I)} \]

\[ \leq C(\alpha, \beta) h^{2m+2} \|u_0\|_{L^{\infty}(I)} + C_5(\alpha, \beta) h^{2m+2} \|u_0\|_{L^{\infty}(I)} \]

\[ + C_6 h^{2m+2} \|\alpha\|_{L^{\infty}(I)} + C_7(\alpha, \beta) h^{2m+2} \|\alpha\|_{L^{\infty}(I)} \]

\[ \leq C(\alpha, \beta) h^{2m+2} \|\alpha\|_{L^{\infty}(I)} + C_3(\alpha, \beta) h^{2m+2} \|\alpha\|_{L^{\infty}(I)} \]

\[ f = 1, \ldots, N - 1, \quad l = 1, \ldots, n, \]

where we take, of course, the right and left limit for the function values at \( x_{i \pm 1} \) and \( x_i \) respectively. We have yet to estimate \( \|\hat{e}_l\|_{L^{\infty}(I)} \).

Let \( \hat{\Pi}_l \) be the interpolation of \( \hat{a} \) defined by (2.34). Then we can prove from [4], [11] and [2] that

\[ \|\hat{e}_l\|_{L^{\infty}(I)} \leq \|\hat{\Pi}_l \hat{a} - \hat{a}\|_{L^{\infty}(I)} + |\hat{a} - \Pi_l \hat{a}|_{L^{\infty}(I)} \]

\[ \leq C(\alpha, \beta) \|\hat{\Pi}_l \hat{a} - \hat{a}\|_{L^{\infty}(I)} + C_3(\alpha, \beta) \|\hat{a}\|_{L^{\infty}(I)} \]

\[ \leq C(\alpha, \beta) \|\hat{\Pi}_l \hat{a} - \hat{a}\|_{L^{\infty}(I)} + C_3(\alpha, \beta) \|\hat{a}\|_{L^{\infty}(I)} \]

\[ f = 1, \ldots, N, \]

where \( \hat{a} \) is the interpolation of \( a \) defined by (2.34). Then we can prove from [4], [11] and [2] that

\[ \|\hat{e}_l\|_{L^{\infty}(I)} \leq \|\hat{\Pi}_l \hat{a} - \hat{a}\|_{L^{\infty}(I)} + |\hat{a} - \Pi_l \hat{a}|_{L^{\infty}(I)} \]

\[ \leq C(\alpha, \beta) \|\hat{\Pi}_l \hat{a} - \hat{a}\|_{L^{\infty}(I)} + C_3(\alpha, \beta) \|\hat{a}\|_{L^{\infty}(I)} \]

\[ f = 1, \ldots, N, \]
From (2.22) plus the Poincaré inequality, we can derive that
\[ h^{-1} \| \mathbf{e} \|_{W^{1,1}(T)} \leq Ch^{-1} \| \mathbf{e} \|_{L^2(T)} \leq C(\alpha) \| u_{\ell,2} \|_{L^2(T)} + \| \mathbf{e} \|_{L^2(T)}. \]

Hence
\[ (2.50) \quad \| \mathbf{e} \|_{W^{1,1}(T)} \leq C(\alpha) \| u_{\ell,2} \|_{L^2(T)} + \| \mathbf{e} \|_{L^2(T)} + \| \mathbf{e} \|_{W^{1,1}(T)}. \]

\[ \| \mathbf{e} \|_{L^2(T)} \] was already estimated (formula (2.24)), for the estimation of \( \| \mathbf{e} \|_{W^{1,1}(T)} \), we simply use the differential equation (2.2a) to obtain
\[ (2.51) \quad \| \mathbf{e} \|_{W^{1,1}(T)} \leq C(\alpha) \| u_{\ell,2} \|_{W^{1,1}(T)}. \]

In summary, we have obtained from (2.48)–(2.51) that
\[ (2.52) \quad \sum_{i=1}^{n} h^{2m} \omega_{i} \langle L_{D} \phi_{i} \rangle \pm \delta \phi_{i} \| \mathbf{e} \|_{L^2(T)} \leq C(\alpha) h^{4+2} \| u_{\ell,2} \|_{W^{1,1}(T)}, \quad i = 1, \ldots, n. \]

We have to prove the solvability of the linear system (2.52). It is easily proved that
\[ (2.53) \quad \left| \left( \omega_{i} L_{D} \phi_{i} - \frac{2}{h} B(\phi_{i}, \phi_{i}) \right) h^{2m} \right| \leq Ch^{2}, \]
if \( h \) is small enough. Consequently, the matrix \( (h^{2m} \omega_{i} L_{D} \phi_{i}) \) approximates a symmetric positive definite matrix whose eigenvalues are of \( O(h^{2}) \). This means that its eigenvalues are nearly positive, i.e., the real parts are positive of \( O(h^{2}) \) and the imaginary parts are of \( O(h) \). Since \( \mathcal{E} \in \mathcal{P}_{\alpha} \mathcal{P}_{\alpha} \), we can show from (2.52) by elementary matrix calculus that
\[ (2.54) \quad \| \mathbf{e} \|_{L^2(T)} \leq C(\alpha) h^{4+2} \| u_{\ell,2} \|_{W^{1,1}(T)}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, N. \]

Application of Lemma 3 to (2.54) leads to the bound
\[ |e(t, \xi_{j})| \leq h^{4+2} e^{-\alpha t} \| u_{\ell,2} \|_{L^2(T)} \int_{0}^{\alpha} e^{-\alpha t} C(\alpha) \, dt. \]

Again, the fact that \( e(0, \xi_{j}) = 0 \) is not sufficient to extend the above bound down to \( t = 0 \), as is shown in the appendix.

**Theorem 2.** Let the conditions of Theorem 1 hold with the restriction that \( k \geq 2m \). Then \( e(t) \) has the bounds (1.13) and (2.39) plus the additional bound
\[ (2.55) \quad |e(t, \xi_{j})| \leq F(\tau, \mu) e^{-\alpha t} h^{4+2} \| u_{\ell,2} \|_{W^{1,1}(T)}, \quad t \leq \tau, \quad j = 1, \ldots, N, \quad l = 1, \ldots, n. \]

**3. Quadrature rules.** When solving (1.11), one is usually forced to approximate \( B(U, V) \) by some quadrature [12]. The choice of this rule is, as usual, dictated not only by its accuracy but also by its impact on the convergence properties. It may sometimes be useful to approximate \( B(U, V) \) by a quadrature rule as well, e.g., in the case \( m = 1 \) where the choice of \((k+1)\)-point Lobatto quadrature delivers a purely explicit system of ordinary differential equations [2]. However, in this paper, we confine ourselves to the numerical quadrature of \( B(U, V) \).

**3.1. \( q \)th order rules.** Let \( q \geq 2r \) be a constant integer and let \( -1 \leq z_{1} < z_{2} < \cdots < z_{p} \leq 1 \) be \( p \) distinct points on \( I \) and let, for \( f \in W^{q}(I) \),
\[ (3.1) \quad \int_{-1}^{1} f(z) \, dz = \sum_{i=1}^{p} w_{i} f(z_{i}). \]
be an approximation which is exact if $f \in P_{r-1}(I)$. Given a partition $\Delta$ of $I$ we define, for $\alpha, \beta \in W^k(\Delta)$,
\[
(\alpha, \beta)^* = \frac{h}{2} \sum_{i=1}^{N} w_i (\alpha \beta) \left( x_i - \frac{h}{2} + 1 + z_i \right).
\]
(3.2)
\[
(\alpha, \beta) = \sum_{i=1}^{N} (\alpha, \beta)^*.
\]
\[
B_h(\alpha, \beta) = \sum_{i=0}^{m} (p_i D^i \alpha, D^i \beta).
\]
As examples, we can take $r$-point Gauss–Legendre or $(r + 1)$-point Lobatto quadrature.

**Lemma 7.** For any $U, V \in S(\Delta)$, we have for sufficiently small $h$
\[
|B(U, V) - B_h(U, V)| \leq Ch^{r+1} \|D^r(U) D^r V\|_{L^\infty(L)}.
\]

**Proof.** Application of Bramble and Hilbert's lemma [3] gives
\[
|B(U, V) - B(U, V)| \leq Ch^{r+1} \sum_{i=1}^{N} \sum_{j=0}^{m} \|D^r(p_i D^j U D^j V)\|_{L^\infty(L)}.
\]
(3.4)
\[
\leq Ch^{r+1} \|U\|_{L^\infty(L)} \|V\|_{L^\infty(L)}.
\]
By applying Lemma 7 for $i = j = m$, we can easily prove the following corollary.

**Corollary 1.** If $h$ is sufficiently small then the bilinear mapping $B_h : S(\Delta) \times S(\Delta) \to \mathbb{C}$ is strongly coercive.

As the last preliminary of this section, we prove the next lemma.

**Lemma 8.** For $v \in H^{r+1}(I) \cap H^r_0(I) \cap W^k(\Delta)$, let $V \in S(\Delta)$ be an approximation of $v$ with the error bound
\[
\|v - V\| \leq Ch^{r+1-\mu} \|v\|_{\mu+1}, \quad l = 0, \cdots, m.
\]
Then we have
\[
\|V\|_{\mu+1} \leq C \|v\|_{\mu+1}.
\]
(3.6)

**Proof.** Let $\Pi : H^{r+1}(\Delta) \cap H^r_0(I) \cap W^k(\Delta) \to S(\Delta)$ be defined by (2.9). Then
\[
\|V\|_{\mu+1} \leq \|V - \Pi v\|_{\mu+1} + \|v - \Pi v\|_{\mu+1} + \|v\|_{\mu+1}
\]
(3.7)
\[
\leq C_1 h^{-1} \|V - \Pi v\|_{\mu+1} + C_2 h^{-1} \|v\|_{\mu+1} + \|v\|_{\mu+1}
\]
\[
\leq C_1 \|v\|_{\mu+1} + C_2 h^{-1} (\|v\|_{\mu+1} + \|v - \Pi v\|_{\mu+1}) \leq C \|v\|_{\mu+1}.
\]

**3.2. Preservation of the orders of convergence.** In this section, we shall prove that the replacement of $B(\cdot, \cdot)$ by $B_h(\cdot, \cdot)$ does not affect the validity of Theorems 1 and 2 except that the constant $\mu$ will be slightly smaller. This is due to the fact that
\[
\mu < \Lambda^*_h = \inf_{V \in S(\Delta)} \frac{B_h(V, V)}{\|V\|_{\mu+1}},
\]
and $\Lambda^*_h$ need no longer be greater than $\lambda_1$.

Let $Y : J \to S(\Delta)$ be the solution of the initial boundary problem
\[
\frac{\partial Y}{\partial t} + B_h(Y, V) = 0, \quad V \in S(\Delta), \quad t \in J, \quad Y(0) = U_0 = \Pi u_0.
\]
(3.9)
where $\Pi$ is defined by (2.34) and $B_h$ by (3.2). We define

$$\eta(t) = U(t) - Y(t),$$

(3.10)

where $U$ is the solution of (1.11). We again define the points $P_1, P_2, \cdots, P_s$ by (2.8) where we take care that (3.8) holds, in other words that (see Fig. 1) $\eta = \mathcal{L}\eta(s)$ has no poles inside $P_1 P_2 P_3 P_4$. Then we can prove, in analogy to Lemma 2, that

$$D^l \eta(t, x) = \frac{e^{-it\omega}}{\pi} \int_0^\infty e^{-is\omega} \text{Im} [(1 + i) e^{-is\omega} D^j \eta(-\alpha - \mu - ia, x)] \, ds,$$

(3.11)

$$l = 0, \cdots, m - 1.$$

As before, we are only interested in the case $s \in P_1 P_2 P_3$. By applying $\mathcal{L}$ to (3.9) and subtracting the result from (2.3) we get

$$B_h(\eta, V) + s(\eta, V) = B_h(\hat{V}, V) - B(\hat{V}, V), \quad V \in S(\Omega).$$

(3.12)

If we substitute $V = \eta$ and apply Lemmas 7 and 8 plus formula (2.24), we get

$$|B_h(\eta, \eta) + s(\eta, \eta)| \leq C h^{m-\alpha} \|\eta\|_{m, \alpha}$$

(3.13)

Since $B_h(\tilde{\eta}, \tilde{\eta}) \equiv \langle \lambda(\tilde{\eta}, \tilde{\eta}) \rangle$ and $B_h$ is strongly coercive, we can prove from (3.13) that

$$\|\eta\|_{m, \alpha} \leq C(\alpha) h^{m-\alpha} \|\eta\|_{m, \alpha} + 1.$$

(3.14)

For $\tilde{\eta}$ now we can prove the local bounds

$$|D^l \tilde{\eta}(x)| = |B(\tilde{\eta}, \tilde{\eta}) + s(\tilde{\eta}, \tilde{\eta})|$$

(3.15)

$$\leq |B(\tilde{\eta}, \tilde{\eta}) - V| + |s(\tilde{\eta}, \tilde{\eta} - V)| \leq C(\alpha) \|\tilde{\eta}\|_{m, \alpha} + C(\alpha) \|\tilde{\eta}\|_{m, \alpha} + C(\alpha) \|\tilde{\eta}\|_{m, \alpha}.$$

We take $V$ such that (2.28) holds. For $\tilde{\eta}$, we see that

$$\|\tilde{\eta} - \tilde{\eta} - V\| \leq \|\tilde{\eta}\| + \|\tilde{\eta}\| \leq C(\alpha) h^{-1} \|\tilde{\eta}\|_{m, \alpha},$$

hence after application of Lemma 8, we get

$$\|\tilde{\eta}\|_{m, \alpha} \leq C(\alpha) \|\tilde{\eta}\|_{m, \alpha} + 1 \leq C(\alpha) \|\tilde{\eta}\|_{m, \alpha} + 1.$$

(3.16)

From (3.14)-(3.16), it now easily follows that

$$|D^l \eta(t, x)| \leq C(\alpha) h^{m-\alpha} \|u\|_{m, \alpha}, \quad l = 0, \cdots, m - 1,$$

(3.17)

$$j = 1, \cdots, N - 1,$$

and as an immediate result of (3.11) and (3.17)

$$|D^l \eta(t, x)| \leq F(\tau, \mu) e^{-\tau\alpha} \|u\|_{m, \alpha},$$

(3.18)

where $F$ depends on $\tau$ and $\mu$ only. We again refer to the appendix for a proof that $\tau$ should be positive.

For the local bounds of $\eta(t)$ at the Jacobi points, we confine our attention to the case $k \geq 2m$. Let $S(\Omega)$ and $\xi_0$ be defined by (2.40) and (2.33). Then for arbitrary $j$,
we can prove from (3.12) that
\[
(\tilde{\eta}, LV + \varepsilon V) = B_3(\tilde{Y}, V) - B(Y, V) \\
+ \sum_{i=1}^{m-1} \sum_{k=0}^{i} \left[ (-1)^i D^i(p_i D^i V) D^{i-1} \tilde{\eta} \right]_{x_i, k+1}, \quad V \in S(I).
\]
(3.19)

If we apply the quadrature rule (2.44) to (3.19) put \( V = \phi_n \) where \( \phi_n \) is defined by (2.42) and multiply by \( 2h^{2m-1} \), we obtain, in analogy to (2.50),
\[
\left| \sum_{i=1}^{m-1} \omega_i h^{2m-1} (L\phi_n(\tilde{\eta}_i) + s_i \tilde{\eta}(\tilde{\eta}_i)) \right| \\
= 2h^{2m-1} \left| - \sum_{i=1}^{m-1} \left( \frac{h}{2} \right)^{i+1} \sum_{k=0}^{i} \theta_i D^i(\tilde{\eta}(L\phi_n + s_i \tilde{\eta}))_{x_i, k+1} \right| \\
+ \left( \tilde{\eta}, L\phi_n + s_i \tilde{\eta} \right)_{x_i, 1} \tilde{\eta}(x_i) \\
= C_1(\alpha) h^{2m-1} \left[ \| \phi_n \|_{W^{m-1}(I)} + \sum_{i=1}^{m-1} \sum_{k=0}^{i} |D^i \tilde{\eta}(x_i)_{k+1}| \right] \\
+ C_2 h^{2m-2} \| D^{2m-1} \tilde{\eta}(L\phi_n + s_i \tilde{\eta}) \|_{L^\infty(I)} + C_3 h^{2m-2} \sum_{i=1}^{m} \| D^{2m} (p_i D^m \phi_n) \|_{L^\infty(I)} \\
(3.20)
\]

For the last inequality we used Lemma 8 and the inequality
\[
\| \tilde{\eta} \|_{W^{k}(I)} \leq C h^{-m-k} \| \tilde{\eta} \|_{W^{m-k}(I)} \leq C \| \tilde{\eta} \|_{W^{m-k}(I)} \leq C(\alpha) h^{2m-2} \| y \|_{L^\infty(I)} \\
\leq C(\alpha) h^{2m-2} \| y \|_{L^\infty(I)} + C(\alpha) h^{2m-2} \| u_0 \|_{L^\infty(I)},
\]
which can be proved by Sobolev's embedding theorems [11] and (3.7). From (3.20) and the results of § 2.4, we easily prove that
\[
\| \tilde{\eta}(\tilde{\eta}_i) \| \leq C(\alpha) h^{2m-2} \| u_0 \|_{L^\infty(I)}, \quad i = 1, \ldots, n.
\]
(3.21)

and application of (3.11) gives
\[
\| \tilde{\eta}(t, \tilde{\eta}_i) \| \leq F_1(\tau, \mu) e^{-\alpha h^{2m}} \| u_0 \|_{L^\infty(I)}, \quad t \geq \tau,
\]
(3.22)
where \( F \) depends on \( \tau \) and \( \mu \) only.

We have to estimate \( \| \eta(t) \| \) yet. Since \( \eta \in S(\Delta) \), this job is very easy, because all the nodal values of \( \eta(t): D^i \tilde{\eta}(t, x_i) \) and \( \tilde{\eta}(t, x_i) \) have been shown to be of \( O(h^{2m+k}F(t, \mu) e^{-\alpha}) \). This implies automatically that
\[
\| \tilde{\eta}(t) \|_{L^\infty(I)} \leq F_2(\tau, \mu) e^{-\alpha} h^{2m} \| u_0 \|_{L^\infty(I)},
\]
(3.23)
and
\[
\| \eta(t) \| \leq F_2(\tau, \mu) e^{-\alpha} h^{2m} \| u_0 \|_{L^\infty(I)}, \quad t \geq \tau.
\]
where \( F_1 \) and \( F_2 \) depend on \( \tau \) and \( \mu \) only. For \( n = 0 \), we have to replace \( k + 2 \) by \( k + 1 \) in (3.23). Thus we have proved the next theorem.

**Theorem 3.** Let \( Y : J \rightarrow S(\Delta) \) be the solution of (3.9) and let \( u : J \rightarrow H_0^1(I) \cap H_{\text{loc}}^\infty(I) \cap W^{\infty}(\Delta) \) be the solution of (1.1) with \( q \geq 2r \). Then, if \( h \) is small enough, the error function \( \varepsilon(t) = u(t) - Y(t) \) has the bounds

\[
\begin{align*}
&|\varepsilon(t)| \leq |\varepsilon(t)| + F_1(\tau, \mu) e^{-\mu \tau} ||u||_{L^\infty} , \quad \nu = \min (k + 2, 2r), \\
&|\varepsilon(t)| \leq F_2(\tau, \mu) e^{-\mu \tau} ||u||_{L^\infty} , \\
&|\varepsilon(t)| \leq F_3(\tau, \mu) e^{-\mu \tau} ||u||_{L^\infty} , \quad t = 1, \ldots, N, \quad \varepsilon(t) \equiv \tau.
\end{align*}
\]

where \( |\varepsilon(t)| \) has the bound (1.12), \( \mu \) has the bound (3.8) and where \( F_1, F_2 \) and \( F_3 \) depend on \( \tau > 0 \) and \( \mu \) only.

4. **Conclusions.** In the preceding sections we saw that earlier superconvergence results [2], [7], [8], [9], [10] can be generalized to \( 2n \)th order problems if the spatial operator is independent of time and linear. In that case the Laplace transformation enabled us to transfer the local convergence result of \( e(t) \) to its object function \( e(t, x) \). It also was made clear how the superconvergence of \( e(t) \) at the knots and interior nodal points crucially depends on the convergence properties of \( e(0) \). Furthermore, it was shown that Jacobi points play an important role in this matter: they are to be chosen as interior nodal points for the Hermite interpolation of \( u(0) \) and the local order of convergence is better at these points than at other interior points. En passant, we also gave a proof for superconvergence phenomena in the case of a \( 2n \)th order elliptic problem. That the use of \( q \)th order quadrature rules, necessary to evaluate the stiffness matrix, left all the convergence results of §2 unaltered was to be expected, although the supremum error of \( \eta(t) \) is lower than usual.

**Appendix.** Below, we will illustrate by a simple example that it is in principle impossible to extend (2.38) to \( t = 0 \).

Consider the problem

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad t \in I, \quad x \in I, \quad u(0, x) = 1 - x^4.
\end{align*}
\]

This is an example from the class \( m = 1 \) with \( p_0 = 0 \) and \( p_1 = 1 \). We apply the Faedo-Galerkin method for \( k = 2 \), using the Lagrangian basis for \( S(\Delta) \). The result is the system of ordinary differential equations

\[
\begin{align*}
d_a &= 0, \\
8d_{21} + 2(d_{21} + d_{21+1}) - (d_{21} + d_{21+2}) &= \frac{10}{h} \left( 14a_{21} - 8(a_{21} + a_{21+1}) + a_{21} - a_{21+2} \right), \quad j = 1, \ldots, N - 1, \\
d_{2N} &= 0, \\
8d_{21} + 2a_{21} + a_{21} &= \frac{40}{h} \left( 2a_{21} - 2a_{21} - a_{21+1} \right), \quad j = 1, \ldots, N, \quad t \equiv 0,
\end{align*}
\]

where the overdot stands for \( d/dt \), \( a_{21} \) stands for \( U(t, x_j - (k/2)) \) and \( a_{21} \) stands for \( U(t, x_j) \).
If we put
\[ \epsilon_{2j-1}(t) = (u - U)(t, x_j - \frac{h}{2}), \quad j = 1, \ldots, N, \]
(3)
\[ \epsilon_{2j}(t) = (u - U)(t, x_j), \quad j = 0, \ldots, N, \]
and apply (1) and (2) for \( t = 0 \) and subtract the results, we find for \( \dot{\epsilon}(0)(j = 0, \ldots, 2N) \) the linear system
\[ \dot{\epsilon}_0(0) = 0, \]
(4)
\[ 8\dot{\epsilon}_1(0) + 2[\dot{\epsilon}_{2j-1}(0) + \dot{\epsilon}_{2j+1}(0)] - \dot{\epsilon}_{2j-2}(0) - \dot{\epsilon}_{2j+2}(0) = -2h^2, \quad j = 1, \ldots, N-1, \]
\[ \dot{\epsilon}_{2N}(0) = 0. \]
If we eliminate the odd-indexed unknowns from (4) by static condensation, we obtain the linear system
\[ \dot{\epsilon}_0(0) = 0, \]
(5)
\[ \dot{\epsilon}_1(0) - \frac{1}{6}[\dot{\epsilon}_{2j-3}(0) + \dot{\epsilon}_{2j+3}(0)] = -\frac{1}{5}h^2, \quad j = 1, \ldots, N-1, \]
\[ \dot{\epsilon}_{2N}(0) = 0. \]
The matrix of system (5) has the representation
\[ M = I - A, \]
where \( A \) has only nonnegative entries and a spectral radius smaller than 1. From this we can conclude that
\[ M^{-1} = \sum_{n=0}^{\infty} A^n, \]
(7)
which means that \( M^{-1} \) has only nonnegative entries and hence we can conclude that
\[ \dot{\epsilon}_j(0) = -C_i h^2, \quad j = 1, \ldots, N-1, \]
(8)
where \( C_i \) is of \( O(1) \). For \( \dot{\epsilon}_{2j-1} \) we can find a similar expression. All the above formulae now imply that
\[ \epsilon_j(t) = O(th^4), \quad j = 1, \ldots, 2N-1, \]
as \((h, t) \to (0, 0)\). From this, we can prove that
\[ [\epsilon_j(t)h^{-5}]_{\substack{\text{a.h} \to \infty \quad \text{as} \quad t \to 0}} \]
is not bounded, as \( h^{10} \). This proves that the Theorems 1–2 cannot be extended to \( t = 0 \). Since any fourth or higher order quadrature rule yields the same system (2), the same story holds for Theorem 3.

**Remark.** In [2, pp. 171–172], a wrong statement on the uniformity of superconvergence for \( t \geq 0 \) was made, which can also be shown by this example.
REFERENCES

A PROGRAM TO SOLVE ROTATING PLASMA PROBLEMS

M. BAKKER
Stichting Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

and

M.S. VAN DEN BERG *
FOM Instituut voor Atoom en Moleculphysica, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

Received 17 September 1979

PROGRAM SUMMARY

Title of program: PLASMA

Catalogue number: ABVE

Program obtainable from: CPC Program Library, Queen's University of Belfast, Northern Ireland (see application form in this issue)

Computer: Cyber 173-12; Installation: SARA (Stichting Academisch Rekencentrum Amsterdam), Amsterdam

Operating system: NOS/BE

Programming language: FORTRAN IV

High speed storage locations: 128 Kwords

No. of bits in a word: 60

Overlay structure: none

Other peripherals used: card reader, line printer, scratch disc store

No. of cards in combined program and test deck: 1072

Card punching code: 029

Keywords: plasmadynamics, plasmacentrifuge, finite element method

Nature of physical problem
Rotation of a partially ionized gas in a cylinder, caused by crossed electric and magnetic fields.

Method of solution
The system of PDEs is discretized to a symmetric system of linear equations by means of the finite element method using continuous piecewise bilinear functions. The linear system is scaled and solved by means of the conjugate gradient method.

Restrictions on program complexity
A workspace is needed of about 16 times the dimension of the linear system to be solved. The maximal size of the workspace depends on the computer system.

Running time
Proportional to the square of the dimension of the linear problem.

* Now at National Aerospace Laboratory NLR, Anthony Fokkerweg 2, 1059 CM Amsterdam.
LONG WRITE-UP

1. Introduction

When a radial electric current flows in a cylindrical system through an electrically conducting viscous medium in the presence of an externally applied axial magnetic field, the resulting Lorentz force will put the medium into azimuthal motion. Equilibrium is achieved when the Lorentz force balances the opposing viscous force. This principle is applied in case of plasma centrifuges for mass separation of gaseous mixtures.

A calculation model is developed for a system, in which a current flows between two cathodes and a number of anodes. The cathodes are located at the centre of the two endplates closing the system and the ringshaped anodes at a prescribed distance on the cylinder wall. The electrode configuration is symmetrical with respect to the symmetry plane of the cylinder (fig. 1).

The result is a current distribution with a radial, axial and (in the presence of an axial magnetic field) also an azimuthal component. The distribution of the coefficient of viscosity and the electrical conductivity of the incompressible medium is assumed to be uniform.

We define

\[ u = \frac{V_\theta}{U_0}, \quad \phi = \Phi/(U_0 B_r R), \quad u = v/r, \quad x = r/L, \quad y = z/L, \]

\[ \gamma = R/L, \quad Ha = B_x R (\sigma/\eta)^{1/2}, \quad \sigma/\sigma_1 = 1 + \beta^2, \]

where \( V_\theta \) — azimuthal velocity, \( U_0 \) — characteristic velocity, \( \Phi \) — electric potential, \( B_x \) — magnetic induction, \( R, L \) — radius and half length of cylinder, \( \sigma, \sigma_1 \) — normal and tangential components of conductivity, \( \eta \) — dynamic viscosity, \( \gamma \) — inverse aspect ratio, \( Ha \) — Hartmann constant, \( \beta \) — Hall parameter.

Fig. 1. The physical model of the rotating plasma problems.
\[ \frac{3}{x} \frac{\partial u}{\partial x} + \frac{2}{y} \frac{\partial^2 u}{\partial y^2} - \frac{H a^2}{1 + \beta^2} \left( u - \frac{1}{x} \frac{\partial \phi}{\partial x} \right) = 0, \]
\[ \frac{1}{x} \frac{\partial \phi}{\partial x} + \gamma (1 + \beta^2) \frac{\partial^2 \phi}{\partial y^2} - x u_x = 2 u = 0, \quad 0 \leq x, y \leq 1, \]

Boundary conditions (see fig. 2)
\[ \gamma = 0: \frac{\partial u}{\partial x} = \frac{\partial \phi}{\partial x} = 0, \quad \Gamma_2: x = 1: u = 0, \quad \frac{\partial \phi}{\partial x} = f_1(y), \]
\[ \gamma = 0: \frac{\partial u}{\partial y} = \frac{\partial \phi}{\partial y} = 0, \quad \Gamma_3: y = 1: u = 0, \quad \frac{\partial \phi}{\partial y} = f_3(x). \]

As a matter of fact, Van den Berg [1] derives a system of PDEs for \( u \) and \( \phi \). Since, however, the equation is potentially troublesome at \( x = 0 \), it seems numerically preferable to work with (2), we multiply both sides of (2b) with \( x \) and integrate partially over \( R = [0, 1] \times [0, 1] \), it turns out that \( f_1(y) \)
\( i \) are not independent functions but satisfy the integral relation
\[ \gamma > \gamma \int \int_0^1 x f_3(x) \, dx = 0. \]

As the homogeneity of the current from the cathodes, \( f_3(x) \) is a piecewise constant function,
\[ \begin{align*}
\begin{cases}
C_2 > 0, & 0 \leq x < x_0, \\
0, & \text{elsewhere}.
\end{cases}
\end{align*} \]

Rational formulation of the problem

Define the function spaces
\[ \begin{align*}
\text{I} = \{ f: f \text{ continuous on } R; \partial f/\partial x \text{ and } \partial f/\partial y \text{ continuous almost everywhere on } R \}, \\
\text{II} = \{ f: f \in C^1(R), f = 0 \text{ on } \Gamma_2 \text{ and } \Gamma_3 \}.
\end{align*} \]
and the (linear and bilinear) functionals

\[
I(p, q) = \int_R x^3 (p^2 + \gamma^2 p_x^2) \, dx \, dy + \frac{\text{Ha}^2}{1 + \beta^2} \int_R x [(xp - q_x)^2 + \gamma^2 (1 + \beta^2) q_x^2] \, dx \, dy, \quad p \in C_0^3(R), \quad q \in C^0(R),
\]

\[
b(q) = \frac{\text{Ha}^2}{1 + \beta^2} \int_R G(\sigma) q(\sigma) \, d\sigma, \quad q \in C^0(R),
\]

\[
G(\sigma) = \begin{cases} 
0, & \sigma \in \Gamma_1 \cup \Gamma_4, \\
\gamma^2 x \sigma(\sigma)(1 + \beta^2), & \sigma \in \Gamma_1, \\
f(\gamma), & \sigma \in \Gamma_2,
\end{cases}
\]

\[
a_1(p, q) = \int_R x^2 (p_x q_x + \gamma^2 p_y q_y + \frac{\text{Ha}^2}{1 + \beta^2} pq) \, dx \, dy,
\]

\[
a_2(p, q) = \frac{\text{Ha}^2}{1 + \beta^2} \int_R x^2 pq \, dx \, dy,
\]

\[
a_3(p, q) = \frac{\gamma^2}{1 + \beta^2} \int_R x (p_x q_x + \gamma^2 (1 + \beta^2) p_y q_y) \, dx \, dy.
\]

**Theorem 1.** If \((u, \phi)\) is a solution of (2)–(7), it minimizes

\[
E(p, q) = I(p, q) - 2b(q), \quad (p, q) \in C_0^3(R) \times C^0(R),
\]

and satisfies the relations

\[
a_1(u, p) - a_2(u, \phi) = 0, \quad p \in C_0^3(R),
\]

\[
-a_3(u, q) + a_3(\phi, q) = b(q), \quad q \in C^0(R).
\]

Moreover, \(u\) is unique and \(\phi\) is unique up to an additive constant.

**Proof.** If we multiply (2a) with \(x^2 p\) and (2b) with \([x^2(1 + \beta^2)] q\), then after partial integration we obtain (14). By using (14), we also find that

\[
E(p, q) - E(u, \phi) = I(u - p, \phi - q)\geq 0, \quad p \in C_0^3(R), \quad q \in C^0(R),
\]

which proves that \(E(u, \phi)\) is a minimum. It is not a strong minimum, for \(I(u - p, \phi - q) = 0\), if and only if \(u \equiv p\) and \(\phi - q \equiv \text{constant}.

**Remark.** One can make \(\phi\) unique by introducing an additional constraint. For technical reasons, we define \(\phi(0, 1) = 0\).

3. Finite element solution

It is a standard result (see e.g. Mitchell and Wait [3] or Courant and Hilbert [2]) that one can approximate \((u, \phi)\) by minimizing \(E(p, q)\) over a finite-dimensional subspace \(S_p \times S\) of \(C_0^3(R) \times C^0(R)\). For \(S\), we select the space of piecewise bilinear functions, given a partition of \(R\) in rectangles \([x_{i-1}, x_i] \times [y_{j-1}, y_j]\) (see e.g. Strang and Fix [5] or Mitchell and Wait [2]). This partition should be given by partitions of the \(x\)-interval and the
The minimization of $E$ over the finite element space results in the system of linear equations

$$
\begin{pmatrix}
A_1 & -A_2 \\
-A_2^T & A_3 \\
\end{pmatrix}
\begin{pmatrix}
\mathbf{U} \\
\mathbf{c} \\
\end{pmatrix} = 
\begin{pmatrix}
0 \\
1 \\
\end{pmatrix}
$$

(15)

with

$$A_1 = \langle \phi_1(\mathbf{B}_1, \mathbf{B}_2) \rangle, \quad A_2 = \langle \phi_2(\mathbf{B}_1, \mathbf{B}_2) \rangle, \quad A_3 = \langle \phi_3(\mathbf{B}_1, \mathbf{B}_2) \rangle, \quad \mathbf{c} = \langle \mathbf{c}(\mathbf{B}_i) \rangle,$$

(16)

where $\langle \mathbf{B}_1 \rangle$ and $\langle \mathbf{B}_2 \rangle$ are the basis functions of $S_0$ and $S$, respectively, and where $a_1, a_2, a_3$ and $b$ are given by $C = (B_2). A_1, A_2$ and $A_3$ are 9-diagonal, while $A_1$ and $A_3$ are also symmetric; hence the overall matrix is symmetric and 27-diagonal.

Further trimming of the matrix

System (15) can be made still sparser if the energy functional $E$ is approximated by a suitable quadrature rule.

On any rectangle $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$, we define

$$
\int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} z^m \mathbf{F}(x, y) \, dx \, dy = \omega_1 \mathbf{F}(x_{i-1}, y_{j-1}) + \omega_2 \mathbf{F}(x_{i-1}, y_j) + \omega_3 \mathbf{F}(x_i, y_{j-1}) + \omega_4 \mathbf{F}(x_i, y_j),
$$

(17)

where the weights $\omega_1, \ldots, \omega_4$ are chosen such that (17) is exact, if $\mathbf{F}$ is bilinear on $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$. If we use (17) to approximate the integral $I$ defined by (7), it is consequently used to approximate the entries of the matrices $A_1, A_2$ and $A_3$. The result is a further trimming of these matrices; $A_1$ and $A_3$ are reduced to penta-diagonal matrices and $A_3$ even to a tri-diagonal one. The overall reduction is from 27 to 11 diagonals. The vanishing of so many matrix entries is due to the fact that the corresponding integrands of (16) vanish on all the grid-points $(x_i, y_j)$, hence (17) yields a zero value. It can be proved that the order of accuracy of the finite element solution is not affected by using (17) (see ref. [5], ch. IV).

Solution of the linear system

The linear system is solved by means of the conjugate gradient method (CG method). This iterative method (see Reid [4]) is very suitable for symmetric nonnegative definite systems of sparse structure.

4. Test-example

In the test-problem $f_1(y)$ and $f_2(x)$ are piecewise constant functions defined by

$$f_1(y) = \begin{cases} 
C_1, & 0.25 \leq y \leq 0.3125, \\
0, & \text{elsewhere}, 
\end{cases}, \quad f_2(x) = \begin{cases} 
C_2, & 0 \leq x \leq 0.0625, \\
0, & \text{elsewhere}, 
\end{cases}$$

From the value $\int_0^1 f_1(y) \, dy$, it turns out that $C_1$ and $C_2$ have the values $-0.23944$ and $0.67188$, respectively.

For plotting the graphs of $v$ and $\phi$ (see figs. 3 and 4), the library DISSIPAL was used. Calls to this library have, however, been omitted so that the user can supply his own plotting routines. We see that near the cathodes both $v$ and $\phi$ have boundary layers. They also have boundary layers near the anode but less significant ones.

5. Program description

In this section, we give a brief description of the subroutine PLASMA and the other subprograms. For a more detailed description, we refer to the comment lines in the software package.
PLASMA — this is the driving subroutine which reads the input data, solves the problem and enables the user to manipulate the output data. See also next section and fig. 5.

EVAL — this subroutine computes the matrix and the right hand side of problem (15).

SCALE — this subroutine scales system (15) to

\[ x = D y, \quad S y = D A D y = D b, \]

where \( D \) is a diagonal matrix whose entries are given by

\[ d_i = \sigma_i^{-1/2}; \]

CONGR — this subroutine solves (18) iteratively by the CG method.

MATVEC — this subroutine computes the matrix-vector product \( S u \), given a vector \( u \).

INPROD — this subroutine computes the inner product of two vectors.

FATAL — this subroutine terminates the program if inconsistencies of the input record are discovered, or if the workspace (see section 6) is not large enough.

F1 — in this subprogram, \( f_1 \) is assigned its value. Only the profile is to be given, i.e. the function value, up to a
Fig. 4. Graph of $e(x, y)$.

Fig. 5. Hierarchical structure of subprograms.
constant factor. This factor is computed from the value of

\[ Q = \int_{0}^{1} f_{1}(y) \, dy \]  \hspace{1cm} (19)

which is to be given on the input record.

OUT - in this subroutine, the user is allowed to manipulate the input and output data, as he desires. A default subroutine is given in the program.

6. Workspace

PLASMA has a work-array of dimension NWORK as formal parameter. NWORK should be at least 31 MN, where M and N are the respective lengths of the x-grid and y-grid.

7. The input record

The following parameters have to be read in:

1) \( Q, \alpha, \gamma, H_{a}, x_{0} \) in FORMAT (F12.4),
2) \( M \) in FORMAT (I4),
3) \( x_{1}, ..., x_{M} \) in FORMAT (F12.4),
4) \( N \) in FORMAT (I4),
5) \( y_{1}, ..., y_{N} \) in FORMAT (F12.4),
6) \( ND \) in FORMAT (I4),
7) \( y_{d1}, ..., y_{dND} \) in FORMAT (F12.4),

where \( Q, \alpha, \gamma, H_{a} \) and \( x_{0} \) are given by (19), (1) and (17), respectively, and \( M, N \) and \( ND \) are the number of x-points, the number of y-points and the number of discontinuity points \( f_{1}(y) \), respectively. See fig. 6 for the parameters used in the test run.

The user has to take care of the following:

a) the x-grid and y-grid have to be strictly monotone and, of course, \( x_{1} \) and \( y_{1} \) should be zero, while \( x_{M} \) and \( y_{N} \)

<table>
<thead>
<tr>
<th>x-coordinates</th>
<th>y-coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.004</td>
<td>1.005</td>
</tr>
<tr>
<td>1.006</td>
<td>1.007</td>
</tr>
<tr>
<td>1.008</td>
<td>1.009</td>
</tr>
<tr>
<td>1.010</td>
<td>1.011</td>
</tr>
</tbody>
</table>

Fig. 6. Input parameters.
should be one. If not, the program is terminated and a message of the reason is printed;
b) the array of discontinuity points $\psi$ should contain at least one point, even if $f_i(\psi)$ is continuous. In this case, the user is advised to give zero or one. Furthermore, each of the discontinuity points should occur in the $y$-grid.
If not, the program is terminated with a message of the reason;
c) the $x$-grid should contain $x_0$, otherwise the program is terminated too;
d) around the discontinuity points of $f_1(\psi)$ and $f_2(\psi)$, the user is advised to refine the grid. This especially holds for $x_0$, since there the jump of $\delta(\psi)$ is generally larger than near the discontinuity points of $f_1(\psi)$. This may be illustrated by fig. 4. However, he should take care that the rectangles of the partition do not become too "lean", i.e. the ratio of length and width should remain reasonable. For a discussion, we refer to ref. [5].

8. The common blocks

8.1. COMMON/PRBLM/

GAMSO $\gamma^2$,
GAMSQB $\gamma^2(1 + \beta^2)$,
HASQ $\text{Ha}^2/(1 + \beta^2)$,
NXO grid number of $x_0$: $X(NXO) = x_0$,

AA value of $(1 + \beta^2) \int_0^1 f_i(\psi) \, d\psi$,

XO $x_0$,

FAC scaling factor of $f_i(\psi)$.

8.2. COMMON/WKSP/

Except the three last members, this block contains integer pointers, locating the arrays within the global array WORK.

MATRIX — locates the matrix of the system
WORK(MATRIX + 3*L - 2) — Lth nonzero entry,
WORK(MATRIX + 3*L - 1) — its row number,
WORK(MATRIX + 3*L) — its column number, L = 1, ..., NEL/3;
NX — locates the $x$-grid
WORK(NX + L) — Lth $x$-point, L = 1, ..., M;
NXY — locates the $y$-grid
WORK(NXY + L) — Lth $y$-point, L = 1, ..., N;
NFYD — locates the jumping points of $f_1(\psi)$
WORK(NFYD + L) — Lth jumping point, L = 1, ..., ND;
NFR, NSC, NG, NH — locate auxiliary arrays of dimension 2MN;
NU — locates the approximate solution of (2)
WORK(NU + (J-1)*M + 1) — approximation of $\phi$ in $(x(I), y(J))$,
WORK(NU + MN + (J-1)*M + 1) — approximation of $\phi$ in $(x(I), y(J))$;
NEL — three times the number of nonzero entries in the strict upperdiagonal part;
NBBOUND — the actual dimension of workspace needed;
NWORK — the dimension of WORK.
Acknowledgements

The authors wish to thank Professor J. Kistemaker for his assistance and Mr. F. Vitalis for his careful reading of the manuscript. This work is part of the research program of the Stichting voor Fundamenteel Onderzoek der Materie (Foundation for Fundamental Research on Matter) and was made possible by financial support from the Nederlandse Organisatie voor Zuiver Wetenschappelijk Onderzoek (Netherlands Organization for the Advancement of Pure Research).

References

A PROGRAM TO SOLVE A SOLUTE DIFFUSION PROBLEM WITH SEGREGATION AT A MOVING INTERFACE

M. BAKKER
Stichting Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

and

D. HOONHOUT
FOM-Instituut voor Atoom- en Moleculijjystica, Kruislaan 407, 1098 DB Amsterdam, The Netherlands

Received 19 December 1980

PROGRAM SUMMARY

Title of program: DIFSEG
Catalogue number: ABVZ
Program available from: CPC Program Library, Queen's University, Belfast, N. Ireland (see application form in this issue)
Installation: SARA (Academic Computing Centre Amsterdam)
Plotting library used: CALCOMP, available at SARA
Operating system: NOS/BE
Programming language: FORTRAN IV
High speed storage requested: 128 K
No. of bits in a word: 60
Overlay structure: none
Other peripherals used: card reader, line printer, plotter and scratch disk store (the last one only if the plotter is used)

No. of cards in combined program and test deck: 911

Key words: silicon, laser-annealing, solute-segregation, diffusion equation, finite difference method, moving boundary problem

Nature of the physical problem
The diffusion and segregation of impurities implanted in silicon which accompany the annealing of implantation damage by means of pulsed-laser irradiation.

Method of solution
The diffusion equation is semidiscretized by the finite difference method and the resulting ODE is integrated by an adapted Runge-Kutta method.

Running time
Roughly proportional to the dimension of the ODE.

Unusual features
The user should have no difficulty in replacing the CALCOMP plotting routines by equivalent routines at his own installation, using figs. 4-7 as examples.

LONG WRITE-UP

1. Introduction

The introduction of dopant atoms in silicon is normally accomplished via diffusion or via ion implantation. In the latter case, a monoenergetic beam of ionized dopant atoms, accelerated to an energy in the range 10–500 keV,
is directed onto a monocrystalline silicon target. Thus the dopant atoms penetrate the silicon to depths in the order of 10–500 nm in a well controlled manner. However, ion implantation introduces a high degree of disorder in the surface layer of the silicon crystal. Therefore, an annealing procedure is needed to restore the crystallographic structure of the implanted layer and to make the implanted dopant atoms electrically active.

Traditionally, this is done by a prolonged thermal treatment at 700–1100°C. A new, promising approach to the annealing problem is the irradiation of the implanted silicon with a pulsed high-power laser, mostly a Q-switched ruby- or Nd:Yag-laser. It is found that after pulsed-laser annealing the crystallinity of the implanted layer is restored, but the depth distribution of the dopant has changed drastically [3,4]. One model to explain this is laser-induced melting of the implanted layer followed by rapid resolidification during cool-down. In this picture, the high diffusivity in liquid silicon, and the segregation of dopant atoms at the moving solidification front are held responsible for the changes in the dopant depth distribution.

This paper describes a numerical solution of a mathematical model for solute diffusion in the liquid and segregation at the phase boundary. A comparison between this model and experimental dopant profiles, measured after pulsed ruby-laser annealing has been published elsewhere [2,3].

In previous computer models, presented by Baer et al. [1] and White et al. [7], segregation and diffusion in the liquid are represented as separate steps, although it is probably more realistic to have both effects interact continuously. It is not clear in these models which boundary conditions at the walls separating the liquid from the solid and from the vacuum are used to solve the diffusion equation. Furthermore, it is not clear how the width of the intervals, in which the depth scale is divided, affects the results. These questions are answered explicitly for our computer model.

As the material is implanted and laser-irradiated over an area of the order of a few cm², which is extremely large compared to the depth range of the implantation, we can describe the physical process as a one-dimensional problem. The mathematical model can be described as follows. Let \( C(x, t) \) be the depth distribution of the dopant as function of the depth \( x \) and the time \( t \). Prior to laser-annealing, \( C(x, t) \) can be approximated by a Gaussian distribution with its peak at depth \( R_p \) and with standard deviation \( \Delta R_p \). Due to the laser-pulse, a melt front moves back from \( x = a \) to \( x = 0 \) at a velocity \( u \) which is assumed to be constant. Thus the position of the front is given by \( g(t) = a - ut \). At the resolidification front, the dissolved dopant atoms segregate in such a way that the ratio of the concentration in liquid and solid in the immediate vicinity of the front is equal to \( k \), the interfacial distribution coefficient. It is assumed that \( k \) is constant during the process while we will only consider cases for which \( k < 1 \). In the liquid phase, the dopant atoms diffuse according to a diffusion coefficient \( D \). On the time-scale of pulsed-laser annealing, solute diffusion in the solid can be neglected.

Thus, the calculation of the dopant depth distribution \( C(x, t) \) after completion of the process boils down to the solution of the one-dimensional diffusion equation

\[
\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2} , \quad 0 < x < g(t) , \quad (1.1)
\]

with

\[
g(t) = a - ut . \quad (1.2)
\]

The absence of diffusion in the solid phase implies

\[
\frac{\partial C}{\partial t} = 0 ; \quad g(t) < x < a . \quad (1.3)
\]

The boundary, initial and interface conditions are the following:

a) for \( t = 0 \), the depth distribution for all \( x \) in the region \( 0 < x < a \) is a Gaussian function with its maximum at \( x = R_p \) and a standard deviation \( \Delta R_p \). In formula

\[
C(x, 0) = \exp[-((x - R_p)/\Delta R_p)^2/2] ; \quad (1.4)
\]
b) no mass transport across the vacuum—liquid interface is allowed:

\[
\frac{\partial C}{\partial x} = 0 \quad x = 0, \quad t > 0
\]  \quad (1.5)

c) the segregation taking place at the resolidification front at \( x = g(t) \) is expressed as

\[
\lim_{x \to g(t)} C(x, t) = k \lim_{x \to g(t)} C(x, t), \quad t > 0
\]  \quad (1.6)

As a measure of segregation we compute the so-called surface fraction \( F(s) \) defined by

\[
F(s) = \int_0^{s} C(x, t) \, dx / \int_0^{\infty} C(x, t) \, dx
\]  \quad (1.7)

\( s < a, \quad t \geq (a - s)/v \) \quad (1.8)

i.e. we compute the amount of mass in the region \([0, s]\) after resolidification. In this paper, we always chose the values

\[
s = 240 \, \text{Å}, \quad a = 3000 \, \text{Å}, \quad v = 4 \, \text{m/s}
\]  \quad (1.9)

but the user is enabled to give other values on the input record.

2. Numerical analysis of the problem

In this chapter we discuss the several difficulties and pitfalls predictable or not which are met during the process of numerically solving the problem. To that end, we first scale the problem by the introduction of the dimensionless variables

\[
\rho = \frac{x}{a}, \quad \alpha = \frac{D}{at}, \quad \tau = \frac{D}{at}, \quad \mu = \frac{R_e}{a}, \quad \sigma = \frac{\Delta R_e}{a}
\]  \quad (2.1)

Scaling of (1.1)–(1.6) by (2.1) leads to the dimensionless problem (with \( c(\rho, \tau) = C(\rho a, a^2 + D^{-1}) \))

\[
\frac{\partial c}{\partial \tau} = \begin{cases} 0, & 1 - \alpha r < \rho < 1, \quad 0 \leq \tau \leq a^{-1}, \\ \frac{\partial^2 c}{\partial \rho^2}, & 0 < \rho < 1 - \alpha r, \end{cases}
\]  \quad (2.2a)

with boundary conditions

\[
\frac{\partial c}{\partial \rho} = 0, \quad \rho = 0; \quad \lim_{\rho \to 1 - \alpha r} c = k \lim_{\rho \to 1 - \alpha r} c, \quad \tau > 0
\]  \quad (2.2b)

and initial conditions

\[
c(\rho, \tau) = \exp(-\frac{1}{2}[(\rho - \mu)/\alpha]^2)
\]  \quad (2.2c)

2.1. Derivation of a mixed boundary condition at the interface

From a mathematical point of view, (2.2) is incomplete since the boundary and initial conditions (2.2b) and (2.2c) are insufficient. However, we can use the additional external condition

\[
\int_0^1 c(\rho, \tau) \, d\rho = \int_0^1 c(\rho, 0) \, d\rho, \quad \tau \geq 0
\]  \quad (2.3)
which can physically be interpreted as the conservation of mass during the resolidification process. This means that

\[
\frac{d}{dt} \int_0^t c(\rho, \tau) \, d\rho = 0, \quad \tau \geq 0.
\] (2.4)

Further elaboration of (2.4) yields, if we substitute \( g(\tau) = 1 - \alpha \tau \), and use (2.2a) and (2.2b)

\[
\frac{d}{dt} \left[ \int_0^t g(\tau) c(\rho, \tau) \, d\rho + \int_{\gamma(t)} \left( c(\rho, \tau) \, d\rho \right) \right]
\]

\[
= \frac{d}{dt} \lim_{\rho \to \gamma(t)} c(\rho, \tau) + \int_0^t \frac{d}{dt} \left( \frac{\partial c}{\partial \rho} (\rho, \tau) \right) \, d\rho - \frac{d}{dt} \left( \int_{\gamma(t)} \frac{\partial c}{\partial \rho} (\rho, \tau) \, d\rho \right) + \int_{\gamma(t)} \frac{d}{dt} \left( c(\rho, \tau) \right) \, d\rho
\]

\[
= \frac{d}{dt} (1 - k) \lim_{\rho \to \gamma(t)} c + \int_0^t \frac{d}{dt} \left( \frac{\partial^2 c}{\partial \rho^2} (\rho, \tau) \right) \, d\rho = \frac{d}{dt} (1 - k) \lim_{\rho \to \gamma(t)} c + \lim_{\rho \to \gamma(t)} \frac{dc}{d\rho} = 0.
\] (2.5)

Since \( g(\tau) = 1 - \alpha \tau \), we have the mixed boundary condition

\[
\lim_{\rho \to \gamma(t)} \frac{dc}{d\rho} = \alpha(1 - k) \lim_{\rho \to \gamma(t)} c,
\] (2.2d)

which makes the initial boundary problem complete.

2.2. Semidiscretization in the space variable

We divide the interval \([0, 1]\) in \(N\) segments of equal length (see fig. 1).

At time \(\tau\), the interface is situated in the interval \([\rho_i, \rho_{i+1}]\) with

\[
\rho_i = h(i - 1), \quad i = 1, \ldots, N + 1 \quad h = 1/N, \quad M = 1 + N(1 - \alpha)\] (2.6)

Let \( c(\rho_i, \tau) = \rho_{i+1}(\tau) \), \( i = 1, \ldots, N + 1 \).

For \( i = 1, \ldots, M - 1 \), we use the standard finite difference approximation [2] for (2.2a)

\[
\frac{dc_i}{d\tau} = c_{i-1} - 2c_i + c_{i+1}, \quad i = 1, \ldots, M - 1, \quad \frac{dc_i}{d\tau} = \frac{2}{h^2} (c_{i-1} - c_i).
\] (2.7a)

For \( i > M \), it is also simple:

\[
\frac{dc_i}{d\tau} = 0, \quad i > M.
\] (2.7b)

![Graph of c(\rho, \tau).](image)
For $i = M$, approximation of $\frac{dc_M}{dr}$ is more problematic, since $c$ is discontinuous on $[\rho_M, \rho_{M+1}]$, see fig. 2.

In order to get a suitable approximation of $\frac{\partial^2 c}{\partial \rho^2}$ at $\rho = \rho_M$, we put

$$\delta = 1 - \alpha r - \rho_M, \quad \frac{\partial^2 c}{\partial \rho^2} (\rho_M, r) = \omega_1 c(\rho_{M-1}, r) + \omega_2 c(\rho_M, r) + \omega_3 c(\rho_M + \delta, r),$$  \hspace{1cm} (2.7c)

where $\omega_1, \omega_2$ and $\omega_3$ are chosen in such a way that (2.7c) has maximum accuracy, hence we find

$$\omega_1 = \frac{2}{h(h + \delta)}, \quad \omega_2 = \frac{2}{\delta(h + \delta)}, \quad \omega_3 = -\omega_1 - \omega_2, \quad 0 < \delta < h.$$  \hspace{1cm} (2.8)

The crucial question is: how to compute $c(\rho_M + \delta, r)$? We do this by extrapolation. Let $c_{\text{int}}$ be the (unknown) approximation of $c(\rho_M + \delta, r)$. On the interval $[\rho_{M-1}, \rho_M + \delta]$ we define a parabola which interpolates $c(\rho, r)$ at the points $\rho_{M-1}, \rho_M$ and $\rho_M + \delta$, and which satisfies (2.2d). In formula

$$c(\rho, r) \approx c(\rho, r) = \frac{(\rho - \rho_M)(\rho - \rho_{M-1})}{h(h + \delta)} c(\rho_{M-1}, r) + \frac{(\rho - \rho_{M-1})(\rho_M + \delta - \rho)}{h\delta} c(\rho_M, r)$$

$$+ \frac{(\rho - \rho_M)(\rho - \rho_{M-1})}{h(h + \delta)} c_{\text{int}}, \quad \rho \in [\rho_{M-1}, \rho_M + \delta].$$  \hspace{1cm} (2.9)

$$\frac{\partial c}{\partial \rho} = \alpha(1 - k) c_{\text{int}}, \quad \rho = \rho_M + \delta.$$  

Elaboration of (2.9) yields

$$c_{\text{int}} \approx \frac{c(\rho_M, r) - \beta c(\rho_{M-1}, r)}{[1 - \beta^2 - \alpha(1 - k) 3h]}, \quad \beta = \delta/(h + \delta).$$  \hspace{1cm} (2.10)

At the other hand, if $\delta \to h$, it is better to approximate $c_{\text{int}}$ by linear extrapolation of $c(\rho_{M+1}, r)$ and $c(\rho_{M+2}, r)$ provided, of course, that $M < N - 1$. This means that

$$c_{\text{int}} \approx \frac{[(2 - \delta/h) c(\rho_{M+1}, r) - (1 - \delta/h) c(\rho_{M+2}, r)]}{k}.$$  \hspace{1cm} (2.11)

Numerical experiments have shown that (2.10) should be applied if $\beta < 4/9$ and (2.11) otherwise.

Summarizing, we found for $\frac{dc_M}{dr}$ the formula

$$\frac{dc_M}{dr} = \begin{cases} \frac{2}{h + \delta} \left[ \frac{c_{M-1} - c_M}{h} + \frac{c_{\text{int}} - c_M}{\delta} \right], & \delta > 0, \\ \frac{2}{h} \left[ \frac{c_{M-1} - c_M}{h} + \alpha(1 - k) c_M \right], & \delta = 0, \end{cases}$$

where $c_{\text{int}}$ is defined by (2.10) or (2.11).

**Remark.** In earlier versions of this program, $c_{\text{int}}$ was approximated by linear interpolation between $c_M$ and $c_{M+1}/k$.

![Fig. 2. Graph of $c(\rho_M, r)$](image-url)
This approach was rather unsatisfactory, since the behaviour of \( c(\rho, \tau) \) sometimes differs left and right of the interface (monotonically increasing to the left and monotonically decreasing to the right, see fig. 2) which resulted in a systematical underestimation of \( c_{\text{int}} \). Because of this underestimation, relation (2.3) refused to remain valid, especially as \( \alpha \) was large and \( k \) small, the mass increased by about 15\%, and much more if \( N \) was small.

2.3. Boundary layers, choice of \( N \)

It turns out that \( c(\rho, \tau) \) has a boundary layer left of \( 1 - \alpha \tau \), if \( \alpha \) is large and \( k \) small. Physically, this can be interpreted as a high concentration of mass near the resolidification front due to slow mass transport and strong segregation. One is forced to choose \( N \) large in this case. From (2.10) we can derive a minimum value of \( N \) in order to keep the denominator of (2.10) positive, i.e.

\[
N > \alpha(1 - k) \beta(1 - \beta^2), \quad 0 < \beta < \frac{1}{2}.
\]

This is the minimum value of \( N \). In practice \( N \) should be much larger, certainly if \( k \) is small and \( \alpha \) large; a safe choice would be

\[
N = \alpha/5k.
\]

2.4. Integration of semidiscretized problem

In this section, we discuss the organisational problems which are encountered if we want to solve (2.7) by an ODE integrator.

As we saw in section 2.2, we can consider (2.7) as a system of ODEs with dimension decreasing in time

\[
M(\tau) = \text{entire}(1 - \alpha \tau) + N + 1, \quad 0 < \tau < \alpha^{-1}.
\]

(2.14)

If we treat (2.7) as a system of fixed length by putting \( de/d\tau = 0, i > M \) and let a common ODE integrator loose on it, we run into trouble and the output results are rubbish.

The explanation is simple. Let \( \{\tau_i\}_{i=1}^N \) be a set of time-points defined by

\[
\tau_i = (i - 1)aN = \Delta \tau(i - 1), \quad i = 1, ..., N.
\]

If \( \tau_1 < \tau < \tau_{i+1} \), the de facto dimension of (2.7) is \( N + 1 - i \). If \( \tau \) passes \( \tau_{i+1} \), the dimension of (2.7) is diminished by one, \( c_{N+1-i} \) is multiplied by \( k \) and kept constant forever, etc. An ODE integrator may choose step-sizes of the time-step which exceed \( \Delta \tau \) by an order of magnitude. By this choice several points \( \tau_i \) are passed in one integration step. This is an invitation to disaster, which duly materializes. Hence we impose:

**Condition 1.** Any ODE integrator should hop from one time-point \( \tau_i \) to the next after which \( c_{N+1-i} \) is multiplied by \( k \) and kept constant.

The above precaution alone is insufficient because many ODE integrators do not really hit the end-point of the integration interval but go on until it is passed and then interpolate between the last and previous time-points. This also leads to disaster so we impose:

**Condition 2.** The ODE integrator should hit the end-point exactly. In formula, at time \( \tau \), 

\[
\text{STEPSIZE} = \min(\Delta^* \tau, \tau_{i+1} - \tau),
\]

where \( \Delta^* \tau \) is a step-size computed on the basis of accuracy, stability and possibly other motives as well.

Because most ODE integrators do not satisfy condition 2, we took a house ODE integrator which uses explicit one-step multipoint Runge–Kutta–Chebyshev methods of second-order consistency and applies automatic step-size control based upon accuracy and stability [5]. With a slight alteration condition 2 could be implemented in the package RKC. This delivered very satisfactory results.
Remark. In an earlier stage, we worked with the forward Euler method [6, p. 316] with constant time-step \( \Delta \tau = (\alpha a)^{-1} \), where \( m = 1 + 2N/a \); in formula

\[
c_j(\rho, \tau + \Delta \tau) = c_j(\rho, \tau) + \Delta \tau \frac{d}{d\tau} c_j(\rho, \tau).
\]

(2.15)

This simple algorithm satisfied the two above conditions. It was, however, too inaccurate, which was proved when the quotient

\[
Q(\tau) = \int_0^1 c(\rho, \tau) d\rho \int_0^1 c(\rho, 0) d\rho
\]

(2.16)

was computed. According to (2.3) \( Q \) should be 1, but as \( \tau \to \alpha^{-1} \), \( Q \) took unacceptable values, especially if \( \alpha \) was large and \( \kappa \) small. So we had to use a more sophisticated integrator. Experiments with RKC showed that the deviation of \( Q \) from unity was of the same order of magnitude as the relative tolerance for the time-integration.

3. Description of the program

In this section, we give a brief description of the subroutine DRIVE and other subprograms (see fig. 3). For a more detailed description, we refer to the comment lines in the software package.

MCFOM  This is the main program to be written by the user; it should contain a blank common block of at least \( 7N + 13 \) words, where \( N \) is the number of segments; see section 2; its only statement is a call of DIFSEG with the dimension of the blank common block as actual parameter;

DIFSEG  This subroutine reads and prints the input record and calls DRIVE if the work-space is large enough; otherwise it terminates the program printing a message of the reason;

DRIVE  This is the managing subroutine; it initializes \( c(\rho, 0) \) at the mesh-points, integrates (2.7) until the interface has passed 240 Å; after the integration, it computes the surface fraction and \( Q(\tau) \); if desired, it plots \( C(x, \tau) \) a couple of times together with a graph of the solid state;

RKC  This routine and its auxiliary routines perform the time-integration;

DER  This routine evaluates the vector \( dc/dr \), given an input vector \( c \) at time \( \tau \) according to (2.7). It is called by RKC and its auxiliary routines;

SPECTR  This function assigns the value of the spectral radius of the Jacobian matrix \( (dc/dc) \). Knowledge of this value is essential to numerically stable time-integrations [5];

UNIT  A function subroutine to be implemented by the user, which gives \( C(x, 0) \) its value for given \( x \).

Fig. 3. Organization scheme of the program.
4. The input record

See the comment lines in subroutine DIFSEG.

5. Workspace

In the main program a blank common block of length at least $7N + 13$ should be declared. This length should also be the actual parameter of DIFSEG.
6. Common blocks

See the comment lines in subroutine DIFSEG.

7. Test-examples

The program was run for two sets of input values (see table 1) and a plot was drawn (see figs. 4–7). We see the curves become steeper and steeper as the resolidification front moves to the left.
Table 1
Input records for test examples

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>240 A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>$1 \times 10^{-4}$ cm²/s $5 \times 10^{-6}$ cm²/s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>3000 A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>4 m/s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>500</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Plot? yes
k 0.05 0.1
$R_p$ 1100 A
$\Delta R_p$ 500 A
Acknowledgements

The authors wish to thank Prof. Kistemaker of Stichting FOM and Prof. Basyen of Stichting Mathematisch Centrum for their assistance. This work was made possible by Stichting FOM, Stichting Mathematisch Centrum and Stichting ZWO.

References

### TEST RUN OUTPUT

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D)</td>
<td>0.05</td>
</tr>
<tr>
<td>Diffusion Coefficient in c****/sec</td>
<td>0.65 \times 10^{-3}</td>
</tr>
<tr>
<td>Surface Depth in Angstrom</td>
<td>240</td>
</tr>
<tr>
<td>Melting Depth in Angstrom</td>
<td>3000</td>
</tr>
<tr>
<td>Number of Segments in which ([0,1]) is divided</td>
<td>50</td>
</tr>
<tr>
<td>Speed of re-solidification front in m/sec</td>
<td>4.00</td>
</tr>
<tr>
<td>Mean Value in Angstrom</td>
<td>1100</td>
</tr>
<tr>
<td>Sigma in Angstrom</td>
<td>500</td>
</tr>
<tr>
<td>Surface Fraction</td>
<td>0.975</td>
</tr>
<tr>
<td>(m_{\text{total}}/m_{\text{solid}})</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D)</td>
<td>0.10</td>
</tr>
<tr>
<td>Diffusion Coefficient in c****/sec</td>
<td>0.65 \times 10^{-3}</td>
</tr>
<tr>
<td>Surface Depth in Angstrom</td>
<td>240</td>
</tr>
<tr>
<td>Melting Depth in Angstrom</td>
<td>3000</td>
</tr>
<tr>
<td>Number of Segments in which ([0,1]) is divided</td>
<td>50</td>
</tr>
<tr>
<td>Speed of re-solidification front in m/sec</td>
<td>4.00</td>
</tr>
<tr>
<td>Mean Value in Angstrom</td>
<td>1100</td>
</tr>
<tr>
<td>Sigma in Angstrom</td>
<td>500</td>
</tr>
<tr>
<td>Surface Fraction</td>
<td>0.975</td>
</tr>
<tr>
<td>(m_{\text{total}}/m_{\text{solid}})</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D)</td>
<td>0.05</td>
</tr>
<tr>
<td>Diffusion Coefficient in c****/sec</td>
<td>1.0E-03</td>
</tr>
<tr>
<td>Surface Depth in Angstrom</td>
<td>240</td>
</tr>
<tr>
<td>Melting Depth in Angstrom</td>
<td>3000</td>
</tr>
<tr>
<td>Number of Segments in which ([0,1]) is divided</td>
<td>50</td>
</tr>
<tr>
<td>Speed of re-solidification front in m/sec</td>
<td>4.00</td>
</tr>
<tr>
<td>Mean Value in Angstrom</td>
<td>1100</td>
</tr>
<tr>
<td>Sigma in Angstrom</td>
<td>500</td>
</tr>
<tr>
<td>Surface Fraction</td>
<td>0.975</td>
</tr>
<tr>
<td>(m_{\text{total}}/m_{\text{solid}})</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D)</td>
<td>0.10</td>
</tr>
<tr>
<td>Diffusion Coefficient in c****/sec</td>
<td>1.0E-03</td>
</tr>
<tr>
<td>Surface Depth in Angstrom</td>
<td>240</td>
</tr>
<tr>
<td>Melting Depth in Angstrom</td>
<td>3000</td>
</tr>
<tr>
<td>Number of Segments in which ([0,1]) is divided</td>
<td>50</td>
</tr>
<tr>
<td>Speed of re-solidification front in m/sec</td>
<td>4.00</td>
</tr>
<tr>
<td>Mean Value in Angstrom</td>
<td>1100</td>
</tr>
<tr>
<td>Sigma in Angstrom</td>
<td>500</td>
</tr>
<tr>
<td>Surface Fraction</td>
<td>0.975</td>
</tr>
<tr>
<td>(m_{\text{total}}/m_{\text{solid}})</td>
<td>1.000</td>
</tr>
</tbody>
</table>
SAMENVATTING

Dit proefschrift bestaat uit een inleiding gevolgd door vijf artikelen die in wetenschappelijke tijdschriften zijn verschenen. Al deze artikelen behandelen de numerieke oplossing van elliptische en parabolische problemen door middel van de Methode der Eindige Elementen (EEM).

De EEM is in de jaren vijftig tot ontwikkeling gebracht door ingenieurs in de ruimtevaart. Deze ingenieurs pasten de EEM toe bij het numeriek oplossen van problemen in de sterkteleer. Pas in de tweede helft der jaren zestig werd, mede dankzij de ingenieurs Argyris, Clough en Zienkiewicz, duidelijk dat de EEM een moderne toepassing was van een oude approximatiemethode: de methode van Ritz-Galerkin. Tevens bleek dat de theorie der Sobolev-ruimtes zich uitstekend leende voor de wiskundige onderbouw van de EEM. De EEM komt op het volgende neer:

- het definitiegebied van de te berekenen functie wordt opgedeeld in kleine partijen, ook wel genaamd segmenten; voorbeelden zijn intervallen, driehoeken, viervlakken, etc.;
- op ieder afzonderlijk segment wordt de functie benaderd door een polynoom van vaste graad;
- van de benadering wordt geëist, afhankelijk van het op te lossen probleem, dat zij continu, differentieerbaar, etc. is op het gehele gebied en dat de hogere (partiële) afgeleiden op de randen van de segmenten discontinu zijn.

Bovenstaande benaderingstechniek heeft een aantal prettige eigenschappen, zoals robuustheid, nauwkeurigheid en efficiëntie. Bovendien leidt ze veelal tot grote stelsels algebraische vergelijkingen of (impliciete) stelsels gewone differentiaalvergelijkingen, waarvan de bijbehorende matricies (de bandmatrices zijn die systematisch opgebouwd worden.

In de eerste drie artikelen [A-C] wordt aandacht besteed aan de numeriek oplossen van elliptische en parabolische problemen met één ruimtevariabele. De volgende onderwerpen komen aan de orde:

(a) de extragrote nauwkeurigheid van de EEM-oplossing op specifieke punten (superconvergentie), namelijk op de randen van de segmenten [A-C] en op de nulpunten van bepaalde verschoven Jacobi-polynomen [B-C];
(b) de constructie van eenvoudige approximaties van u(x,0) voor parabolische problemen [A,C];
(c) de invloed van numerieke kwadratuur op de nauwkeurigheid van de EEM-oplossing [A,C];
(d) de toepassing van de Laplace-transformatie op beginrandwaardeproblemen, teneinde de foutenanalyse te vereenvoudigen [A,C].

In het vierde artikel wordt een driedimensionaal potentiaal- en rotatieprobleem op een cilindersymmetrisch gebied numeriek opgelost door middel van de EEM. Door toepassing van een geschikte numerieke kwadratuur wordt de bandmatrix van het op te lossen stelsel gereduceerd van een 27-diagonale tot een 11-diagonale matrix.

In het vijfde artikel wordt een diffusieprobleem met twee fasen en een bewegend front gecombineerd en door middel van de EEM. De hieruit ontstane gewone differentiaalvergelijking wordt met een aangepaste methode van Runge-Kutta opgelost.