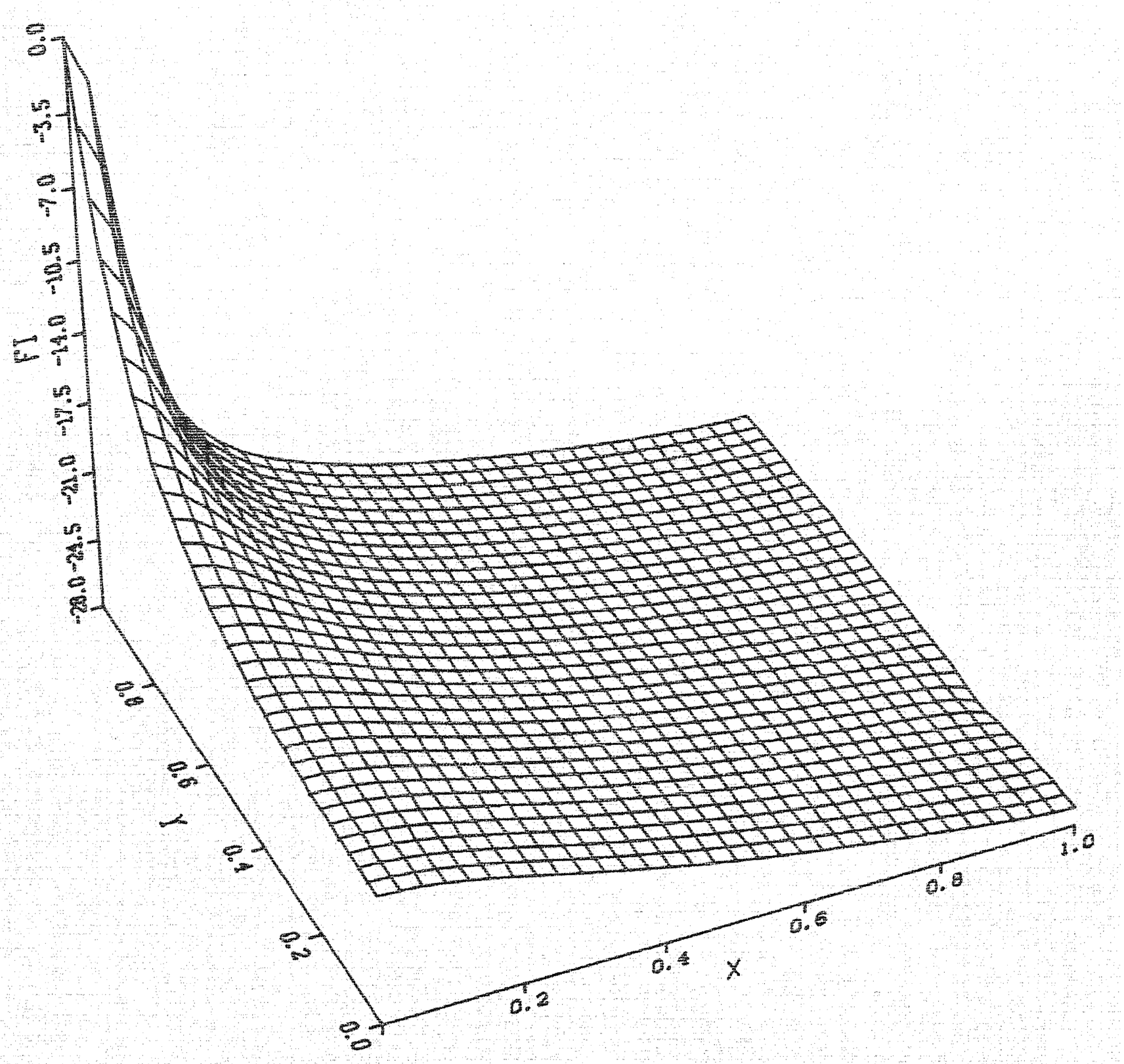
ASPECTS OF THE FINITE ELEMENT METHOD



MIENTE BAKKER

UvA

STELLINGEN

bij het proefschrift

ASPECTS OF THE FINITE ELEMENT METHOD

van

M. BAKKER

3 november 1982

I

De C^0-Collocatie-Galerkin-methode kan gezien worden als een C^0-Galerkin-methode met gebruik van Lobatto-kwadratuur.

J.C. Diaz, *A Collocation-Galerkin Method for the Two Point Boundary Value Problem Using Continuous Piecewise Polynomial Spaces*, SIAM J. Numer. Anal. 14 (1977), pp. 844-858.

II

Zij gegeven het Stefan-probleem

$$c\rho \frac{\partial T}{\partial t} = \frac{\partial}{\partial z}\left(\kappa(T)\frac{\partial T}{\partial z}\right) + \alpha I_0(1-R)e^{-\left(\frac{t-t_0}{h}\right)^2-\alpha z}, \quad t > 0;\ z \in (0,z_0)\backslash\{s(t)\};$$

$$\frac{\partial T}{\partial z}(0,t) = 0; \quad T(z_0,t) = T(z,0) = T_0 \quad (= 300^0K);$$

$$s(0) = 0; \quad T(s(t),t) = T_m, \quad \text{als } s(t) > 0;$$

$$\frac{ds(t)}{dt} = \begin{cases} 0 & , \quad \text{als } T(0,t) \le T_m \quad (= 1685^0K) \\ \frac{1}{L_0\rho}\left[\kappa(T)\frac{\partial T}{\partial z}\right]_{T=T_m^-}^{T=T_m^+} & , \quad \text{als } T(0,t) > T_m; \end{cases}$$

$$\kappa(T) > 0, \quad T \ge T_0 ;$$

c, ρ, α, I_0, t_0, h, z_0, R positieve parameters.

Voor de oplossing van bovenstaand probleem is de methode van Bonerot en Jamet [2] minder geschikt. Een goede methode om dit probleem op te lossen is de substitutie

$$z = \begin{cases} xs(t), & x \in (0,1), \quad \text{als } z \in (0,s(t)); \\ (1-x)s(t) + xz_0, & x \in (0,1), \quad \text{als } z \in (s(t),z_0); \end{cases}$$

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

[1] M. Bakker, F.W. Saris & Z.L. Wang, *Laser-Annealing as a Moving Boundary Problem*, verschijnt begin 1983.

[2] R. Bonerot & P. Jamet, *A Second Order Finite Element Method for the One-Dimensional Stefan Problem*, Internal. J. Numer. Methods Engrg. 8 (1974), pp. 811-820.

III

De bewering van M.F. Wheeler c.s. dat C^0-Galerkin-methoden voor cirkel- en bolsymmetrische gebieden superconvergent zijn in de roosterpunten is onjuist wat het middelpunt van de bol of de cirkel betreft.

G.F. Carey, D. Humphrey & M.F. Wheeler, *Galerkin and Collocation-Galerkin Methods with Superconvergence and Optimal Fluxes*, Internat. J. Numer. Methods Engrg. 17 (1981), pp. 939-950.

IV

Het spectrum van de Boltzman 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.

P. Résibois & M. de Leener, *Classical Kinetic Theory of Fluids*, Wiley and Sons, New York, 1977.

V

In [2] verklaart Wood dat bij laser-annealing de grootte van de absorptiecoefficient vanaf $2.0_{10}4$ cm^{-1} niet meer van invloed is op de smeltdiepte en dat dit wordt bevestigd door de resultaten van het computerprogramma HEATING5. Deze bewering is in tegenspraak met de resultaten van het computerprogramma WANG [1].

[1] M. Bakker, F.W. Saris & Z.L. Wang, *Laser-Annealing as a Moving Boundary Problem*, verschijnt begin 1983.

[2] R.F. Wood & G.E. Giles, *Macroscopic Theory of Pulsed Laser Annealing*, Phys. Rev. B (3) (1981), pp. 2923-2942.

VI

$$\frac{1}{\pi}\int_0^{\pi} J_{2\nu}(2z\sin\phi)e^{z\cos\phi}d\phi = I_\nu^2(z), \quad z \in \mathbb{C}.$$

VII

Laat Turkin's functie $T_m(z,\alpha)$ gedefinieerd zijn door

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

waarin $J_\nu(z)$ de bekende Besselfunktie voorstelt. Voor $T_m(z,\alpha)$ geldt de inhomogene recursie

$$(2) \qquad T_{m-1}(z,\alpha) + T_{m+1}(z,\alpha) = \frac{2(\alpha-m)}{z} T_m(z,\alpha) + \frac{2}{z}\delta_{0m};\ m \in \mathbb{Z}$$

waarby δ_{jm} het symbool van Kronecker is. Op grond van (1) en (2) kan de bewering van Newberger dat

$$(3) \qquad T_m(z,\alpha) = (-1)^{m+1} \frac{\pi}{\sin\pi\alpha} J_{m-\alpha}(z)J_\alpha(z)$$

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

B.S. Newberger, New sum rule for products of Bessel functions with application to Plasma Physics, J. Math. Phys. 23 (1982), 1278-1281.

VIII

Als de voorstellen van de tweede Commissie-Wagner om de ontslagprocedures te vereenvoudigen, worden uitgevoerd, zal het verschil in rechtspositie tussen overheidsdienaren en andere werknemeners 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 onderwijzend 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 RECTOR MAGNIFICUS
DR. R.W. BRESTERS,
HOOGLERAAR IN DE FACULTEIT DER WISKUNDE EN NATUURWETENSCHAPPEN,
IN HET OPENBAAR TE VERDEDIGEN
IN DE AULA DER UNIVERSITEIT
(TIJDELIJK IN DE LUTHERSE KERK, INGANG SINGEL 411, HOEK SPUI)
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

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. Hoonhout, 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 Zienkiewitz.

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.

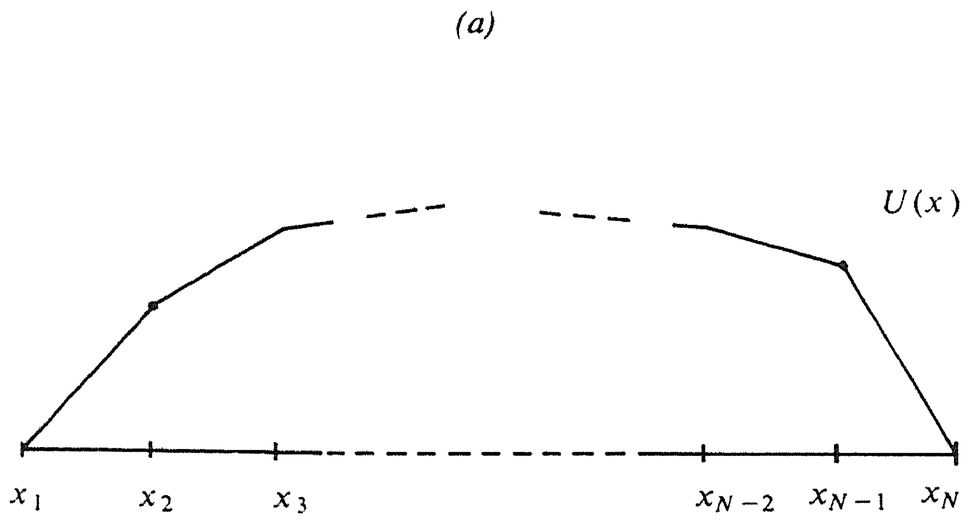


Figure 1.

(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

$$\frac{d^2u}{dx^2} = f(x),\ x\epsilon(0,1);\ u(0) = u(1) = 0 \tag{1}$$

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

$$E(v) = \int_0^1 [(v')^2 - 2fv]dx\ : \tag{2a}$$

over the space

$$V = \{v \mid v\epsilon H^1(0,1);\ v(0) = v(1) = 0\}. \tag{2b}$$

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 $E(\sum_{i=1}^{N} q_i\phi_i(x))$ for all possible real numbers $q_1, \cdots, q_N$. 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 $(\partial^2 E / \partial\phi_i \partial\phi_j)$ 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 ϕ_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 ϕ_i can be piecewise quadratics or piecewise cubics. In the case of two-dimensional problems, the ϕ_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 $\mathbf{R}^1$, these elements are plain intervals, in $\mathbf{R}^2$, they are triangles, quadrangles, etc., in $\mathbf{R}^3$, 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 ϕ_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 ϕ_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 non-linear) system of the form

$$F_i(c_1, \cdots, c_N) = 0, i = 0, \cdots, N; \tag{3}$$

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]

$$||u - U||_{L^2(0,1)} \leqslant C(u)h^2; \tag{4}$$

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 $W_0^{m,p}(\Omega)$, i.e. a Banach space of functions that, together with their first m distributional partial derivatives, are L^p integrable over the domain Ω 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

$$U(x,t) = \sum_{i=1}^{N} c_i(t)\phi_i(x). \tag{5}$$

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

$$G\frac{d\mathbf{c}}{dt} + A\mathbf{c} = \mathbf{b},\ t \geqslant 0;\ \mathbf{c}(0) = \mathbf{c}_0; \tag{6}$$

where $\mathbf{c}(0)$ 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

$$||u(t) - U(t)||_{L^2(\Omega)} \leqslant C(t,u)h^{k+1},\ t \geqslant 0;$$

where k is the degree of the Finite Element space. A frequent choice of $c_i(0)$ is the interpolate 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 *stiffnes 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.

[A] *On the Numerical Solution of Parabolic Equations in a Single Space Variable by Means of the Continuous Time Galerkin Method,* SIAM J. Numer. Anal. **17** (1980), 162-177; ZB 438.65094; MR 81m65167.

[B] *A Note on C^0 Galerkin Methods for Two-Point Boundary Problems,* Numer. Math. **38** (1982), 447-453; ZB 419.65049.

[C] *Galerkin Methods for Even-Order Parabolic Equations in One Space Variable,* SIAM J. Numer. Anal. **19** (1982), 571-587; ZB 422.65059.

[D] *A Program to Solve Rotating Plasma Problems,* Comp. Phys. Comm. **20** (1980), 429-439; ZB 437.76099.

[E] *A Program to Solve a Solute Diffusion Problem with Segregation at a Moving Interface,* Comp. Phys. Comm. **22** (1981), 439-450; ZB 439.76070.

OTHER REFERENCES.

[1] M. BAKKER, *Numerical Solution of Mildly Nonlinear Two-Point Boundary Value Problems by Means of Galerkin's Method,* MC report NW 27/76, Mathematisch Centrum, Amsterdam, 1976.

[2] M.S. VAN DEN BERG, *Theory on a Partially Ionized Gas Centrifuge,* Thesis, FOM, Amsterdam, 1982.

[3] J.H. BRAMBLE & S.R. HILBERT, *Estimation of Linear Functionals on Sobolev Spaces with Application to Fourier Transforms and Spline Interpolation,* SIAM J. Numer. Anal. 7 (1970), 112-124.

[4] J. CHRISTIANSEN & R.D. RUSSELL, *Error Analysis for Spline Collocation Methods with Application to Knot Selection,* Math. Comp. **32** (1978), 415-419.

[5] P.G. CIARLET, *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, 1978.

[6] R. COURANT, *Variational Methods for the Solution of Problems of Equilibrium and Vibrations,* Bull. Amer. Math. Soc. **49** (1943), 1-23.

[7] R. COURANT, K. FRIEDRICHS & H. LEWY, *Ueber die Partiellen Differenzengleichungen der Mathematischen Physik,* Math. Ann. **100** (1928), 32-74.

[8] C. DE BOOR & B. SWARTZ, *Collocation at Gaussian Points,* SIAM J. Numer. Anal. **10** (1973), 582-606.

[9] J. DOUGLAS, JR., & T. DUPONT, *Galerkin Methods for Parabolic Equations,* SIAM J. Numer. Anal. 7 (1970), 575-626.

[10] J. DOUGLAS, JR., & T. DUPONT, *Some Superconvergence Results for Galerkin Methods for the Approximate Solution of Two-Point Boundary Problems,* from: J.J.H. Miller (ed.), Topics in Numerical Analysis, Academic Press, London, 1973.

[11] J. DOUGLAS, JR., & T. DUPONT, *Galerkin Approximation for the Two-Point Boundary Problem Using Continuous Piecewise Polynomial Spaces,* Numer. Math. **22** (1974), 99-109.

[12] J. DOUGLAS, JR., T. DUPONT & M.F. WHEELER, *Some Superconvergence Results for an H^1 Galerkin Procedure for the Heat Equation,* Report MRC 1382, Madison, WI, 1973.

[13] C.A. FELIPPA & R.W. CLOUGH, *The Finite Element Method in Solid Mechanics,* from: G. Birkhoff (ed.), Numerical Solution of Field Problems in Continuum Physics II, Duke University, SIAM-AMS, 1970

[14] G. FORSYTHE & W. WASOW, *Finite Difference Methods for Partial Differential Equations,* Wiley, New York, 1960.

[15] B.G. GALERKIN, *Rods and Plates. Series on Some Problems of Elastic Equilibrium of Rods and Plates,* Vestn. Inzhn. Tech. **19** (1915), 897-908 (in Russian).

[16] A. GEORGE & D. McINTYRE, *On the Application of the Minimum Degree Algorithm to Finite Element Systems,* SIAM J. Numer. Anal. **15** (1978), 90-112.

[17] P.W. HEMKER, *A Numerical Study of Stiff Two-Point Boundary Problems,* MC Tract 80, Mathematisch Centrum, Amsterdam, 1977.

[18] P.W. HEMKER, *On the Comparison of Line-Gauss-Seidel and ILU Relaxation in Multigrid Algorithms,* to appear in: J.J.H. Miller (ed.), Boundary and Interior Layers II, Computational and Asymptotic Methods, Boole Press, Dublin, 1982.

[19] D. HOONHOUT, *Pulsed-Laser Annealing of Ion-Implanted Silicon,* Thesis, FOM, Amsterdam, 1980.

[20] M. MACHURA & R. SWEET, *A Survey of Software for Partial Differential Equations,* ACM Trans. Math. Software **6** (1980), 461-488.

[21] J.W. MEYERINK & H.A. VAN DER VORST, *An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric M-Matrix,* Math. Comp. **31** (1977), 148-162.

[22] A.R. MITCHELL & R. WAIT, *The Finite Element Method in Partial Differential Equations,* Wiley, New York, 1977.

[23] J.T. ODEN, *Finite Elements of Nonlinear Continua,* McGraw-Hill, New York, 1972.

[24] J.T. ODEN & J.N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements,* Wiley, New York, 1976.

[25] F.J. PETERS, *Sparse Matrices and Substructures, With a Novel Implementation of Finite Element Algorithms,* MC tract 119, Mathematisch Centrum, Amsterdam, 1980.

[26] P.A. RAVIART, *The Use of Numerical Integration in Finite Element Methods for Solving Parabolic Equations,* from: J.J.H. Miller (ed.), Topics in Numerical Analysis, Academic Press, London, 1973.

[27] J.K. REID (ed.), *Large Sparse Sets of Linear Equations,* Academic Press, London, 1971.

[28] G.STRANG & G.J. FIX, *An Analysis of the Finite Element Method,* Prentice-Hall, Englewood Cliffs, N.J., 1973.

[29] V. THOMEE, *Some Convergence Results for Galerkin Methods for Parabolic Boundary Value Problems,* from: C. DeBoor (ed.), Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, London, 1974.

[30] M. VAN VELDHUIZEN, *A Refinement Process for Collocation Approximations,* Numer. Math. **26** (1976), 397-407.

[31] J.H. WILKINSON & C. REINSCH, *Handbook for Automatic Computation,* Volume 2, Linear Algebra, Springer-Verlag, 1971.

[32] D.M. YOUNG, *Iterative Solution of Large Linear Systems,* Academic Press, London, 1971.

THE FIVE PAPERS

SIAM J. NUMER. ANAL.
Vol. 17, No. 1, February 1980

0036-1429/80/1701-0016 $01.00/0

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

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

$$x \in [0, 1] = I; \qquad t \in [0, \infty) = J; \tag{1.1}$$

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

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 Faedo-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 $H^m(E)$, $m \geqq 0$, by

$$H^m(E) = \{v | D^j v \in L^2(E), j = 0, \cdots, m\}, \tag{2.1}$$

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

$$(u, v)_{H^m(E)} = \sum_{j=0}^{m} (D^j u, D^j v)_{L^2(E)};$$

$$\|u\|_{H^m(E)} = [(u, u)_{H^m(E)}]^{1/2}, \tag{2.2a}$$

where

$$(\alpha, \beta)_{L^2(E)} = \int_E \alpha(x)\overline{\beta(x)}\, dx, \qquad \alpha, \beta \in L^2(E). \tag{2.2b}$$

* Received by the editors August 22, 1978, and in revised form April 16, 1979.
† Mathematisch Centrum, Amsterdam, The Netherlands.

In the sequel, we use (u, v) instead of $(u, v)_{L^2(I)}$ and $\|u\|_m$ instead of $\|u\|_{H^m(I)}$.

The subspace $H_0^1(I)$ of $H^1(I)$ is defined by

$$H_0^1(I) = \{v | v \in H^1(I);\ v(0) = v(1) = 0\}. \tag{2.3}$$

Next we define the bilinear functional $B: H_0^1(I) \times H_0^1(I) \to \mathbb{C}$ associated with the operator L, L defined by (1.1), as follows:

$$B(u, v) = (pDu, Dv) + (qu, v); \qquad u, v \in H_0^1(I). \tag{2.4}$$

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

$$B(u, u) \geqq C\|u\|_1^2, \qquad u \in H_0^1(I). \tag{2.5}$$

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

2.2. Galerkin's method. Let $u: J \to H_0^1(I) \cap H^2(I)$ 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 $H_0^1(I)$. This U is given by the equation

$$\left(\frac{\partial}{\partial t}U(t,), V\right) + B(U(t,), V) = 0, \qquad V \in S,\ t \geqq 0, \quad U(0,) = U_0 \in S, \tag{2.6}$$

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. $x_j = jh = jN^{-1}$, $j = 0, \cdots, N$. By I_j we denote the segment $[x_{j-1}, x_j]$. We define the space $M_0^r(\Delta)$ (r a constant positive integer) by

$$M_0^r(\Delta) = \{V | V \in H_0^1(I);\ V \in P_r(I_j), j = 1, \cdots, N\}, \tag{2.7}$$

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

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

$$\|v\|_{l,\Delta} = \left[\sum_{j=1}^{N} \|v\|_{H^l(I_j)}^2\right]^{1/2}; \tag{2.8a}$$

$$H^m(\Delta) = \{v | v \in H^m(I_j), j = 1, \cdots, N\}. \tag{2.8b}$$

LEMMA 1. *Let $U_0 \in M_0^r(\Delta)$ be an approximation to u_0 satisfying*

$$\|u_0 - U_0\|_0 = O(h^{r+1}\|u_0\|_{r+1}) \tag{2.9}$$

and let $U: J \to M_0^r(\Delta)$ 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^2 bound

$$\|e(t)\|_0 \leqq e^{-\lambda_1 t}\|e(0)\|_0 + Ch^{r+1}[\|u(t)\|_{r+1} + e^{-\lambda_1 t}\|y_0\|_{r+1} + \int_0^t e^{\lambda_1(\tau - t)}\|u_\tau(\tau)\|_{r+1}\, d\tau], \tag{2.10}$$

where λ_1 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 $O(h^{2r})$ vs. $O(h^{r+1})$. 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 $v: J \to V$, V a set of functions defined on I, we define the Laplace transform $\mathscr{L}v = \hat{v}(s) \in V$ by

$$\hat{v}(s, x) = \int_0^\infty e^{-st} v(t, x)\, dt, \qquad s \in \mathbb{C}, \quad x \in I. \tag{3.1}$$

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

$$\begin{aligned} &L\hat{u}(s) + s\hat{u}(s) = u_0, \qquad x \in I; \\ &\hat{u}(s, 0) = \hat{u}(s, 1) = 0. \end{aligned} \tag{3.2a}$$

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 $(u, v)_{H^m(I)}$. We write (3.2a) in its Galerkin form to obtain

$$B(\hat{u}(s), v) + s(\hat{u}(s), v) = (u_0, v), \qquad v \in H_0^1(I). \tag{3.2b}$$

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

$$B(\hat{U}(s), \hat{V}) + s(U(s), V) = (u_0, V), \qquad V \in M_0^r(\Delta). \tag{3.3}$$

At the other hand, the solution $U(t) \in M_0^r(\Delta)$ of (2.6) has the Laplace transform $\hat{U}_1(s)$ which is the solution of the boundary value problem

$$B(\hat{U}_1(s), V) + s(\hat{U}_1(s), V) = (U_0, V), \qquad V \in M_0^r(\Delta). \tag{3.3a}$$

One easily verifies that if $U_0 \in M_0^r(\Delta)$ is defined by

$$(u_0 - U_0, V) = 0, \qquad V \in M_0^r(\Delta), \tag{3.4}$$

then $\hat{U}_1(s) \equiv \hat{U}(s)$, which is illustrated in diagram 1 (see next page). Also, one easily verifies that U_0 satisfies (2.9), since it minimizes $\|u_0 - V\|_0$ over $M_0^r(\Delta)$.

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

$$\begin{aligned} &u(t, x) = \sum_{n=1}^{\infty} a_n e^{-\lambda_n t} \phi_n(x); \\ &\hat{u}(s, x) = \sum_{n=1}^{\infty} a_n (s + \lambda_n)^{-1} \phi_n(x); \\ &a_n = (u_0, \phi_n), \qquad n = 1, 2, \cdots; \\ &U(t, x) = \sum_{n=1}^{M} A_n e^{-\Lambda_n t} \Phi_n(x); \\ &\hat{U}(s, x) = \sum_{n=1}^{M} A_n (s + \Lambda_n)^{-1} \Phi_n(x); \\ &A_n = (U_0, \Phi_n), \qquad n = 1, 2, \cdots, M = rN - 1, \end{aligned} \tag{3.5}$$

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

$$B(\Phi_n, V) = \Lambda_n(\Phi_n, V), \qquad V \in M_0^r(\Delta), \qquad n = 1, \cdots, M, \tag{3.6}$$

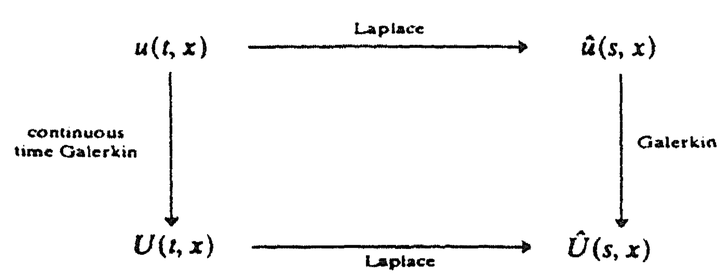


DIAGRAM 1.

with orthonormal eigenfunctions $\Phi_1, \Phi_2, \cdots$. Note that $\Lambda_1 > \lambda_1$ since

$$\Lambda_1 = \inf_{V \in M_0^1(\Delta)} \frac{B(V, V)}{(V, V)} > \inf_{v \in H_0^1(I)} \frac{B(v, v)}{(v, v)} = \lambda_1. \tag{3.7}$$

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

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

$$v(t, x) = \mathscr{L}^{-1}\hat{v}(s, x) = \frac{1}{2\pi i} \int_{\Gamma} e^{st} \hat{v}(s, x)\, ds, \tag{3.8}$$

where $\Gamma = \{s | s = \sigma + i\tau, -\infty < \tau < +\infty,\}$ is a contour in the convergence half-plane of $\hat{v}(s)$. For $\hat{v}(s) = \hat{e}(s)$, the convergence half-plane is given by Re $s > -\lambda_1$. One intuitively feels now that one can use local convergence results for $\hat{e}(s, x)$ to establish similar results for $e(t, x)$ 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 $|s| \to \infty$.

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 Γ we take the imaginary axis ($\sigma = 0$). We now define the contours $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$ and Γ_5 by (see Fig. 1)

$$\begin{aligned}
\Gamma_1 &= \{s | s = +i\tau;\ -R \leqq \tau \leqq R\};\\
\Gamma_2 &= \{s | s = \alpha + iR;\ -R \leqq \alpha \leqq 0\};\\
\Gamma_3 &= \{s | s = \alpha(-1+i);\ 0 \leqq \alpha \leqq R\};\\
\Gamma_4 &= \{s | s = -\alpha(1+i);\ 0 \leqq \alpha \leqq R\};\\
\Gamma_5 &= \{s | s = \alpha - iR;\ -R \leqq \alpha \leqq 0\};
\end{aligned} \tag{3.9}$$

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

$$\int_{\overline{P_1P_2P_3P_1}} \hat{e}(s, x)\, e^{st}\, ds = 0, \qquad x \in I, \quad t \in J, \tag{3.10}$$

where by $\overline{P_1P_2 \cdots P_n}$ we mean the polygonal line starting in P_1 connecting P_1 with P_2,

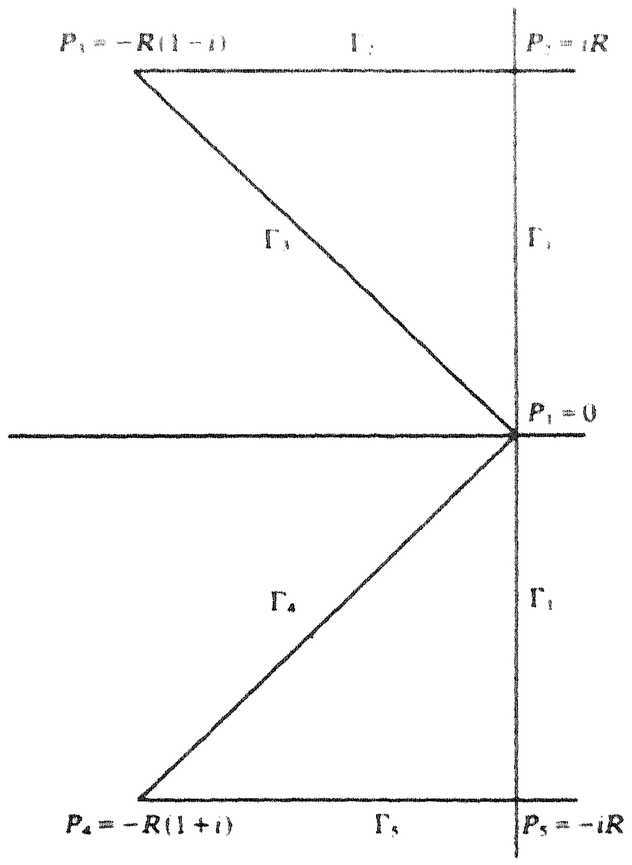


FIG. 1

etc. and ending at P_n. We now consider $\overline{P_2P_3} = \Gamma_2$. If we apply (3.2b) for $v = \hat{u}(s)$, $s = -\alpha + iR$, $0 \leqq \alpha \leqq R$, we get

$$B(\hat{u}(s), \hat{u}(s)) + (iR - \alpha)(\hat{u}(s), \hat{u}(s)) = (\hat{u}(s), u_0). \tag{3.11}$$

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

$$\begin{aligned} &|B(\hat{u}(s), \hat{u}(s)) - \alpha(\hat{u}(s)), \hat{u}(s))|^2 + R^2|(\hat{u}(s), \hat{u}(s))|^2 \\ &\quad = |(\hat{u}(s), u_0)|^2, \quad s = -\alpha + iR. \end{aligned} \tag{3.12}$$

It is easily seen from (3.12) that

$$R|(\hat{u}(s), \hat{u}(s))| \leqq |(\hat{u}(s), u_0)|; \qquad \frac{R}{\sqrt{\alpha^2 + R^2}}|B(\hat{u}(s), \hat{u}(s))| \leqq |(\hat{u}(s), u_0)|; \tag{3.13}$$

hence

$$\|\hat{u}(s)\|_0 \leqq R^{-1}\|u_0\|_0$$

and

$$\begin{aligned} &\|\hat{u}(s)\|_1^2 \leqq CB(\hat{u}(s), \hat{u}(s)) \leqq C\|\hat{u}(s)\|_0\|u_0\|_0 \leqq CR^{-1}\|u_0\|_0^2; \\ &\|\hat{u}(s)\|_1 \leqq CR^{-1/2}\|u_0\|_0; \qquad |\hat{u}(s, x)| \leqq \|\hat{u}(s)\|_1 \leqq CR^{-1/2}\|u_0\|_0. \end{aligned} \tag{3.14}$$

The next-to-last inequality is obtained by applying Poincaré's inequality. In a similar way one proves that

$$\|\hat{U}(s, x)\| \leqq CR^{-1/2}\|U_0\|_0, \tag{3.15}$$

hence

$$|\hat{e}(s, x)| \leqq CR^{-1/2}(\|u_0\|_0 + \|U_0\|_0), \quad \text{on } \Gamma_2. \tag{3.16}$$

After this, we obtain

$$\left|\int_{\overline{P_2P_3}} \hat{e}(s, x)\, e^{st}\, ds\right| \leqq \int_{\overline{P_2P_3}} |\hat{e}(s, x)|\, |e^{st}|\, |ds|$$

$$\leqq CR^{-1/2}(\|u_0\|_0 + \|U_0\|_0) \int_0^R e^{-\alpha t}\, d\alpha$$

$$\leqq CR^{-1/2}t^{-1}(\|u_0\|_0 + \|U_0\|_0),$$

which implies that

$$\lim_{R\to\infty} \int_{\overline{P_2P_3}} \hat{e}(s, x)\, e^{st}\, ds = 0, \qquad t > 0. \tag{3.17}$$

In a similar way one proves that

$$\lim_{R\to\infty} \int_{\overline{P_4P_5}} \hat{e}(s, x)\, e^{st}\, ds = 0, \qquad t > 0,$$

so we have proved that

$$\begin{aligned} e(t, x) &= \frac{1}{2\pi i} \lim_{R\to\infty} \int_{\overline{P_4P_1P_2}} \hat{e}(s, x)\, e^{st}\, ds \\ &= \frac{1}{2\pi} \int_0^\infty [\hat{e}(-\alpha - \alpha i, x)\, e^{-\alpha it} + \hat{e}(-\alpha + \alpha i, x)\, e^{+\alpha it}]\, e^{-\alpha t}\, d\alpha, \end{aligned} \tag{3.18}$$

which is an absolutely convergent integral, since $\hat{e}(s, x)$ is bounded in α. We now have to obtain local error bounds for $\hat{e}(s, x)$ on the lines $\{s | s = -\alpha(1 \pm i), \alpha \geqq 0\}$ only, if we want to derive local error bounds for $e(t, x)$.

We return to problems (3.2) and (3.4), but now only for $s = -\alpha(1 \pm i)$. 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*

$$\|\hat{e}(s)\|_l \leqq C(1 + h\alpha^{l/2})\|\hat{u}(s)\|_{r+1} h^{r+1-l}, \qquad l = 0, 1. \tag{3.19}$$

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

$$B(\hat{e}(s), V) + s(\hat{e}(s), V) = 0, \qquad V \in M_0^r(\Delta). \tag{3.20}$$

Next, we introduce the elliptic projection $\hat{U}_2(s)$ of $\hat{u}(s)$ defined by

$$B(\hat{u}(s) - \hat{U}_2(s), V) = 0, \qquad V \in M_0^r(\Delta). \tag{3.21}$$

We know (see e.g. [16]) that

$$\|\hat{u}(s) - \hat{U}_2(s)\|_l \leqq Ch^{r+1-l}\|\hat{u}(s)\|_{r+1}, \qquad l = 0, 1. \tag{3.22}$$

If we set $\hat{\delta}(s) = \hat{U}(s) - \hat{U}_2(s)$ and subtract (3.21) and (3.20), we obtain after putting

$V = \hat{\delta}(s)$

$$B(\hat{\delta}, \hat{\delta}) + s(\hat{\delta}, \hat{\delta}) = s(\hat{u} - \hat{U}_2, \hat{\delta}). \tag{3.23}$$

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

$$|B(\hat{\delta}, \hat{\delta}) - \alpha(\hat{\delta}, \hat{\delta})|^2 + \alpha^2|(\hat{\delta}, \hat{\delta})|^2 = 2\alpha^2|(\hat{u} - \hat{U}_2, \hat{\delta})|^2. \tag{3.24}$$

From (3.24), we easily get

$$\begin{aligned} &\alpha^2\|\hat{\delta}\|_0^4 \leqq 2\alpha^2\|\hat{\delta}\|_0^2\|\hat{u} - \hat{U}_2\|_0^2; \qquad \|\hat{\delta}\|_0 \leqq \sqrt{2}\|\hat{u} - \hat{U}_2\|_0 \leqq C\|\hat{u}\|_{r+1}h^{r+1}; \\ &\|\hat{e}\|_0 \leqq \|\hat{\delta}\|_0 + \|\hat{u} - \hat{U}_2\|_0 \leqq Ch^{r+1}\|\hat{u}\|_{r+1}. \end{aligned} \tag{3.25}$$

In a similar way we prove

$$\tfrac{1}{2}|B(\hat{\delta}, \hat{\delta})|^2 \leqq 2\alpha^2\|\hat{\delta}\|_0^2\|\hat{u} - \hat{U}_2\|_0^2 \leqq \alpha^2 Ch^{4r+4}\|\hat{u}\|_{r+1}^2,$$

hence

$$\begin{aligned} &\|\hat{\delta}(s)\|_1 \leqq C\sqrt{\alpha}\, h^{r+1}\|\hat{u}(s)\|_{r+1}; \\ &\|\hat{e}(s)\|_1 \leqq \|\hat{\delta}(s)\|_1 + \|\hat{u}(s) - \hat{U}_2(s)\|_1 \leqq C(1 + \sqrt{\alpha}\, h)h^r\|\hat{u}_s\|_{r+1}, \end{aligned} \tag{3.26}$$

which proves the lemma. □

We now introduce the Green's function $\hat{G}(s; x, \cdot) \in H_0^1(I) \cap H^{r+1}(0, x) \cap H^{r+1}(x, 1)$ associated to problem (3.2) defined by

$$B(v, \hat{G}(s; x, \cdot)) + s(v, \hat{G}(s; x, \cdot)) = v(x), \qquad v \in H_0^1(I), \quad x \in I. \tag{3.27}$$

For $j = 1, \cdots, N-1$, we define $\hat{G}_j(s) \in H_0^1(I) \cap H^{r+1}(\Delta)$ by

$$\hat{G}_j(s; \xi) = \hat{G}(s; x_j, \xi), \qquad \xi \in (0, x_j) \cup (x_j, 1). \tag{3.28}$$

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

$$\begin{aligned} \hat{e}(s, x_j) &= B(\hat{e}(s), \hat{G}_j(s)) + s(\hat{e}(s), \hat{G}_j(s)) \\ &= B(\hat{e}(s), \hat{G}_j(s) - V) + s(\hat{e}(s), \hat{G}_j(s) - V), \qquad V \in M_0^r(\Delta), \end{aligned} \tag{3.29}$$

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

$$\begin{aligned} &\|\hat{G}_j(s) - V\|_l \leqq Ch^{r+1-l}\|\hat{G}_j(s)\|_{r+1,\Delta}, \qquad l = 0, 1, \cdots, r; \\ &\|V\|_{r,\Delta} \leqq C\|\hat{G}_j\|_{r+1,\Delta}. \end{aligned} \tag{3.30}$$

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

$$\begin{aligned} |\hat{e}(s, x_j)| &\leqq |B(\hat{e}(s), \hat{G}_j(s) - V)| + |s|\,|(\hat{e}(s), \hat{G}_j(s) - V)| \\ &\leqq C\|\hat{e}(s)\|_1\|\hat{G}_j(s) - V\|_1 + 2\alpha\|\hat{e}(s)\|_0\|\hat{G}_j(s) - V\|_0 \\ &\leqq Ch^{2r}(1 + h\sqrt{\alpha} + h^2\alpha)\|\hat{u}(s)\|_{r+1}\|\hat{G}_j(s)\|_{r+1,\Delta}. \end{aligned} \tag{3.31}$$

Superconvergence seems to be within grasp, but we need yet explicit dependence on u_p.

From (3.2a), we find that

$$\begin{aligned} &B(\hat{u}, \hat{u}) + s(\hat{u}, \hat{u}) = (u_0, \hat{u}); \\ &(L\hat{u}, L\hat{u}) + sB(\hat{u}, \hat{u}) = B(u_0, \hat{u}); \\ &B(L\hat{u}, L\hat{u}) + s(L\hat{u}, L\hat{u}) = (Lu_0, L\hat{u}), \end{aligned}$$

from which we easily obtain that

$$\|\hat{u}\|_m \leqq C\alpha^{-1}\|u_0\|_m, \qquad m = 0, 1, 2,$$

if $|s|$ is large enough. After this, we obtain by induction

$$\|\hat{u}\|_m \leqq C\alpha^{R(m)}\|u_0\|_m, \qquad m = 0, 1, \cdots, r+1; \tag{3.32}$$
$$R(m) = [(m-3)/3].$$

Concerning $\hat{G}_j(x, s)$, a direct computation shows that

$$\|\hat{G}_j\|_{\Delta,m} \leqq C\alpha^{(m-1)/2}, \qquad m \geqq 0, \tag{3.33}$$

if α is large enough.

Combining (3.32) and (3.33), we obtain

$$|e(t, x_j)| \leqq C\|u_0\|_{r+1}h^{2r}\int_0^\infty e^{-\alpha t}F_j(\alpha), \qquad j = 1, \cdots, N-1; \quad t > 0, \tag{3.34}$$

where $F_j(\alpha)$ is of at most polynomial growth in α.

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 $u: J \to H^{r+1}(I) \cap H_0^1(I)$ be the solution of (1.1), with $u_0 \in H^{r+1}(I) \cap H_0^1(I)$ and let $U: J \to M_0^r(\Delta)$ 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 $t \geqq 0$ and the local bound (3.34) for any $t > 0$.* □

Remark. One can improve (3.33) by taking the contour $\{s | s = -\alpha - \mu \pm i\alpha, \alpha \geqq 0\}$ as new integration path, where μ is a positive number between 0 and λ_1. In that case, one obtains the error bound

$$|e(t, x_j)| \leqq Ce^{-\mu t}h^{2r}\|u_0\|_{r+1}\int^\infty e^{-\alpha t}F_j(\alpha; \mu)\, d\alpha, \qquad t > 0.$$

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 $M_0^r(\Delta)$, with preservation of sparseness.

4.1. Lobatto quadrature. It is known (cf. e.g. [5]) that if $f \in H^{2r}(I)$, the integral

$$\int_0^1 f(\sigma)\, d\sigma$$

can be approximated by the $(r+1)$-point Lobatto quadrature

$$Q[f] = \sum_{l=0}^{r}{}' w_l f(\sigma_l), \tag{4.1}$$

where $\sigma_0 = 0$, $\sigma_r = 1$ and $\sigma_1, \cdots, \sigma_{r-1}$ are distinct points inside I. The weights w_l are positive. Examples are the trapezium rule ($r = 1$) and Simpson's rule ($r = 2$). The approximation (4.1) is exact when $f \in P_{2r-1}(I)$, otherwise the error is of $O(D^{2r}f(\xi)$, $\xi \in (0, 1))$, or of $O(\|D^{2r}f\|_0)$. Next, we define $(\alpha, \beta)_h$ by

$$(\alpha, \beta)_h = \sum_{j=1}^{N} (\alpha, \beta)_j^*;$$

$$(\alpha, \beta)_j^* = h \sum_{l=0}^{r} w_l \alpha(\xi_{j,l}) \beta(\xi_{j,l}), \qquad j = 1, \cdots, N; \tag{4.3}$$

$$\xi_{j,l} = x_{j-1} + h\sigma_l; \qquad j = 1, \cdots, N; \qquad l = 0, \cdots, r.$$

Note that $\xi_{j+1,0} = \xi_{j,r}$. Now $(\alpha, \beta)_h = (\alpha, \beta)$ if $\alpha\beta \in P_{2r-1}(I_j)$, $j = 1, \cdots, N$; otherwise if $\alpha\beta \in H^{2r}(I_j)$,

$$|(\alpha, \beta)_h - (\alpha, \beta)| = O(h^{2r}) \sum_{j=1}^{N} \|\alpha\beta\|_{H^{2r}(I_j)}. \tag{4.4}$$

Inequality (4.4) can be proved by direct application of the lemma of Bramble and Hilbert [1, p. 114].

LEMMA 3. *For any* $U \in M_0^r(\Delta)$,

$$|U|_h = [(U, U)_h]^{1/2} \tag{4.5}$$

is a norm equivalent to $\|U\|_0$.

Next, we define the bilinear functional $B_h : M_0^r(\Delta) \times M_0^r(\Delta) \to \mathbb{C}$ by

$$B_h(U, V) = (pU', V')_h + (qU, V)_h, \qquad U, V \in M_0^r(\Delta). \tag{4.6}$$

LEMMA 4. *For sufficiently small h, the following inequalities hold*:

$$h^l \|U\|_{l,\Delta} \leqq C h^m \|U\|_{m,\Delta}, \qquad 0 \leqq m \leqq l \leqq r; \tag{4.7}$$

$$|B_h(U, V) - B(U, V)| \leqq C h^{l+m} \|U\|_{l,\Delta} \|V\|_{m,\Delta}; \tag{4.8}$$

$$|(U, V)_h - (U, V)| \leqq C h^{l+m} \|U\|_{l,\Delta} \|V\|_{m,\Delta}; \tag{4.9}$$

$$|(f, V)_h - (f, V)| \leqq C h^{r+m} \|f\|_{2r,\Delta} \|V\|_{m,\Delta};$$

$$U, V \in M_0^r(\Delta), \quad 0 \leqq l, \quad m \leqq r. \tag{4.10}$$

Proof. Let us represent $D^m U(x)$, $x \in I_j$ by

$$D^m U(x) = \sum_{i=0}^{n-m} a_i L_i^*(x), \tag{4.11}$$

where $L_i^*(x)$ are the orthonormal Legendre polynomials shifted to I_j. Then

$$\|D^m U\|_{L^2(I_j)} = \left[\sum_{i=0}^{n-m} a_i^2 \right]^{1/2}.$$

$D^l U(x)$ is then represented by

$$D^l U(x) = \sum_{i=0}^{r-m} a_i D^{l-m} L_i^*(x) = \sum_{i=l-m}^{n-m} a_i P_{i-l+m}^*(x) h^{l-m},$$

where $P_i^*(x)$ are the Jacobi polynomials $P_i^{(l-m,l-m)}(x)$ shifted to the interval I_j. We see

that

$$\|D^l U\|_{I_j}^2 = h^{2(l-m)} \left\| \sum_{i=l-m}^{r-m} \|a_i P_{i-l-m}^*\|_0^2 \right.$$

$$= h^{2(l-m)} \sum_{i,k=l-m}^{r-m} a_i a_k (P_{i-l+m}^*, P_{k-l+m}^*)_{L^2(I_j)}$$

$$\leqq h^{2(l-m)} \sum_{i,k=l-m}^{r-m} |a_i a_k| \sum_{i,k=0}^{r-l} |(P_i^*, P_k^*)_{L^2(I_j)}|$$

$$\leqq C h^{2(l-m)} \sum_{i=l-m}^{r-m} a_i^2 \sum_{i=0}^{r-l} \|P_i^*\|_{L^2(I_j)}^2$$

$$\leqq C h^{2(l-m)} \|D^m U\|_{I_j}^2 \sum_{i=0}^{r-l} \|P_i^*\|_{L^2(I_j)}^2.$$

It remains to be proved that $\sum_{i=0}^{r-l} \|P_i^*\|_{L^2(I_j)}^2$ does not depend on h. This is, however clear since for any i, P_i^* is represented by

$$P_i^*(x) = \sum_{n=0}^{i} c_{in} L_n^*(x),$$

where C_{in} are constants not depending on h, hence

$$\|P_i^*\|_{L^2(I_j)}^2 = \sum_{n=0}^{i} C_{in}^2,$$

which proves that

$$h^l \|D^l U\|_{L^2(I_j)} \leqq C h^m \|D^m U\|_{L^2(I_j)};$$
$$h^l \|D^l U\|_{0,\Delta} \leqq C h^m \|D^m U\|_{0,\Delta}.$$

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

$$|B_h(U, V) - B(U, V)| \leqq C h^{2r} \|U\|_{r,\Delta} \|V\|_{r,\Delta} (\|p\|_{2r,\Delta} + \|q\|_{2r,\Delta});$$
$$|(U, V)_h - (U, V)| \leqq C h^{2r} \|U\|_{r,\Delta} \|V\|_{r,\Delta}; \qquad |(f, V)_h - (f, V)| \leqq C h^{2r} \|f\|_{2r,\Delta} \|V\|_{r,\Delta},$$

after which the inequalities are easily proved by applying (4.7). □

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

Proof. From (4.8), we know that

$$|B(U, U) - B_h(U, U)| \leqq C \|U\|_1^2 h^2,$$

hence

$$B_h(U, U) \geqq (C_1 - C h^2) \|U\|_1^2 \geqq C \|U\|_1^2,$$

if h is sufficiently small. □

After these technical lemmas, we arrive at:

THEOREM 2. *Let $p, q \in H^{2r}(\Delta)$, let $u \in H_0^1(I) \cap H^{r+3}(I) \cap H^{2r}(\Delta)$ and let $Y_0 \in M_0^r(\Delta)$ be an approximation to u_0, defined by*

$$(Y_0, V)_h = (u_0, V)_h, \qquad V \in M_0^r(\Delta). \tag{4.12}$$

Then the solution $Y:[0,\infty)\to M_0^r(\Delta)$ *of*

$$B_h(Y, V)+(Y_t, V)_h=0, \qquad t\geqq 0, \qquad V\in M_0^r(\Delta); \qquad Y(0)=Y_0 \tag{4.13}$$

has the following error bounds

$$\begin{aligned} &\|u-Y\|_0\leqq\|e(t)\|_0+Ch^{r+1}\|u\|_{r+3}\sqrt{t}, \qquad t\geqq 0;\\ &|(u-Y)(t,x_j)|\leqq C_j(t)h^{2r}\|u_0\|_{2r,\Delta}, \qquad j=1,\cdots,N-1, \quad t\geqq 0, \end{aligned} \tag{4.14}$$

where $\|e(t)\|_0$ *is bounded according to Lemma* 1.

Note that (4.12) implies $Y_0(\xi_{j,l})=u_0(\xi_{j,l})$, $j=1,\cdots,N$; $l=0,\cdots,r$.

Proof. Let $U, Z:J\to M_0^r(\Delta)$ be the solutions of (2.6) with initial functions U_0 and Y_0, defined by (3.4) and (4.12) respectively and let η and ε be defined by

$$\varepsilon(t,x)=U(t,x)-Y(t,x);$$

$$\eta(t,x)=Z(t,x)-Y(t,x); \qquad x\in I, t\in J.$$

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

$$\left(\frac{\partial\eta}{\partial t},\eta\right)_h+B_h(\eta,\eta)=B_h(Z,\eta)-B(Z,\eta)+(Z_t,\eta)_h-(Z_t,\eta) \tag{4.15}$$

$$\leqq Ch^{r+1}\|\eta\|_1\{\|Z_t\|_{r,\Delta}+\|Z\|_{r,\Delta}\}.$$

Let $\Pi u\in M_0^r(\Delta)$ be a projection of u defined by

$$(u-\Pi u, V)=0, \qquad V\in M_0^r(\Delta).$$

Then from Ciarlet and Raviart, we know that

$$\|u-\Pi u\|_{l,\Delta}\leqq Ch^{r+1-l}\|u\|_{r+1}, \qquad l=0,\cdots,r.$$

Hence

$$\begin{aligned} \|Z\|_{r,\Delta}&\leqq\|Z-\Pi u\|_{r,\Delta}+\|u-\Pi u\|_{r,\Delta}+\|u\|_{r,\Delta}\\ &\leqq Ch^{-r+1}\|Z-\Pi u\|_1+Ch\|u\|_{r+1}+\|u\|_r\\ &\leqq Ch\|u\|_{r+1}+\|u\|_r\leqq C\|u\|_{r+1}. \end{aligned}$$

In a similar way, we prove that

$$\|Z_t\|_{r,\Delta}\leqq C\|u_t\|_{r+1}\leqq C\|u\|_{r+3};$$

hence after applying Corollary 1, we get

$$\frac{d}{dt}|\eta|_h^2+C_1\|\eta\|_1^2\leqq C_2\|\eta\|_1\|u\|_{r+3}h^{r+1}. \tag{4.16}$$

After application of Gronwall's inequality, we get

$$\begin{aligned} &\frac{d}{dt}|\eta|_h^2\leqq Ch^{2r+2}\|u\|_{r+3}^2, \qquad t\geqq 0;\\ &|\eta(0)|_h=0, \qquad t=0. \end{aligned} \tag{4.17}$$

This differential inequality has the solution

$$|\eta|_h^2\leqq Cth^{2r+2}\|u\|_{r+3}^2 \tag{4.18}$$

which implies that

$$\|u - Y\|_0 \leqq \|e(t)\|_0 + Ch^{r+1}\sqrt{t}\|u\|_{r+3}, \tag{4.19}$$

from which one easily proves the first part of (4.14). In order to prove superconvergence at the knots, we prove that

$$|\varepsilon(t, x_j)| \leqq C(t)h^{2r}, \qquad j = 1, \cdots, N-1. \tag{4.20}$$

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

$$B_h(\hat{Y}(s), V) + s(\hat{Y}(s), V) = (Y_0, V)_h, \qquad V \in M_0^r(\Delta). \tag{4.21}$$

As in Theorem 1, one easily proves that (see Fig. 1)

$$\varepsilon(t, x) = \frac{1}{2\pi i} \lim_{R\to\infty} \int_{\overline{P_4P_1P_3}} \hat{\varepsilon}(s, x)\, e^{st}\, ds, \tag{4.22}$$

where $\hat{\varepsilon}(s, x) = \hat{Y}(s, x) - \hat{U}(s, x)$. To that end, it suffices to show that

$$|\hat{Y}(-\alpha \pm iR)| \leqq CR^{-1/2}, \qquad -R \leqq \alpha \leqq 0, \tag{4.23}$$

if h is sufficiently small. If we apply (4.21) for $s = -\alpha \pm iR$, we get

$$B_h(\hat{Y}, \hat{Y}) - \alpha(\hat{Y}, \hat{Y})_h \pm iR(\hat{Y}, \hat{Y})_h = (Y_0, \hat{Y})_h.$$

One immediately sees that

$$|\hat{Y}|_h^2 \leqq R^{-1}|(\hat{Y}_0, \hat{Y})_h|; \qquad B_h(\hat{Y}, \hat{Y}) \leqq \alpha(\hat{Y}, \hat{Y}) + |(Y_0, \hat{Y})_h| \tag{4.24}$$

and hence after application of Lemma 3,

$$\|\hat{Y}\|_0 \leqq CR^{-1}\|Y_0\|_0; \qquad B_h(\hat{Y}, \hat{Y}) \leqq C(\alpha + R)R^{-2}\|Y_0\|_0^2.$$

After application of Lemma 4 and Poincaré's inequality, we get

$$|\hat{Y}(s, x)| \leqq C\|\hat{Y}\|_1 \leqq CR^{-1/2}\|Y_0\|_0,$$

which proves (4.23).

From now on, we are only interested in the behavior of $\hat{\varepsilon}(s, x)$ on the lines $\{s | s = -\alpha(1 \pm i), \alpha \geqq 0\}$. If we subtract (4.21) from (3.3), put $V = \hat{\varepsilon}$ and apply (3.4) and (4.12), we get

$$B_h(\hat{\varepsilon}, \hat{\varepsilon}) - \alpha(1 \pm i)[(\hat{\varepsilon}, \hat{\varepsilon})_h + (\hat{U}, \hat{\varepsilon})_h - (\hat{U}, \hat{\varepsilon})] = (u_0, \hat{\varepsilon})_h - (u_0, \hat{\varepsilon}) + B(\hat{U}, \hat{\varepsilon}) - B_h(\hat{U}, \hat{\varepsilon}).$$

One easily verifies that, if h is sufficiently small,

$$|B_h(\hat{\varepsilon}, \hat{\varepsilon}) - \alpha(1 \pm i)(\hat{\varepsilon}, \hat{\varepsilon})_h| \geqq \tfrac{1}{2}\sqrt{2}\, B_h(\hat{\varepsilon}, \hat{\varepsilon}) \geqq C\|\hat{\varepsilon}\|_1^2. \tag{4.25}$$

Furthermore, application of Lemma 4 proves that

$$\begin{aligned} &|(u_0, \hat{\varepsilon})_h - (u_0, \hat{\varepsilon})| \leqq C\|u_0\|_{2r,\Delta}\|\hat{\varepsilon}\|_1 h^{r+1}; \\ &|B(\hat{U}, \hat{\varepsilon}) - B_h(\hat{U}, \hat{\varepsilon})| \leqq C\|\hat{U}\|_{r,\Delta}\|\hat{\varepsilon}\|_1 h^{r+1}; \\ &|(\hat{U}, \hat{\varepsilon}) - (\hat{U}, \hat{\varepsilon})_h| \leqq C\|\hat{U}\|_{r,\Delta}\|\hat{\varepsilon}\|_1 h^{r+1}; \end{aligned} \tag{4.26}$$

so we easily get, after application of (3.32),

$$\begin{aligned} \|\hat{\varepsilon}\|_1 &\leqq C(1+\alpha)h^{r+1}(\|\hat{U}\|_{r,\Delta} + \|u_0\|_{2r,\Delta}) \\ &\leqq C(1+\alpha)h^{r+1}(C_1\alpha^{R(r+1)} + 1)\|u_0\|_{2r,\Delta}. \end{aligned} \tag{4.27}$$

For the estimation of $\hat{\varepsilon}(s, x_j)$, we again use the Green's function:

$$
\begin{aligned}
|\hat{\varepsilon}(s, x_j)| &= |B(\hat{\varepsilon}, \hat{G}_j) + s(\hat{\varepsilon}, \hat{G}_j)| \\
&\leqq |B(\hat{\varepsilon}, \hat{G}_j - V)| + |s(\varepsilon, G_j - V)| + |B(\hat{\varepsilon}, V) + s(\hat{\varepsilon}, V)| \\
&\leqq C\|\hat{\varepsilon}\|_1 \|\hat{G}_j - V\|_1 + |s|\,\|\hat{\varepsilon}\|_0 \|\hat{G}_j - V\|_0 + |(u_0, \hat{V}) - (u_0, \hat{V})_h| \\
&\quad + |B_h(\hat{Y}, V) - B(\hat{Y}, V)| + |s(\hat{Y}, V)_h - s(\hat{Y}, V)| \\
&\leqq C\|\hat{\varepsilon}\|_1 \|\hat{G}_j - V\|_1 + \alpha\sqrt{2}\|\hat{\varepsilon}\|_1 \|\hat{G}_j - V\|_1 + C\|u_0\|_{2r,\Delta} \|V\|_{r,\Delta} h^{2r} \\
&\quad + C\|\hat{Y}\|_{r,\Delta} \|V\|_{r,\Delta} h^{2r} + \alpha\sqrt{2}\|\hat{Y}\|_{r,\Delta} \|V\|_{r,\Delta} h^{2r}, \qquad V \in M_0^r(\Delta).
\end{aligned} \tag{4.28}
$$

If we take V such that (3.30) holds, then

$$
\|\hat{G}_j - V\|_1 \leqq Ch^r \|\hat{G}_j\|_{r+1,\Delta}; \qquad \|V\|_{r,\Delta} \leqq C\|\hat{G}_j\|_{r+1,\Delta}. \tag{4.29}
$$

Furthermore

$$
\begin{aligned}
\|\hat{Y}\|_{r,\Delta} &\leqq \|\hat{\varepsilon}\|_{r,\Delta} + \|\hat{U}\|_{r,\Delta} \\
&\leqq Ch^{-r+1}\|\hat{\varepsilon}\|_1 + C\|\hat{u}\|_{r+1} \\
&\leqq C_1[h^2(1+\alpha)(C_2\alpha^{R(r+1)} + 1) + C_2\alpha^{R(r+1)}]\|u_0\|_{2r,\Delta} \\
&\leqq C_1[\alpha^{R(r+1)} + C_2 h^2 \alpha^{1+R(r+1)}]\|u_0\|_{2r,\Delta}.
\end{aligned} \tag{4.30}
$$

So we finally have

$$
|\hat{\varepsilon}(s, x_j)| \leqq F_j(\alpha) h^{2r} \|u_0\|_{2r,\Delta} \tag{4.31}
$$

where $F_j(\alpha)$ is of at most polynomial growth in α. From this, we easily derive

$$
\begin{aligned}
|(u - Y)(t, x_j)| &\leqq |e(t, x_j)| + |\varepsilon(t, x_j)| \\
&\leqq C_j(t) h^{2r} \|u_0\|_{2r,\Delta}, \qquad j = 1, \cdots, N-1; \quad t > 0.
\end{aligned}
$$

Again, $C_j(t)$ may tend to infinity as $t \to 0$. However, since

$$
|(u - Y)(0, x_j)| = 0, \qquad j = 1, \cdots, N-1,
$$

the error bound can be extended to the closed interval J. Hence, we have

$$
|(u - Y)(t, x_j)| \leqq C_j(t) h^{2r} \|u_0\|_{2r,\Delta}, \tag{4.32}
$$

where C_j is bounded on J, $C_j(0) = 0$ and where C_j is of $O(t^{-1})$, as $t \to \infty$. This completes the proof. □

Remark. By exactly the same methods, but now applied to the contour $\{s | s = -\mu \pm \alpha i\}$, this result can be improved to

$$
|(u - Y)(t, x_j)| \leqq C_j(t; \mu)\, e^{-\mu t} h^{2r} \|u_0\|_{2r,\Delta}.
$$

provided that

$$
\mu < \inf_{V \in M_0^r(\Delta)} \frac{B_h(V, V)}{|V|_h^2} = \Lambda^*,
$$

where $|\Lambda_1^* - \lambda_1|$ is known to be of $O(h^{2r})$.

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 $z_0, z_1, \cdots, z_{rN}$ by

$$\begin{aligned} &z_{jr} = x_j, \qquad j = 0, \cdots, N; \\ &z_{jr+l} = \xi_{j+1,l}, \qquad j = 0, \cdots, N-1; \quad l = 1, \cdots, r-1. \end{aligned} \tag{4.33}$$

Then one can define a basis $\{\phi_i\}_{i=1}^{M}$ for $M_0^r(\Delta)$ by

$$\phi_i(z_j) = \delta_{ij}, \qquad 1 \leqq i, j \leqq M. \tag{4.34}$$

If one represents $Y(t, x)$ by

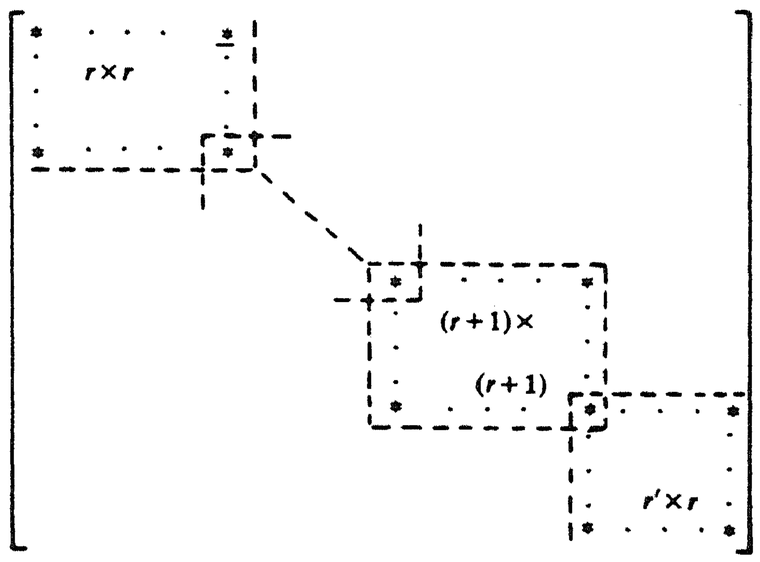
$$Y(t, x) = \sum_{i=1}^{M} a_i(t)\phi_i(x),$$

then $\mathbf{a}$ satisfies the initial value problem:

$$\begin{aligned} &K_1 \frac{d}{dt}\mathbf{a} + K_2\mathbf{a} = 0, \qquad t \geqq 0; \\ &a_i(0) = u_0(z_i), \qquad t = 0, \quad i = 1, \cdots, M. \\ &K_1 = ((\phi_i, \phi_j)_h); \\ &K_2 = (B_h(\phi_i, \phi_j)). \end{aligned} \tag{4.35}$$

As is easily checked, the $M \times M$ matrix K_2 is $(2r+1)$-diagonal of the structure

(4.36)



However, what is more important,

$$(\phi_i, \phi_j)_h = \delta_{ij}|\phi_i|_h^2, \tag{4.37}$$

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

$$K_3 \frac{d\mathbf{b}}{dt} + K_4\mathbf{b} = 0, \tag{4.38}$$

where K_3 and K_4 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

$$\frac{\partial u}{\partial t} = 2\frac{\partial^2 u}{\partial x^2} + (x^{10} + 180x^8 - x)\, e^{-t}, \qquad t \geqq 0;$$

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

The exact solution is $(x - x^{10})\, e^{-t}$. I 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.

$r = 1$

x^N	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)	7.20(−3)

$r = 2$

x^N	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)	1.77(−5)

$r = 3$

x^N	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)

REFERENCES

[1] J. H. BRAMBLE AND S. R. HILBERT, *Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation*, this Journal, 7 (1970), pp. 112–124.

[2] P. G. CIARLET AND P. A. RAVIART, *General Lagrange and Hermite interpolation in R^N with applications of finite element methods*, Arch. Rational Mech. Anal., 46 (1972), pp. 177–199.

[3] P. G. CIARLET, M. H. SCHULTZ AND R. S. VARGA, *Numerical methods of high-order accuracy for nonlinear boundary value problems I: One-dimensional problems*, Numer. Math., 9 (1967), pp. 394–430.

[4] J. H. CERRUTTI AND S. V. PARTER, *Collocation methods for parabolic partial differential equations in one space dimension*, Ibid., 26 (1976), pp. 227–254.

[5] P. J. DAVIS AND P. RABINOWITZ, *Numerical Integration*, Blaisell, New York-Toronto-London, 1967.

[6] G. DOETSCH, *Einführung in Theorie und Anwendung der Laplace–transformation*, Birkhäuser Verlag, Basel, 1958.

[7] J. DOUGLAS, JR. AND T. DUPONT, *Some superconvergence results for Galerkin methods for the approximate solution of two-point boundary problems*, Topics in Numerical Analysis, J. J. H. Miller, ed.

[8] J. DOUGLAS, JR., T. DUPONT AND M. F. WHEELER, *Some superconvergence results for an H^1-Galerkin procedure for the heat equation*, Report MRC 1382, Madison, WI, 1973.

[9] ———, *A quasi-projection approximation method applied to Galerkin procedures for parabolic and hyperbolic questions*, Report MRC 1461, Madison, WI, 1974.

[10] J. DOUGLAS, JR. AND T. DUPONT, *Galerkin approximations for the two-point boundary problem using continuous, piecewise polynomial spaces*, Numer. Math., 22 (1974), pp 99–109.
[11] ———, *Collocation Methods for Parabolic Equations in a Single Space Variable*, Springer-Verlag, Heidelberg, 1974.
[12] J. GOLDSTEIN, *Semigroups of operators and abstract Cauchy problems*, Tulane University, New Orleans, LA, 1970.
[13] P. W. HEMKER, *Galerkin methods and Lobatto points*, MC report 24/75, Mathematisch Centrum, Amsterdam, 1975.
[14] ———, *A numerical study of stiff two-point boundary problems*, MC tract 80, Mathematisch Centrum, Amsterdam, 1977.
[15] J. J. H. MILLER, ed., *Topics in Numerical Analysis*, Academic Press, London-New York, 1973.
[16] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley, New York-London-Sydney-Toronto, 1976.
[17] P. A. RAVIART, *The use of numerical integration in finite element methods for solving parabolic equations*, Topics in Numerical Analysis, J. J. H. Miller, ed.
[18] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[19] V. THOMÉE, *Some convergence results for Galerkin methods for parabolic boundary value problems*, Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974.

Numer. Math. 38, 447-453 (1982)

Numerische Mathematik

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 $O(h^{2k})$ 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 $O(h^{k+2})$, 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

$$\begin{aligned} &Lu \equiv -(p(x)u')' + q(x)u = f(x), \qquad x \in [0,1] = I; \\ &u(0) = u(1) = 0. \end{aligned} \tag{1}$$

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

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

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

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

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

0029-599X/82/0038/0447/$01.40

$$
\begin{aligned}
W^{m,2}(E) &= \{v \mid D^j v \in L^2(E),\ j=0,\ldots,m\};\\
H^m(E) &= W^{m,2}(E);\\
H_0^1(I) &= \{v \mid v \in H^1(I);\ v(0)=v(1)=0\};\\
M_0^k(\Delta) &= \{v \mid v \in H_0^1(I);\ v \in P_k(I_j),\ j=1,\ldots,N\};
\end{aligned}
\tag{2}
$$

$$
\|v\|_{W^{m,2}(E)} = \left[\sum_{j=0}^{m} \|D^j v\|^2_{L^2(E)}\right]^{\frac{1}{2}};
$$

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

where D^j denotes d^j/dx^j. If $E=I$, we write (α,β) instead of $(\alpha,\beta)_{L^2(I)}$ and $\|\alpha\|_m$ instead of $\|\alpha\|_{H^m(I)}$.

Let $U \in M_0^k(\Delta)$ be the unique solution of

$$
B(U,V) = (f,V),\quad V \in M_0^k(\Delta), \tag{3}
$$

where B: $H_0^1(I) \times H_0^1(I) \to \mathbb{R}$ is defined by

$$
B(u,v) = (pu',v') + (qu,v);\quad u,v \in H_0^1(I). \tag{4}
$$

We assume that B is strongly coercive, i.e. there exists a $C>0$ such that

$$
B(v,v) \geqq C\|v\|_1^2, \qquad v \in H_0^1(I). \tag{5}
$$

In the sequel, C, C_1, are generic positive constants not necessarily the same.

Lemma 1. *Let $u \in H_0^1(I) \cap H^{k+1}(I)$ be the solution of* (1) *and let $U \in M_0^k(\Delta)$ be the solution of* (3). *Then the error function $e(x)=u(x)-U(x)$ has the bounds*

$$
\begin{aligned}
&\|e\|_l \leqq Ch^{k+1-l}\|u\|_{k+1}, \qquad l=0,1;\\
&|e(x_j)| \leqq Ch^{2k}\|u\|_{k+1}, \qquad j=1,\ldots,N-1;\\
&\|e\|_{L^\infty(I)} \leqq Ch^{k+1}\|u\|_{k+1}.
\end{aligned}
\tag{6}
$$

Proof. See [6], [4] and [7]. □

In the next §, 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

$$
(1-\sigma_l^2)\frac{d}{d\sigma}P_k(\sigma_l) = 0, \qquad l=0,\ldots,k, \tag{7}
$$

where $P_k(\sigma)$ is the k-th degree Legendre polynomial. Associated to this polynomial is the quadrature formula (see [1, formula 25.4.32])

$$\int_{-1}^{+1} f(\sigma)\,d\sigma = \sum_{l=0}^{k} w_l f(\sigma_l) - \frac{(k+1)\,k^3\,2^{2k+1}[(k-1)!]^4}{(2k+1)[(2k)!]^3} f^{(2k)}(s),\ s\in(-1,+1)$$
$$w_l = \frac{2}{k(k+1)\,[P_k(\sigma_l)]^2},\qquad l=0,\ldots,k. \tag{8}$$

From (7) and (8), we define

$$\xi_{jl} = x_{j-1} + \frac{h}{2}(1+\sigma_l);\qquad l=0,\ldots,k;\ j=1,\ldots,N;$$
$$(\alpha,\beta)_j^* = \frac{h}{2}\sum_{l=0}^{k} w_l\,\alpha(\xi_{jl})\,\beta(\xi_{jl});\qquad \alpha,\beta\in W^{2k,\infty}(I_j);\qquad j=1,\ldots,N; \tag{9}$$
$$(\alpha,\beta)_h = \sum_{j=1}^{N} (\alpha,\beta)_j^*.$$

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

$$B(e,V)=0,\qquad V\in M_0^k(\Delta). \tag{10}$$

For any I_j, we define

$$M_0^k(I_j) = \{V \mid V\in M_0^k(\Delta),\ \mathrm{supp}\,(V)=I_j\}. \tag{11}$$

We temporarily drop the subscript j from the numbers ξ_{lj}. We define a natural basis $\{\phi_i\}_{i=1}^{k-1}$ for $M_0^k(I_j)$ by

$$\phi_i(\xi_l)=\delta_{il},\qquad 1\leqq i,\ 1\leqq k-1, \tag{12}$$

where δ_{il} is the Kronecker symbol. If we elaborate (10) for $V=\phi_i$, $i=1,\ldots,k-1$, we get

$$(e, L\phi_i) = [p(x)\,e(x)\,\phi_i'(x)]_{\xi_0}^{\xi_k},\qquad i=1,\ldots,k-1. \tag{13}$$

Approximation of $(e, L\phi_i)$ by Lobatto quadrature yields

$$\sum_{l=1}^{k-1} w_l L\phi_i(\xi_l)\,e(\xi_l)$$
$$= 2h^{-1}[p(x)\,e(x)\,\phi_i'(x)]_{\xi_0}^{\xi_k} - w_0\,e(\xi_0)\,L\phi_i(\xi_0) \tag{14}$$
$$- w_k\,e(\xi_k)\,L\phi_i(\xi_k) + Ch^{2k}D^{2k}(eL\phi_i)(\xi\in I_j),\qquad i=1,\ldots,k-1.$$

This is a linear system for $e(\xi_1),\ldots,e(\xi_{k-1})$. We have to prove the non-singularity of $(w_l L\phi_i(\xi_l))$ and to compute the order of the solution. We know that

$$hB(\phi_i,\phi_l) = h(L\phi_i,\phi_l)$$
$$= h^2\sum_{\nu=1}^{k-1} w_\nu L\phi_i(\xi_\nu)\,\phi_l(\xi_\nu) + Ch^{2k+2}D^{2k}(L\phi_i(\xi)\,\phi_l(\xi)),\ \xi\in I_j$$
$$= h^2\,w_l L\phi_i(\xi_l) + Ch^{2k+2}D^{2k}(L\phi_i(\xi)\,\phi_l(\xi)),\ \xi\in I_j.$$

Hence we have

$$|hB(\phi_i, \phi_l) - h^2 w_l L\phi_i(\xi_l)| \leqq Ch^2. \tag{15}$$

This means that $M_1 = (h^2 w_l L\phi_i(\xi_l))$ 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 $(hB(\phi_i, \phi_l))$ by M_2, we find that

$$M_1 = M_2 + h^2 M_3 = M_2(I + h^2 M_2^{-1} M_3).$$

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

$$M_1^{-1} = M_2^{-1} + h^2 M_4,$$

where the entries of M_4 are of $O(1)$. This proves that M_2^{-1} has entries of $O(1)$ and so we have that $(w_l L\phi_i(\xi_l))^{-1}$ 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 $O(h^{2k-2}\|u\|_{k+1})$. For the last term, we prove that

$$\|D^{2k}(eL\phi_i)\|_{L^\infty(I_j)} \leqq C\|e\|_{W^{2k,\infty}(I_j)} \|L\phi_i\|_{W^{2k,\infty}(I_j)}. \tag{16}$$

From [3], it can be proved that

$$\|D^l e\|_{L^\infty(I_j)} \leqq \begin{cases} Ch^{k+1-l}\|u\|_{k+1}, & l \leqq k; \\ \|D^l u\|_{L^\infty(I_j)}, & l > k. \end{cases} \tag{17}$$

Furthermore,

$$\|L\phi_i\|_{W^{2k,\infty}} \leqq Ch^{-k}, \tag{18}$$

hence we summarily have

$$\left|\sum_{l=1}^{k-1} w_l L\phi_i(\xi_l)\, e(\xi_l)\right| \leqq Ch^k[\|u\|_{k+1} h^{k-2} + \|u\|_{W^{2k,\infty}(I_j)}], \quad i = 1, \ldots, k-1. \tag{19}$$

This was the last step in the proof of

Theorem 1. *Let* $u \in H_0^1(I) \cap H^{k+1}(I) \cap \bigcap_{j=1}^{N} W^{2k,\infty}(I_j)$ *be the solution of* (1) *and let* $U \in M_0^k(\Delta)$ *be the solution of* (3). *Then the error function has the local error bound.*

$$|e(\xi_{jl})| \leqq Ch^{k+2}[\|u\|_{k+1} h^{k-2} + \|u\|_{W^{2k,\infty}(I_j)}], \quad j = 1, \ldots, N;\ l = 1, \ldots, k-1. \tag{20}$$

□

3. Lobatto Quadrature

Usually, $B(\ ,\ )$ and $(\ ,\ )$ 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

$$B_h(\alpha, \beta)=(p\alpha', \beta')_h+(q\alpha, \beta)_h; \qquad \alpha, \beta \in \bigcap_{j=1}^{N} W^{2k,\infty}(I_j), \tag{21}$$

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

Lemma 2. *Let* $Y \in M_0^k(\Delta)$ *be the solution of*

$$B_h(Y, V)=(f, V)_h, \qquad V \in M_0^k(\Delta) \tag{22}$$

and let $u \in H_0^1(I) \cap H^{k+1}(I) \cap \bigcap_{j=1}^{N} W^{2k,\infty}(I_j)$ *be the solution of* (1). *Then the error function* $\eta = u - Y$ *has the bounds*

$$|\eta(x_j)| \leqq C h^{2k} \|f\|_{2k,\Delta}; \qquad j=1, \ldots, N-1,$$

if h *is small enough, with*

$$\|f\|_{l,\Delta}=\left[\sum_{j=1}^{N} \|f\|^2_{H^l(I_j)}\right]^{\frac{1}{2}}. \tag{23}$$

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

$$\begin{aligned} |B(\varepsilon, V)| &\leqq |(f, V)-(f, V)_h|+|B_h(Y, V)-B(Y, V)| \\ &\leqq C h^{2k+1} \|V\|_{H^k(I_j)}[\|f\|_{H^{2k}(I_j)}+\|Y\|_{H^k(I_j)}], \qquad V \in M_0^k(I_j). \end{aligned}$$

If we take for V any of the basis functions ϕ_i of $M_0^k(I_j)$, as defined by (12), we have

$$|B(\varepsilon, \phi_i)| \leqq C h^{k+1}[\|f\|_{H^{2k}(I_j)}+\|Y\|_{H^k(I_j)}], \qquad i=1, \ldots, k-1. \tag{25}$$

Since

$$\begin{aligned} \sum_{l=1}^{k-1} w_l\, \varepsilon(\xi_l)\, L\phi_i(\xi_l) &= 2h^{-1} B(\varepsilon, \phi_i) \\ &\quad - w_0\, \varepsilon(\xi_0)\, L\phi_i(\xi_0) - w_k\, \varepsilon(\xi_k)\, L\phi_i(\xi_k) \\ &\quad - \frac{2}{h}[p(x)\, \varepsilon(x)\, \phi_i'(x)]_{\xi_0}^{\xi_k} + C h^{2k} D^{2k}(\varepsilon L\phi_i)\, (\xi \in I_j) \end{aligned} \tag{26}$$

and

$$\begin{aligned} \|D^{2k}(\varepsilon L\phi_i)\|_{L^\infty(I_j)} &\leqq C \|\varepsilon\|_{W^{k,\infty}(I_j)} \|\phi_i\|_{W^{k,\infty}(I_j)} \\ &\leqq C h^{-2k} \|\varepsilon\|_{L^\infty(I_j)} \leqq C h^{-k+1} \|f\|_{2k,\Delta}, \end{aligned} \tag{27}$$

we have

$$\left|\sum_{l=1}^{k-1} w_l \varepsilon(\xi_l) L\phi_i(\xi_l)\right| \leqq C_1 h^k [\|f\|_{H^{2k}(I_j)} + \|Y\|_{H^k(I_j)}] + C_2 h^{2k-2} \|f\|_{2k,\Delta} + C_3 h^{k+1} \|f\|_{2k,\Delta}. \tag{28}$$

The nonsingularity of $(w_l L\phi_i(\xi_l))$ has already been proved, its inverse is of $O(h^2)$, hence we have

$$|\varepsilon(\xi_l)| \leqq C_1 h^{k+2} [\|f\|_{H^{2k}(I_j)} + \|Y\|_{H^k(I_j)}] + C_2 h^{k+3} \|f\|_{2k,\Delta}. \tag{29}$$

Since (see [3]).

$$\|Y\|_{H^k(I_j)} \leqq \|\eta\|_{H^k(I_j)} + \|u\|_{H^k(I_j)} \leqq Ch\|u\|_{k+1} + \|u\|_{H^k(I_j)} \leqq C\|u\|_{k+1}, \tag{30}$$

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

Theorem 2. *Let* $u \in H_0^1(I) \cap H^{k+1}(I) \cap \bigcap_{j=1}^{N} W^{2k,\infty}(I_j)$ *be the solution of* (1) *and let* $Y \in M_0^k(\Delta)$ *be the solution of* (22). *Then the error function* η *has the bounds*

$$|\eta(\xi_{lj})| \leqq C_1 h^{k+2} [\|f\|_{H^{2k}(I_j)} + \|u\|_{k+1}] + C_2 h^{k+3} \|f\|_{2k,\Delta};$$
$$j = 1, \ldots, N; \quad l = 1, \ldots, k-1. \quad \square$$

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 $O(h^4)$ 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

1. Abramowitz, A., Stegun, I.: Handbook of mathematical functions. Dover Publications, 1968
2. Bakker, M.: On the numerical solution of parabolic equations in a single space variable by the continuous time Galerkin method. SIAM J. Num. Anal. **17**, 161–177 (1980)
3. Ciarlet, P.G., Raviart, P.A.: General Lagrange and Hermite interpolation in R^N with applications to finite element methods. Arch. Rational. Mech. Anal. **46**, 177–199 (1972)
4. Douglas, J. Jr., Dupont, T.: Galerkin approximations for the two-point boundary problem using continuous, piecewise polynomial spaces. Num. Mat. **22**, 99–109 (1974)

5. Lesaint, P., Zlamal, M.: Superconvergence of the gradient of finite element solutions. R.A.I.R.O. **13**, 139–166 (1979)
6. Strang, G., Fix, G.J.: An analysis of the finite element method. Englewood Cliffs, New Jersey: Prentice-Hall 1973
7. Wheeler, M.F.: An optional L_∞ error estimate for Galerkin approximations to solutions of two-point boundary problems. SIAM J. Num. Anal. **10**, 914–917 (1973)

Received December 12, 1979; Revised April 12, 1981

SIAM J. NUMER. ANAL.
Vol. 19, No. 3, June 1982

0036-1429/82/1903-0010 $01.00/0

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 $2m$th order spatial operator, a kth degree Faedo–Galerkin method is developed which has local convergence of order $2(k+1-m)$ at the knots for the first $m-1$ spatial derivatives and, if $k \geqq 2m$, convergence of order $k+2$ at specific interior nodal points. These nodal points are the zeros of the Jacobi polynomial $P_n^{m,m}(\sigma)$ $(n = k+1-2m)$ shifted to the segments of the partition. All these convergence properties are preserved if suitable quadrature rules are used.

1. Introduction.

We consider the $2m$th order initial boundary problem

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

$$Lu=\sum_{l=0}^{m}(-1)^l\frac{\partial}{\partial x^l}\left[p_l(x)\frac{\partial^l u}{\partial x^l}\right], \tag{1.1}$$

$$\frac{\partial^l u}{\partial x^l}=0, \qquad x=\pm 1, \quad l=0,\cdots,m-1, \quad t\in J,$$

$$u(0,x)=u_0(x).$$

We suppose that $p_0,\cdots,p_m$ 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 $H^l(E)$, $l\geqq 0$, and their norms by

$$W^l(E)=\{v\,|\,D^j v\in L^\infty(E), j=0,\cdots,l\},$$

$$H^l(E)=\{v\,|\,D^j v\in L^2(E), j=0,\cdots,l\}, \tag{1.2}$$

$$\|v\|_{W^l(E)}=\max_{j=0,\cdots,l}\|D^j v\|_{L^\infty(E)},$$

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

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

$$(\alpha,\beta)_E=\int_E \alpha(x)\,\overline{\beta(x)}\,dx, \qquad \alpha,\beta\in L^2(E). \tag{1.3}$$

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

$$\|\alpha\|_l=\|\alpha\|_{H^l(I)}, \qquad (\alpha,\beta)=(\alpha,\beta)_I. \tag{1.4}$$

Furthermore, we define $H_0^m(I)$ and the bilinear functional $B: H_0^m(I)\times H_0^m(I)\to\mathbb{C}$ by

$$H_0^m(I)=\{v\,|\,v\in H^m(I); D^l v(\pm 1)=0, l=0,\cdots,m-1\},$$

$$B(u,v)=(Lu,v)=(u,Lv)=\sum_{l=0}^{m}(p_l D^l u, D^l v), \qquad u,v\in H_0^m(I). \tag{1.5}$$

* 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 $p_0, \cdots, p_m$ are such that B is strongly coercive; i.e., that there exist positive constants C_1 and C_2 depending on $p_0, \cdots, p_m$ only, such that

$$|B(u, v)| \leqq C_1 \|u\|_m \|v\|_m, \qquad u, v \in H_0^m(I),$$
$$B(v, v) \geqq C_2 \|v\|_m^2, \qquad v \in H_0^m(I). \tag{1.6}$$

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 Faedo–Galerkin method. Let $N \geqq 2$ be a constant integer and define the partition $\Delta = \{x_j\}_{j=0}^N$ of I by

$$h = \frac{2}{N},$$
$$x_j = -1 + hj, \qquad j = 0, \cdots, N, \tag{1.7}$$
$$I_j = [x_{j-1}, x_j], \qquad j = 1, \cdots, N.$$

Let $k \geqq 2m - 1$ be a constant integer. Then we define the finite element space $S(\Delta) \subset H_0^m(I)$ by

$$S(\Delta) = \{V \mid V \in H_0^m(I);\ V \in P_k(I_j),\ j = 1, \cdots, N\}, \tag{1.8}$$

where for any $l \geqq 0$ $P_l(E)$ 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:

$$r = k + 1 - m, \quad n = k + 1 - 2m, \quad M = rN - m. \tag{1.9}$$

In (1.9) n is the number of interior nodal points of $S(\Delta)$ on I_j and M is the dimension of $S(\Delta)$.

In connection with Δ, we define the partition spaces $W^l(\Delta)$ and $H^l(\Delta)$ together with their norms by

$$W^l(\Delta) = \{v \mid v \in W^l(I_j);\ j = 1, \cdots, N\},$$
$$\|v\|_{W^l(\Delta)} = \max_{j=1,\cdots,N} \|v\|_{W^l(I_j)},$$
$$H^l(\Delta) = \{v \mid v \in H^l(I_j);\ j = 1, \cdots, N\}, \tag{1.10}$$
$$\|v\|_{l,\Delta} = \left[\sum_{j=1}^N \|v\|_{H^l(I_j)}^2\right]^{1/2}.$$

After these preliminary definitions, we can define a finite element solution of (1.1). Let $U: J \to S(\Delta)$ be the solution of the initial boundary problem

$$\left(\frac{\partial U}{\partial t}, V\right) + B(U, V) = 0, \qquad V \in S(\Delta), \quad t \geqq 0,$$
$$U(0, x) = U_0(x), \tag{1.11}$$

where $U_0 \in S(\Delta)$ is an approximation of u_0 satisfying

$$\|u_0 - U_0\|_l \leqq Ch^{k+1-l} \|u_0\|_{k+1}, \qquad l = 0, \cdots, m. \tag{1.12}$$

LEMMA 1. *Let $u: J \to H_0^m(I) \cap H^{k+1}(I)$ 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*

$$\|e(t)\|_0 \leqq Ch^{k+1} * \left[\|u(t)\|_{k+1} + e^{-\lambda_1 t} \left\{ \|u_0\|_{k+1} + \int_0^t e^{\lambda_1 \tau} \|Lu(\tau)\|_{k+1}\, d\tau \right\} \right], \tag{1.13}$$

where λ_1 is the smallest eigenvalue of L.

Proof. See [11]. □

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 $u(t) \in H_0^m(I) \cap H^{k+1}(I) \cap W^{2r}(\Delta)$, $t \in J$. In that case $D^l e(t, x_j)(l = 0, \cdots, m-1, j = 1, \cdots, N-1)$ is of $O(h^{2r})$, $t > 0$. Furthermore, if $n \geqq 1$, there are on each I_j n specific interior points, where $e(t)$ is of $O(h^{k+2})$, 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 $k \geqq 2$, 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

$$\mathcal{L}v(s, x) = \hat{v}(s, x) = \int_0^\infty e^{-st} v(t, x)\, dt, \tag{2.1}$$

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)

$$L\hat{u} + s\hat{u} = u_0, \qquad x \in I, \tag{2.2a}$$

$$B(\hat{u}, v) + s(\hat{u}, v) = (u_0, v), \qquad v \in H_0^m(I), \tag{2.2b}$$

and for $\hat{U}$ the weak Galerkin form

$$B(\hat{U}, V) + s(\hat{U}, V) = (U_0, V), \qquad V \in S(\Delta). \tag{2.3}$$

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 will 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 μ, γ and D be positive parameters, let s be a complex number and let the following inequalities hold:*

$$\begin{aligned} &|x_1 + sx_2| \leqq D\sqrt{x_2}, \quad x_1 \geqq \gamma x_2, \quad s = -\alpha + i\beta, \\ &\mu \leqq \alpha \leqq |\beta| + \mu, \qquad 0 < \mu < \gamma. \end{aligned} \tag{2.4}$$

Then x_1 and x_2 have the bounds

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

(2.5)

$$x_2 \leqq \begin{cases} \dfrac{D^2}{(\gamma-\alpha)^2+\beta^2}, & \text{if } \alpha \leqq \gamma, \\ \dfrac{D^2}{\beta^2}, & \text{if } \alpha \geqq \gamma. \end{cases}$$

Proof. We substitute

$$x_1 = y_1 + \alpha y_2, \qquad x_2 = y_2. \tag{2.6}$$

Then, for y_1 and y_2, we have the inequalities

$$y_1^2 + \beta^2 y_2^2 \leqq D^2 y_2, \quad y_1 \geqq (\gamma-\alpha)y_2, \quad y_2 \geqq 0, \tag{2.7}$$
$$\mu \leqq \alpha \leqq |\beta| + \mu, \qquad \mu < \gamma,$$

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 β delivers (2.5). □

We turn to the problems (2.2) and (2.3). Let μ 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

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

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

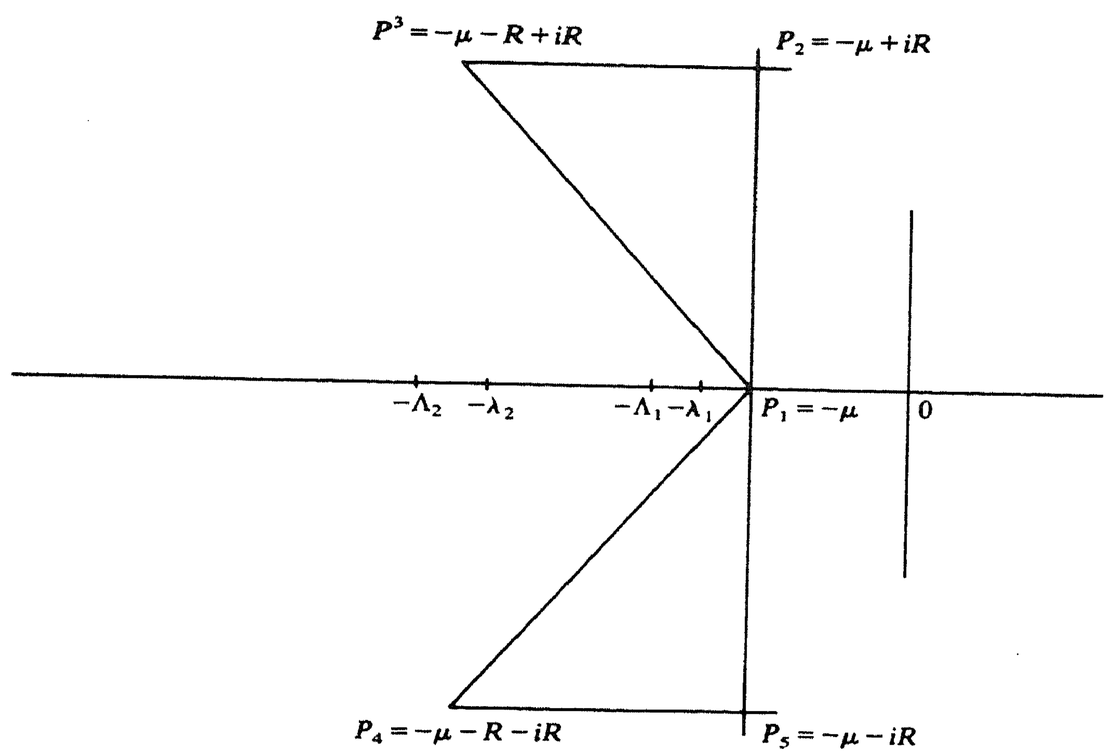


FIG. 1

LEMMA 3. *Let $e(t)=u(t)-U(t)$ and $\hat{e}=\hat{u}-\hat{U}$, 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*

$$D^l e(t,x)=\frac{1}{2\pi i}\lim_{R\to\infty}\int_{\overline{P_4P_1P_3}} D^l\hat{e}(s,x)\exp(st)\,ds$$

$$=\frac{e^{-\mu t}}{\pi}\int_0^\infty e^{-\alpha t}\operatorname{Im}\left[(1-i)\,e^{-i\alpha t}D^l\hat{e}(-\alpha-\mu-i\alpha,x)\right]d\alpha, \tag{2.9}$$

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

Proof. It is known [11] that

$$\hat{u}(s,x)=\sum_{i=1}^{\infty}(u_0,\phi_i)\frac{\phi_i(x)}{s+\lambda_i},\qquad \hat{U}(s,x)=\sum_{i=1}^{M}(U_0,\Phi_i)\frac{\Phi_i(x)}{s+\Lambda_i}, \tag{2.10}$$

where $\lambda_1,\lambda_2,\cdots$, are the positive eigenvalues of L in nondecreasing order, with orthonormal eigenfunctions $\phi_1,\phi_2,\cdots$, and where $\Lambda_1,\Lambda_2,\cdots,\Lambda_M$ (in nondecreasing order) and $\Phi_1,\Phi_2,\cdots,\Phi_M$ are the positive eigenvalues and eigenfunctions of the problem

$$B(\Phi_i,V)=\Lambda_i(\Phi_i,V),\qquad V\in S(\Lambda),\quad i=1,\cdots,M.$$

Note that

$$\Lambda_1=\inf_{V\in S(\Delta)}\frac{B(V,V)}{(V,V)}>\inf_{V\in H_0^m(I)}\frac{B(v,v)}{(v,v)}=\lambda_1. \tag{2.11}$$

From (2.10), we see that $D^l\hat{e}$ 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 $\overline{P_1P_2P_3P_1}$ and $\overline{P_1P_4P_5P_1}$ we have by Cauchy's theorem

$$\int_{\overline{P_1P_2P_3P_1}} D^l\hat{e}(s,x)\exp(st)\,ds=\int_{\overline{P_1P_4P_5P_1}} D^l\hat{e}(s,x)\exp(st)\,ds=0. \tag{2.12}$$

Furthermore, since $\overline{P_5P_1P_2}$ lies in the convergence half-plane of $\hat{e}$, we can apply the complex inversion formula [6] to obtain

$$D^l e(t,x)=\lim_{R\to\infty}\frac{1}{2\pi i}\int_{\overline{P_5P_1P_2}} D^l\hat{e}(s,x)\exp(st)\,ds. \tag{2.13}$$

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

$$\lim_{R\to\infty}\int_{\overline{P_2P_3}} D^l\hat{e}(s,x)\exp(st)\,ds=\lim_{R\to\infty}\int_{\overline{P_4P_5}} D^l\hat{e}(s,x)\exp(st)\,ds=0. \tag{2.14}$$

From (2.2), we can derive that

$$|B(\hat{u},\hat{u})+s(\hat{u},\hat{u})|=|(u_0,\hat{u})|\leqq\|u_0\|_0\|\hat{u}\|_0. \tag{2.15}$$

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

$$|B(\hat{u},\hat{u})|\leqq\frac{1}{2}\|u_0\|_0^2\frac{[\alpha+\sqrt{\alpha^2+R^2}]}{R^2},\qquad |D^l\hat{u}(x)|\leqq C\|\hat{u}\|_m\leqq CR^{-1/2}\|u_0\|_0, \tag{2.16a}$$

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

$$|D^l\hat{U}(s,x)|\leqq CR^{-1/2}\|u_0\|_0,\qquad l=0,\cdots,m-1, \tag{2.16b}$$

if $R \to \infty$ and $s = \pm iR - \alpha - \mu$. 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 $\hat{e}$ immediately to $e(t)$. Since these properties are not standard if $|s| \to \infty$, 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 α 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 $\hat{e} = \hat{u} - \hat{U}$ has the bound*

$$\|\hat{e}\|_l \leqq C(\alpha) h^{k+1-l} \|u_0\|_{k+1}, \qquad l = 0, \cdots, m. \tag{2.17}$$

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

$$B(\hat{u} - \hat{U}, V) + s(\hat{u} - \hat{U}, V) = (u_0 - U_0, V), \qquad V \in S(\Delta). \tag{2.18}$$

Next, we introduce the elliptic projection $\hat{U}_2 \in S(\Delta)$ of $\hat{u}$ by

$$B(\hat{u} - \hat{U}_2, v) = 0, \qquad V \in S(\Delta). \tag{2.19}$$

It is standard [11] that $\|\hat{u} - \hat{U}_2\|_l \leqq Ch^{k+1-l} \|\hat{u}\|_{k+1}$, $l = 0, \cdots, m$. If we put $v = \hat{\varepsilon} = \hat{U}_2 - \hat{U}$ and subtract (2.19) from (2.18), we find

$$|B(\hat{\varepsilon}, \hat{\varepsilon}) + s(\hat{\varepsilon}, \hat{\varepsilon})| = |(u_0 - U_0 - s(\hat{u} - \hat{U}_2), \hat{\varepsilon})| \leqq C\|\hat{\varepsilon}\|_0 h^{k+1} (\|u_0\|_{k+1} + |s| \|\hat{u}\|_{k+1}). \tag{2.20}$$

Application of Lemma 2 to (2.20) yields

$$\begin{aligned} B(\hat{\varepsilon}, \hat{\varepsilon}) &\leqq C(\alpha) h^{2(k+1)} (\|u_0\|_{k+1} + |s| \|\hat{u}\|_{k+1})^2, \\ \|\hat{\varepsilon}\|_l \leqq \|\hat{\varepsilon}\|_m &\leqq C(\alpha) h^{k+1} (\|u_0\|_{k+1} + |s| \|\hat{u}\|_{k+1}). \end{aligned} \tag{2.21}$$

We now have

$$\begin{aligned} \|\hat{e}\|_l &\leqq \|\hat{u} - \hat{U}_2\|_l + \|\hat{\varepsilon}\|_l \\ &\leqq Ch^{k+1-l} [\|\hat{u}\|_{k+1} + C(\alpha) h^l (\|u_0\|_{k+1} + |s| \|\hat{u}\|_{k+1})], \qquad l = 0, \cdots, m. \end{aligned} \tag{2.22}$$

We still need an estimation of $\|\hat{u}\|_{k+1}$. From (2.2), we can derive that, since $L\hat{u} \in H_0^m(I)$,

$$|B(L\hat{u}, L\hat{u}) + s(L\hat{u}, L\hat{u})| = |(Lu_0, L\hat{u})| \leqq \|Lu_0\|_0 \|L\hat{u}\|_0.$$

Application of Lemma 2 yields

$$\begin{aligned} \|\hat{u}\|_{3m} &\leqq C\|L\hat{u}\|_m \leqq C_1(\alpha)\|Lu_0\|_0 \leqq C_1(\alpha)\|u_0\|_{2m}, \\ \|\hat{u}\|_{2m} &\leqq C\|L\hat{u}\|_0 \leqq C_2(\alpha)\|Lu_0\|_0 \leqq C_2(\alpha)\|u_0\|_{2m}. \end{aligned} \tag{2.23}$$

Since

$$\|D^l L\hat{u}\|_0 \leqq |s| \|D^l \hat{u}\|_0 + \|D^l u_0\|, \qquad l = 0, \cdots, n,$$

we can prove by induction that

$$\|\hat{u}\|_{k+1} \leqq C(\alpha)\|u_0\|_{k+1}. \tag{2.24}$$

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 $\mu \uparrow \lambda_1$.

Now that we have established convergence of $\hat{e}$ on the contour $\overline{P_4P_1P_3}$, we can investigate the superconvergence at the knots.

30

For any $x \in (-1, +1)$ and $l \in \{0, 1, \cdots, m-1\}$, we define the generalized Green's function $\hat{G}_l(x, \xi) \in H_0^m(I) \cap H^{k+1}(0, x) \cap H^{k+1}(x, 1)$ associated to L, by

$$\begin{aligned} L_\xi \hat{G}_l(x, \xi) + \bar{s}\hat{G}_l(x, \xi) &= 0, \qquad \xi \in I \backslash \{x\}, \\ B(v, \hat{G}_l(x)) + s(v, \hat{G}_l(x)) &= D^l v(x), \qquad v \in H_0^m(I), \end{aligned} \tag{2.25}$$

where the subscript ξ of L_ξ denotes partial differentiation with respect to ξ. If we denote

$$\hat{G}_{lj}(\xi) = \hat{G}_l(x_j, \xi), \qquad j = 1, \cdots, N-1, \quad l = 0, \cdots, m-1, \tag{2.26}$$

we find for $D^l \hat{e}(x_j)$ the bound

$$\begin{aligned} |D^l \hat{e}(x_j)| &= |B(\hat{e}, \hat{G}_{lj}) + s(\hat{e}, \hat{G}_{lj})| \\ &\leqq |B(\hat{e}, \hat{G}_{lj} - V) + s(\hat{e}, \hat{G}_{lj} - V)| + |B(\hat{e}, V) + s(\hat{e}, V)| \\ &\leqq C(\alpha)\|\hat{e}\|_m \|\hat{G}_{lj} - V\|_m + |(u_0 - U_0, V)|, \\ &\qquad V \in S(\Delta), \quad j = 1, \cdots, N-1, \quad l = 0, \cdots, m-1. \end{aligned} \tag{2.27}$$

Since $\hat{G}_{lj} \in H_0^m(I) \cap H^{k+1}(\Delta)$, we can take V such that

$$\|\hat{G}_{lj} - V\|_m \leqq Ch^r \|\hat{G}_{lj}\|_{k+1,\Delta}, \qquad \|V\|_{W^k(\Delta)} \leqq C\|G_{lj}\|_{W^{k+1}(\Delta)}. \tag{2.28}$$

Then it is easily proved from (2.17) and (2.27) that

$$\begin{aligned} |D^l \hat{e}(x_j)| &\leqq C(\alpha)h^{2r}\|u_0\|_{k+1}\|\hat{G}_{lj}\|_{k+1,\Delta} + |(u_0 - U_0, V)|, \\ &\qquad l = 0, \cdots, m-1, \quad j = 1, \cdots, N-1. \end{aligned} \tag{2.29}$$

We have yet to estimate $|(u_0 - U_0, v)|$ and $\|\hat{G}_{lj}\|_{k+1,\Delta}$. We have to choose V_0 such that $|(u_0 - U_0, V)|$ is of $O(h^{2k})$. 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 $D^l e(t, x_j)$, is simpler and which imposes rather mild extra conditions to u_0 and $u(t)$: they also have to be in $W^{2r}(\Delta)$. Although we chose Δ 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 νth degree Jacobi polynomial $P_\nu^{\alpha,\beta}(x)$ by [1], [13]

$$\begin{aligned} P_\nu^{\alpha,\beta}(x) &= [w(x)]^{-1} D^\nu [(1-x^2)^\nu w(x)], \qquad \nu \geqq 0, \\ w(x) &= (1-x)^\alpha (1+x)^\beta, \qquad x \in (-1, +1), \quad \alpha, \beta > -1. \end{aligned} \tag{2.30}$$

These polynomials have the properties [1], [13]

$$\begin{aligned} (wP_\mu^{\alpha,\beta}, P_\nu^{\alpha,\beta}) &= \delta_{\mu\nu}(wP_\nu^{\alpha,\beta}, P_\nu^{\alpha,\beta}), \qquad \mu, \nu \geqq 0, \\ P_\nu^{\alpha,\beta}(x_{\mu\nu}) &= 0, \qquad -1 < x_{1\nu} < x_{2\nu} < \cdots < x_{\nu\nu} < 1, \end{aligned} \tag{2.31}$$

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 $P_n^{m,m}(\sigma)$; i.e.,

$$P_n^{m,n}(\sigma_l) = 0, \qquad l = 1, \cdots, n. \tag{2.32}$$

Of course, (2.32) only makes sense if $n \geqq 1$. In the sequel, it is tacitly assumed that the formulae which make no sense if $n = 0$ are to be omitted.

Given a partition Δ of I, we define the points ξ_{lj} by

$$\xi_{lj} = x_{j-1} + \frac{h}{2}(1+\sigma_l), \qquad j = 1, \cdots, N, \quad l = 1, \cdots, n. \tag{2.33}$$

Next, we introduce the linear interpolation Π: $H_0^m(I) \cap W^{2r}(\Delta) \to S(\Delta)$ by

$$\begin{aligned} D^l \Pi f(x_j) &= D^l f(x_j), \qquad l = 0, \cdots, m-1, \quad j = 1, \cdots, N-1, \\ \Pi f(\xi_{lj}) &= f(\xi_{lj}), \qquad l = 1, \cdots, n, \quad j = 1, \cdots, N. \end{aligned} \tag{2.34}$$

LEMMA 5. *For any $V \in S(\Delta)$ and $f \in H_0^m(I) \cap W^{2r}(\Delta)$*

$$|(f - \Pi f, V)| \leqq C h^{2r} \|f\|_{W^{2r}(\Delta)} \|V\|_{W^k(\Delta)}. \tag{2.35}$$

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

$$\begin{aligned} (f - \Pi f, V)_{I_j} &= \frac{1}{2} h \int_{-1}^{+1} [(f - \Pi f) V]\left(\frac{1}{2}(x_{j-1} + x_j + h\sigma)\right) d\sigma \\ &= \frac{1}{2} h \int_{-1}^{+1} (1-\sigma^2)^m P_n^{m,m}(\sigma)(gV)\left(\frac{1}{2}(x_{j-1} + x_j + h\sigma)\right) d\sigma, \end{aligned}$$

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

$$|(f - \Pi f, V)_{I_j}| \leqq C h^{2r+1} \|D^{2r}(fV)\|_{L^\infty(I_j)}, \qquad j = 1, \cdots, N. \tag{2.36}$$

Elaboration of (2.36) and summation over all I_j 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

$$|D^l \hat{e}(x_j)| \leqq C(\alpha) h^{2r} \|u_0\|_{r+1} \|\hat{G}_{lj}\|_{k+1,\Delta} + |(u_0 - U_0, V)|,$$
$$j = 1, \cdots, N-1, \quad l = 0, \cdots, m-1,$$

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

$$\begin{aligned} |D^l \hat{e}(x_j)| &\leqq C(\alpha) h^{2r} [\|u_0\|_{k+1} \|\hat{G}_{l_j}\|_{k+1,\Delta} + \|u_0\|_{W^{2r}(\Delta)} \|V\|_{k,\Delta}] \\ &\leqq C(\alpha) h^{2r} \|\hat{G}_{lj}\|_{W^{k+1}(\Delta)} \|u_0\|_{W^{2r}(\Delta)}, \end{aligned} \tag{2.37}$$
$$j = 1, \cdots, N-1, \quad l = 0, \cdots, m-1.$$

It is easily proved that $\|\hat{G}_{lj}\|_{k+1,\Delta}$ is polynomially bounded, hence we can prove by combination of (2.37) and Lemma 3 that

$$|D^l e(t, x_j)| \leqq h^{2r} e^{-\mu t} \|u_0\|_{W^{2r}(\Delta)} \int_0^\infty e^{-\alpha t} C(\alpha)\, d\alpha, \qquad t > 0. \tag{2.38}$$

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

$$D^l e(0, x_j) = 0, \qquad l = 0, \cdots, m-1, \quad j = 1, \cdots, N-1,$$

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 Π leaves all members of $S(\Delta)$ invariant. This concludes the proof of the following theorem.

THEOREM 1. *Let $u: J \to H_0^m(I) \cap H^{k+1}(I) \cap W^{2r}(\Delta)$ be the solution of* (1.1) *and let $U: J \to S(\Delta)$ be the solution of* (1.11) *with U_0 defined by* (2.34). *Then the error function $e(t) = u(t) - U(t)$ 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 τ and μ only, such that*

$$|D^l e(t, x_j)| \leqq F(\tau, \mu)\, e^{-\mu t} h^{2r} \|u_0\|_{W^{2r}(\Delta)}, \qquad l = 0, \cdots, m-1, \quad j = 1, \cdots, N-1, \quad t \geqq \tau. \tag{2.39}$$

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

For any $I_j \in \Delta$, we define

$$S(I_j) = \{V \mid V \in S(\Delta);\ \text{supp}\,(V) = I_j\}. \tag{2.40}$$

It is evident that $S(I_j)$ has dimension n and that

$$D^l V(x) = 0, \qquad x \in \partial I_j, \quad V \in S(I_j), \quad l = 0, \cdots, m-1. \tag{2.41}$$

We define a basis $\{\phi_i\}_{i=1}^n$ of $S(I_j)$ by

$$\phi_i(\xi_{lj}) = \delta_{il}, \qquad 1 \leqq i, \quad l \leqq n. \tag{2.42}$$

If we apply (2.18) for $\phi_1, \cdots, \phi_n$, we find after partial integration that

$$(\hat{e}, L\phi_i + \bar{s}\phi_i)_{Ij} = (u_0 - U_0, \phi_i)_{Ij} + \sum_{l=1}^{m} \sum_{\nu=0}^{l-1} [(-1)^\nu D^\nu (p_l D^l \phi_i) D^{l-1-\nu} \hat{e}]|_{x_{j-1}}^{x_j}, \qquad i = 1, \cdots, n. \tag{2.43}$$

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

$$\int_{-1}^{+1} f(\sigma)\, d\sigma \doteq \int_{-1}^{+1} \Pi f(\sigma)\, d\sigma, \tag{2.44}$$

where $\Pi: W^{2r}(I) \to P_k(I)$ is defined by (2.34) shifted from I_j 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

$$f(\sigma) - \Pi f(\sigma) = (1 - \sigma^2)^m P_n^{m,m}(\sigma) g(\sigma),$$

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

Elaboration of (2.44) yields

$$\int_{-1}^{+1} \Pi f(\sigma)\, d\sigma = \sum_{l=0}^{m-1} [\theta_{l_1} D^l f(-1) + \theta_{l_2} D^l f(+1)] + \sum_{l=1}^{n} \omega_l f(\sigma_l), \tag{2.45}$$

where $\sigma_1, \cdots, \sigma_n$ are the zeros of $P_n^{m,m}(\sigma)$ and θ_{l_1}, θ_{l_2} and ω_l are constant weights.

By applying (2.44) to $f_l(\sigma)=(1-\sigma^2)^m P_n^{m,m}(\sigma)/(\sigma-\sigma_l)$, $l=1,\cdots,n$, one can prove that [13, Chapt. XV]

$$\omega_l=\mu_l(1-\sigma_l^2)^m, \qquad l=1,\cdots,n,$$

where $\mu_1,\cdots,\mu_n$ are the positive Gauss–Christoffel numbers for the n-point Gauss–Jacobi quadrature formula with weight function $(1-\sigma^2)^m$. This proves that $\omega_l>0$, $l=1,\cdots,n$.

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

$$(2.46)\quad (\alpha,\beta)_{I_i}^*=\frac{h}{2}\sum_{l=1}^{n}\omega_l\alpha(\xi_{l_i})\bar\beta(\xi_{l_i})+\frac{h}{2}\sum_{l=0}^{m-1}\left(\frac{h}{2}\right)^l[\theta_{l_1}D^l(\alpha\bar\beta)(x_{j-1})+\theta_{l_2}D^l(\alpha\bar\beta)(x_j)].$$

This quadrature rule has the error bound [3]

$$(2.47)\qquad |(\alpha,\beta)_{I_i}-(\alpha,\beta)_{I_i}^*|\leqq Ch^{2r+1}\|D^{2r}(\alpha\bar\beta)\|_{L^\infty(I_i)}.$$

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

$$\begin{aligned}
&\left|\sum_{l=1}^{n}h^{2m}\omega_l[L\phi_i(\xi_{l_i})+s\delta_{il}]\hat e(\xi_{l_i})\right|\\
&\quad=2h^{2m-1}*\Bigg|-\sum_{l=0}^{m-1}\left(\frac{h}{2}\right)^{l+1}\sum_{\nu=1}^{2}\theta_{l_\nu}D^l[\hat e(L\phi_i+s\phi_i)](x_{j+\nu-2})\\
&\qquad+(\hat e,L\phi_i+\bar s\phi_i)_{I_i}^*-(\hat e,L\phi_i+\bar s\phi_i)_{I_i}+(u_0-U_0,\phi_i)_{I_i}\\
&\qquad+\sum_{l=1}^{m}\sum_{\nu=0}^{l-1}[(-1)^\nu D^\nu(p_iD^l\phi_i)D^{l-1-\nu}\hat e]_{x_{j-1}}^{x_j}\Bigg|\\
(2.48)\qquad&\leqq C_1(\alpha)h^{2m}\|\phi_i\|_{W^k(I_i)}\sum_{l=0}^{m-1}\sum_{\nu=1}^{2}|D^l\hat e(x_{j+\nu-2})|\\
&\quad+C_2h^{2m+2r}\|D^{2r}(\hat e(L\phi_i+s\phi_i))\|_{L^\infty(I_i)}+C_3h^{2m+2r}\|D^{2r}(u_0\phi_i)\|_{L^\infty(I_i)}\\
&\quad+C_4h^{2m-1}\|\phi_i\|_{W^{2m-1}(I_i)}\sum_{l=0}^{m-1}\sum_{\nu=0}^{1}|D^l\hat e(x_{j+\nu-1})|\\
&\leqq C_1(\alpha)h^{k+2}\|u_0\|_{W^{2r}(\Delta)}+C_2(\alpha)h^{k+2}\|\hat e\|_{W^{2r}(\Delta)}\\
&\quad+C_3h^{k+2}\|u_0\|_{W^{2r}(\Delta)}+C_4(\alpha)h^{2r}\|u_0\|_{W^{2r}(\Delta)}\\
&\leqq C_1(\alpha)\|u_0\|_{W^{2r}(\Delta)}h^{k+2}+C_2(\alpha)\|\hat e\|_{W^{2r}(I_i)}h^{k+2},\\
&\qquad\qquad j=1,\cdots,N-1,\quad i=1,\cdots,n,
\end{aligned}$$

where we take, of course, the right and left limit for the function values at x_{j-1} and x_j, respectively. We have yet to estimate $\|\hat e\|_{W^{2r}(I_i)}$.

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

$$\begin{aligned}
\|\hat e\|_{W^{2r}(I_i)}&\leqq\|\hat U-\Pi\hat u\|_{W^k(I_i)}+|\hat u-\Pi\hat u\|_{W^{2r}(I_i)}\\
(2.49)\qquad&\leqq C_1h^{-k}\|\hat U-\Pi\hat u\|_{L^\infty(I_i)}+C_2\|\hat u\|_{W^{2r}(I_i)}\\
&\leqq C_1h^{-k}[\|\hat e\|_{L^\infty(I_i)}+\|\hat u-\Pi\hat u\|_{L^\infty(I_i)}]+C_2\|\hat u\|_{W^{2r}(I_i)}\\
&\leqq C_1h^{-k}\|\hat e\|_{L^\infty(I_i)}+C_2\|\hat u\|_{k+1}+C_3\|\hat u\|_{W^{2r}(I_i)},\qquad j=1,\cdots,N.
\end{aligned}$$

From (2.22) plus the Poincaré inequality, we can derive that

$$h^{-k}\|\hat{e}\|_{L^\infty(I_j)} \leqq Ch^{-k}\|\hat{e}\|_1 \leqq C(\alpha)[\|u_0\|_{k+1} + \|\hat{u}\|_{k+1}].$$

Hence

$$\|\hat{e}\|_{W^{2r}(I_j)} \leqq C(\alpha)[\|u_0\|_{k+1} + \|\hat{u}\|_{k+1} + \|\hat{u}\|_{W^{2r}(I_j)}]. \tag{2.50}$$

$\|\hat{u}\|_{k+1}$ was already estimated (formula (2.24)), for the estimation of $\|\hat{u}\|_{W^{2r}(I_j)}$, we simply use the differential equation (2.2a) to obtain

$$\|\hat{u}\|_{W^{2r}(I_j)} \leqq C(\alpha)\|u_0\|_{W^{2r}(I_j)}. \tag{2.51}$$

In summary, we have obtained from (2.48)–(2.51) that

$$\left|\sum_{l=1}^{n} h^{2m}\omega_l[L\phi_i(\xi_{lj}) + s\delta_{il}]\hat{e}(\xi_{lj})\right| \leqq C(\alpha)h^{k+2}\|u_0\|_{W^{2r}(\Delta)}, \qquad i = 1, \cdots, n. \tag{2.52}$$

We have to prove the solvability of the linear system (2.52). It is easily proved that

$$\left|\left(\omega_l L\phi_i(\xi_{lj}) - \frac{2}{h}B(\phi_i, \phi_l)\right)h^{2m}\right| \leqq Ch^2, \tag{2.53}$$

if h is small enough. Consequently, the matrix $(h^{2m}\omega_l L\phi_i(\xi_{lj}))$ approximates a symmetric positive definite matrix whose eigenvalues are of $O(h^0)$. This means that its eigenvalues are nearly positive, i.e., the real parts are positive of $O(h^0)$ and the imaginary parts are of $O(h^2)$. Since $\bar{s} \in P_4P_1P_3$, we can show from (2.52) by elementary matrix calculus that

$$|\hat{e}(\xi_{lj})| \leqq C(\alpha)h^{k+2}\|u_0\|_{W^{2r}(\Delta)}, \qquad l = 1, \cdots, n, \quad j = 1, \cdots, N. \tag{2.54}$$

Application of Lemma 3 to (2.54) leads to the bound

$$|e(t, \xi_{lj})| \leqq h^{k+2}\, e^{-\mu t}\|u_0\| \int_0^\infty e^{-\alpha t}C(\alpha)\, d\alpha.$$

Again, the fact that $e(0, \xi_{lj}) = 0$ 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* $k \geqq 2m$. *Then* $e(t)$ *has the bounds* (1.13) *and* (2.39) *plus the additional bound*

$$|e(t, \xi_{lj})| \leqq F(\tau, \mu)\, e^{-\mu t}h^{k+2}\|u_0\|_{W^{2r}(\Delta)}, \qquad t \leqq \tau, \quad j = 1, \cdots, N, \quad l = 1, \cdots, n. \tag{2.55}$$

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 (U_t, 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 $q \geqq 2r$ be a constant integer and let $-1 \leqq z_1 < z_2 < \cdots < z_p \leqq 1$ be p distinct points on I and let, for $f \in W^q(I)$,

$$\int_{-1}^{1} f(z)\, dz \doteq \sum_{i=1}^{p} w_i f(z_i) \tag{3.1}$$

be an approximation which is exact if $f \in P_{q-1}(I)$. Given a partition Δ of I we define, for $\alpha, \beta \in W^q(\Delta)$,

$$(\alpha, \beta)_j^* = \frac{h}{2} \sum_{i=1}^{p} w_i (\alpha\bar{\beta})\left(x_{j-1} + \frac{h}{2}(1+z_i)\right),$$

$$(\alpha, \beta)_h = \sum_{j=1}^{N} (\alpha, \beta)_j^*, \tag{3.2}$$

$$B_h(\alpha, \beta) = \sum_{l=0}^{m} (p_l D^l \alpha, D^l \beta)_h.$$

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*

$$|B(U, V) - B_h(U, V)| \leqq Ch^{q-2k+i+j} \|U\|_{i,\Delta} \|V\|_{j,\Delta}, \qquad 0 \leqq i, \quad j \leqq k. \tag{3.3}$$

Proof. Application of Bramble and Hilbert's lemma [3] gives

$$|B_h(U, V) - B(U, V)| \leqq Ch^{q+1} \sum_{j=1}^{n} \sum_{l=0}^{m} \|D^q (p_l D^l U D^l V)\|_{L^\infty(I_j)}$$

$$\leqq Ch^q \|U\|_{k,\Delta} \|V\|_{k,\Delta} \sum_{l=0}^{m} \|p_l\|_{W^q(\Delta)} \tag{3.4}$$

$$\leqq Ch^{q+i+j-2k} \|U\|_{i,\Delta} \|V\|_{j,\Delta}. \qquad \square$$

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 $v \in H^{k+1}(I) \cap H_0^m(I) \cap W^q(\Delta)$, let $V \in S(\Delta)$ be an approximation of v with the error bound*

$$\|v - V\|_l \leqq Ch^{k+1-l} \|v\|_{k+1}, \qquad l = 0, \cdots, m. \tag{3.5}$$

Then we have

$$\|V\|_{k,\Delta} \leqq C\|v\|_{k+1}. \tag{3.6}$$

Proof. Let $\Pi : H^{k+1}(\Delta) \cap H_0^m(I) \cap W^q(\Delta) \to S(\Delta)$ be defined by (2.34). Then [4]

$$\begin{aligned} \|V\|_{k,\Delta} &\leqq \|V - \Pi v\|_{k,\Delta} + \|v - \Pi v\|_{k,\Delta} + \|v\|_k \\ &\leqq C_1 h^{-k} \|V - \Pi v\|_0 + C_2 h \|D^{k+1} v\|_0 + \|v\|_k \\ &\leqq C\|v\|_{k+1} + C_1 h^{-k} [\|v - V\|_0 + \|v - \Pi v\|_0] \leqq C\|v\|_{k+1}. \end{aligned} \tag{3.7}$$

$\square$

3.2. Preservation of the orders of convergence. In this section, we shall prove that the replacement of $B(\cdot, \cdot)$ by $B(\cdot, \cdot)_h$ does not affect the validity of Theorems 1 and 2 except that the constant μ will be slightly smaller. This is due to the fact that

$$\mu < \Lambda_1^* = \inf_{V \in S(\Delta)} \frac{B_h(V, V)}{(V, V)}, \tag{3.8}$$

and Λ_1^* need no longer be greater than λ_1.

Let $Y : J \to S(\Delta)$ be the solution of the initial boundary problem

$$\left(\frac{\partial Y}{\partial t}, V\right) + B_h(Y, V) = 0, \qquad V \in S(\Delta), \quad t \in J, \quad Y(0) = U_0 = \Pi u_0, \tag{3.9}$$

where Π is defined by (2.34) and B_h by (3.2). We define

$$\eta(t) = U(t) - Y(t), \tag{3.10}$$

where U is the solution of (1.11). We again define the points $P_1, P_2, \cdots, P_5$ by (2.8) where we take care that (3.8) holds, in other words that (see Fig. 1) $\hat{\eta} = \mathscr{L}\eta(s)$ has no poles inside $\overline{P_1P_2P_3P_4P_5}$. Then we can prove, in analogy to Lemma 2, that

$$D^l\eta(t,x) = \frac{e^{-\mu t}}{\pi}\int_0^\infty e^{-\alpha t}\,\mathrm{Im}\,[(1+i)\,e^{-i\alpha t}D^l\hat{\eta}(-\alpha-\mu-i\alpha, x)]\,d\alpha, \tag{3.11}$$

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

As before, we are only interested in the case $s \in \overline{P_4P_1P_3}$. By applying $\mathscr{L}$ to (3.9) and subtracting the result from (2.3) we get

$$B_h(\hat{\eta}, V) + s(\hat{\eta}, V) = B_h(\hat{U}, V) - B(\hat{U}, V), \qquad V \in S(\Delta). \tag{3.12}$$

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

$$\begin{aligned} |B_h(\hat{\eta}, \hat{\eta}) + s(\hat{\eta}, \hat{\eta})| &\leqq Ch^{q-k+m}\|\hat{\eta}\|_m\|\hat{U}\|_{k,\Delta} \\ &\leqq Ch^{q-k+m}\|\hat{\eta}\|_m\|\hat{u}\|_{k+1} \leqq C(\alpha)h^{q-k+m}\|u_0\|_{k+1}\|\hat{\eta}\|_m. \end{aligned} \tag{3.13}$$

Since $B_h(\hat{\eta}, \hat{\eta}) \geqq \Lambda_1^*(\hat{\eta}, \hat{\eta})$ and B_h is strongly coercive, we can prove from (3.13) that

$$\|\hat{\eta}\|_m \leqq C(\alpha)h^{q-k+m}\|u_0\|_{k+1}. \tag{3.14}$$

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

$$\begin{aligned} |D^l\hat{\eta}(x_j)| &= |B(\hat{\eta}, \hat{G}_{lj}) + s(\hat{\eta}, \hat{G}_{lj})| \\ &\leqq |B(\hat{\eta}, G_{lj} - V) + s(\hat{\eta}, \hat{G}_{lj} - V)| + |B_h(\hat{Y}, V) - B(\hat{Y}, V)| \\ &\leqq C(\alpha)\|\hat{\eta}\|_m\|\hat{G}_{lj} - V\|_m + Ch^q\|\hat{Y}\|_{k,\Delta}\|V\|_{k,\Delta}. \end{aligned} \tag{3.15}$$

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

$$\|\hat{u} - \hat{Y}\|_l \leqq \|\hat{e}\|_l + \|\hat{\eta}\|_l \leqq C(\alpha)h^{k+1-l}\|\hat{u}\|_{k+1},$$

hence after application of Lemma 8, we get

$$\|\hat{Y}\|_{k,\Delta} \leqq C(\alpha)\|u\|_{k+1} \leqq C(\alpha)\|u_0\|_{k+1}. \tag{3.16}$$

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

$$|D^l\hat{\eta}(x_j)| \leqq C(\alpha)h^q\|u_0\|_{k+1}, \qquad l = 0, \cdots, m-1, \quad j = 1, \cdots, N-1, \tag{3.17}$$

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

(3.18)

$$|D^l\eta(t, x_j)| \leqq F(\tau, \mu)\,e^{-\mu t}h^q\|u_0\|_{k+1}, \qquad t \geqq \tau, \quad l = 0, \cdots, m-1, \quad j = 1, \cdots, N-1,$$

where F depends on τ and μ only. We again refer to the appendix for a proof that τ should be positive.

For the local bounds of $\eta(t)$ at the Jacobi points, we confine our attention to the case $k \geqq 2m$. Let $S(I_j)$ and ξ_{ij} be defined by (2.40) and (2.33). Then for arbitrary j,

we can prove from (3.12) that

$$(\hat{\eta}, LV+\bar{s}V) = B_h(\hat{Y}, V) - B(\hat{Y}, V) + \sum_{l=1}^{m}\sum_{\nu=0}^{l-1}[(-1)^{\nu}D^{\nu}(p_lD^lV)D^{l-1-\nu}\hat{\eta}]|_{x_{j-1}}^{x_j}, \qquad V \in S(I_j). \tag{3.19}$$

If we apply the quadrature rule (2.44) to (3.19) put $V \doteq \phi_i$, where ϕ_i is defined by (2.42) and multiply by $2h^{2m-1}$, we obtain, in analogy to (2.50),

$$\begin{aligned}
&\left|\sum_{l=1}^{n}\omega_l h^{2m}(L\phi_i(\xi_{lj})+s\delta_{il})\hat{\eta}(\xi_{lj})\right| \\
&\quad = 2h^{2m-1} * \Bigg|-\sum_{l=0}^{m-1}\left(\frac{h}{2}\right)^{l+1}\sum_{\nu=1}^{2}\theta_{l\nu}D^l(\hat{\eta}(L\phi_i+\bar{s}\phi_i))(x_{j+\nu-2}) \\
&\qquad + (\hat{\eta}, L\phi_i+\bar{s}\phi_i)_{I_j}^{*} - (\hat{\eta}, L\phi_i+\bar{s}\phi_i)_{I_j} + B_h(\hat{Y}, \phi_i) - B(\hat{Y}, \phi_i) \\
&\qquad + \sum_{l=1}^{m}\sum_{\nu=0}^{l-1}[(-1)^{\nu}D^{\nu}(p_lD^l\phi_i)D^{l-1-\nu}\hat{\eta}]|_{x_{j-1}}^{x_j}\Bigg| \\
&\quad \leqq C_1(\alpha)h^{2m}\|\phi_i\|_{W^k(I_j)}\sum_{l=0}^{m-1}\sum_{\nu=0}^{1}|D^l\hat{\eta}(x_{j-1+\nu})| \\
&\qquad + C_2h^{2m+2r}\|D^{2r}(\hat{\eta}(L\phi_i+\bar{s}\phi_i))\|_{L^{\infty}(I_j)} + C_3h^{q+2m}\sum_{l=0}^{m}\|D^q(p_lD^l\hat{Y}D^l\phi_i)\|_{L^{\infty}(I_j)} \\
&\qquad + C_4h^{2m-1}\|\phi_i\|_{W^{2m-1}(I_j)}\sum_{l=0}^{m-1}\sum_{\nu=0}^{1}|D^l\hat{\eta}(x_{j+\nu-1})| \\
&\quad \leqq C_1(\alpha)h^{q+2m-k}\|u_0\|_{k+1} + C_2(\alpha)h^{k+2}\|\hat{\eta}\|_{W^k(I_j)} \\
&\qquad + C_3h^{q+2m-k}\|\hat{Y}\|_{W^k(I_j)} + C_4(\alpha)h^q\|u_0\|_{k+1} \\
&\quad \leqq C(\alpha)h^{k+2} * [\|\hat{Y}\|_{W^k(I_j)} + \|\hat{\eta}\|_{W^k(I_j)} + \|u_0\|_{k+1}] \\
&\quad \leqq C(\alpha)h^{k+2}\|u_0\|_{k+1}, \qquad i = 1, \cdots, n.
\end{aligned} \tag{3.20}$$

For the last inequality we used Lemma 8 and the inequality

$$\|\hat{\eta}\|_{W^k(I_j)} \leqq Ch^{m-k-1}\|\hat{\eta}\|_{W^{m-1}(I_j)} \leqq Ch^{m-k-1}\|\hat{\eta}\|_m \leqq C(\alpha)h^{q-2k+2m-1}\|u_0\|_{k+1},$$

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

$$|\hat{\eta}(\xi_{ij})| \leqq C(\alpha)h^{k+2}\|u_0\|_{k+1}, \tag{3.21}$$

and application of (3.11) gives

$$|\eta(t, \xi_{ij})| \leqq F(\tau, \mu)\, e^{-\mu t}h^{k+2}\|u_0\|_{k+1}, \qquad t \geqq \tau, \tag{3.22}$$

where F depends on τ and μ only.

We have to estimate $\|\eta(t)\|_0$ yet. Since $\eta \in S(\Delta)$, this job is very easy, because all the nodal values of $\eta(t)$: $D^l\hat{\eta}(t, x_j)$ and $\hat{\eta}(t, \xi_{ij})$ have been shown to be of $O(h^{k+2}F(t, \mu)e^{-\mu t})$. This implies automatically that

$$\begin{aligned}
\|\eta(t)\|_{L^{\infty}(I)} &\leqq F_1(\tau, \mu)\, e^{-\mu t}h^{k+2}\|u_0\|_{k+1}, \\
\|\eta(t)\|_0 &\leqq F_2(\tau, \mu)\, e^{-\mu t}h^{k+2}\|u_0\|_{k+1}, \qquad t \geqq \tau,
\end{aligned} \tag{3.23}$$

where F_1 and F_2 depend on τ and μ 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 $Y: J \to S(\Delta)$ be the solution of (3.9) and let $u: J \to H_0^m(I) \cap H^{k+1}(I) \cap W^q(\Delta)$ be the solution of (1.1) with $q \geqq 2r$. Then, if h is small enough, the error function $\zeta(t) = u(t) - Y(t)$ has the bounds*

$$\|\zeta(t)\|_0 \leqq \|e(t)\|_0 + F_1(\tau,\mu)\, e^{-\mu t h^\nu} \|u_0\|_{k+1}, \qquad \nu = \min(k+2, 2r),$$

$$\|\zeta(t, x_j)\| \leqq F_2(\tau,\mu)\, e^{-\mu t h^{2r}} \|u_0\|_{W^{2r}(\Delta)}, \qquad j = 1, \cdots, N-1,$$

$$|\zeta(t, \xi_{ij})| \leqq F_3(\tau,\mu)\, e^{-\mu t h^{k+2}} \|u_0\|_{W^{2r}(\Delta)}, \qquad i = 1, \cdots, n, \quad j = 1, \cdots, N, \quad t \geqq \tau.$$

where $\|e(t)\|_0$ has the bound (1.12), μ has the bound (3.8) and where F_1, F_2 and F_3 depend on $\tau > 0$ and μ only.

4. Conclusions. In the preceding sections we saw that earlier superconvergence results [2], [7], [8], [9], [10] can be generalized to $2m$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 $\hat{e}(x)$ 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 $2m$th order elliptic problem. That the use of qth 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

$$(1) \qquad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad t \in J, \quad x \in I, \qquad u(0, x) = 1 - x^4.$$

This is an example from the class $m = 1$ with $p_0 \equiv 0$ and $p_1 \equiv 1$. 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

$$\dot{a}_0 = 0,$$

$$8\dot{a}_{2j} + 2(\dot{a}_{2j-1} + \dot{a}_{2j+1}) - (\dot{a}_{2j-2} + \dot{a}_{2j+2})$$

$$(2) \qquad = -\frac{10}{h^2}(14a_{2j} - 8(a_{2j-1} + a_{2j+1}) + a_{2j-2} + a_{2j+2}), \qquad j = 1, \cdots, N-1,$$

$$\dot{a}_{2N} = 0,$$

$$8\dot{a}_{2j-1} + \dot{a}_{2j-2} + \dot{a}_{2j} = -\frac{40}{h^2}(2a_{2j-1} - a_{2j-2} - a_{2j}), \qquad j = 1, \cdots, N, \quad t \geqq 0,$$

where the overdot stands for d/dt, a_{2j-1} stands for $U(t, x_j - (h/2))$ and a_{2j} stands for $U(t, x_j)$.

If we put

$$e_{2j-1}(t) = (u-U)\left(t, x_j - \frac{h}{2}\right), \qquad j = 1, \cdots, N,$$
$$e_{2j}(t) = (u-U)(t, x_j), \qquad j = 0, \cdots, N, \tag{3}$$

and apply (1) and (2) for $t = 0$ and subtract the results, we find for $\dot{e}_j(0)(j = 0, \cdots, 2N)$ the linear system

$$\dot{e}_0(0) = 0,$$
$$8\dot{e}_{2j}(0) + 2[\dot{e}_{2j-1}(0) + \dot{e}_{2j+1}(0)] - \dot{e}_{2j-2}(0) - \dot{e}_{2j+2}(0) = -2h^2, \qquad j = 1, \cdots, N-1,$$
$$\dot{e}_{2N}(0) = 0. \tag{4}$$
$$8\dot{e}_{2j-1}(0) + \dot{e}_{2j-2}(0) + \dot{e}_{2j}(0) = -h^2, \qquad j = 1, \cdots, N.$$

If we eliminate the odd-indexed unknowns from (4) by static condensation, we obtain the linear system

$$\dot{e}_0(0) = 0,$$
$$\dot{e}_{2j}(0) - \frac{1}{6}[\dot{e}_{2j-2}(0) + \dot{e}_{2j+2}(0)] = -\frac{1}{5}h^2, \qquad j = 1, \cdots, N-1, \tag{5}$$
$$\dot{e}_{2N}(0) = 0.$$

The matrix of system (5) has the representation

$$M = I - A, \tag{6}$$

where A has only nonnegative entries and a spectral radius smaller than 1. From this we can conclude that

$$M^{-1} = \sum_{n=0}^{\infty} A^n, \tag{7}$$

which means that M^{-1} has only nonnegative entries and hence we can conclude that

$$\dot{e}_{2j}(0) = -C_j h^2, \qquad j = 1, \cdots, N-1, \tag{8}$$

where C_j is of $O(1)$. For $\dot{e}_{2j-1}$ we can find a similar expression. All the above formulae now imply that

$$e_j(t) = O(th^2), \qquad j = 1, \cdots, 2N-1, \tag{9}$$

as $(h, t) \to (0, 0)$. From this, we can prove that

$$[e_j(t)h^{-4}]_{t=h}, \qquad j = 1, \cdots, 2N-1 \tag{10}$$

is not bounded, as $h \downarrow 0$. 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 $t \geqq 0$ was made, which can also be shown by this example.

REFERENCES

[1] M. ABRAMOWITZ AND I. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1968.

[2] M. BAKKER, *On the numerical solution of parabolic equations in a single space variable by the continuous time Galerkin method*, this Journal, 17 (1980), pp. 162–177.

[3] J. H. BRAMBLE AND S. R. HILBERT, *Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation*, this Journal, 7 (1970), pp. 112–124.

[4] P. G. CIARLET AND P. A. RAVIART, *General Lagrange and Hermite interpolation in R^N with applications of finite element methods*, Arch. Rat. Mech. Anal., 46 (1972), pp. 177–199.

[5] P. J. DAVIS AND P. RABINOWITZ, *Numerical Integration*, Blaisdell, New York–Toronto–London, 1967.

[6] G. DOETSCH, *Einführung in Theorie und Anwendung der Laplace-Tranformation*, Birkhäuser Verlag, New York, 1958.

[7] J. DOUGLAS, JR. AND T. DUPONT, *Collocation Methods for Parabolic Equations in a Single Space Variable*, Springer-Verlag, Heidelberg, 1974.

[8] J. DOUGLAS, JR., T. DUPONT AND M. F. WHEELER, *Some superconvergence results for an H^1-Galerkin procedure for the heat equation*, Rep. MRC 1382, Mathematics Research Center, University of Wisconsin, Madison, 1973.

[9] ———, *A quasi-projection approximation method applied to Galerkin procedures for parabolic and hyperbolic equations*, Rep. MRC 1461, Mathematics Research Center, University of Wisconsin, Madison, 1974.

[10] ———, *A quasi-projection analysis of Galerkin methods for parabolic and hyperbolic equations*, Math. Comp., 32 (1978), pp. 345–362.

[11] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley, New York–London–Sydney–Toronto, 1976.

[12] P. A. RAVIART, *The use of numerical integration in finite element methods for solving parabolic equations*, in Topics in Numerical Analysis, J. J. H. Miller, ed., Academic Press, London, 1973.

[13] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1959.

Computer Physics Communications 20 (1980) 429–439

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 Molecuulfysica, 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

$$v = V_\phi/U_0, \quad \phi = \Phi/(U_0 B_z R), \quad u = v/x, \quad x = r/R, \quad y = z/L,$$
$$\gamma = R/L, \quad \mathrm{Ha} = B_z R(\sigma/\eta)^{1/2}, \quad \sigma/\sigma_\perp = 1 + \beta^2, \tag{1}$$

where V_ϕ – azimuthal velocity, U_0 – characteristic velocity, Φ – electric potential, B_z – magnetic induction, R, L – radius and half length of cylinder, $\sigma_\perp, \sigma$ – normal and tangential components of conductivity, η – dynamic viscosity, γ – inverse aspect ratio, Ha – Hartmann constant, β – Hall parameter.

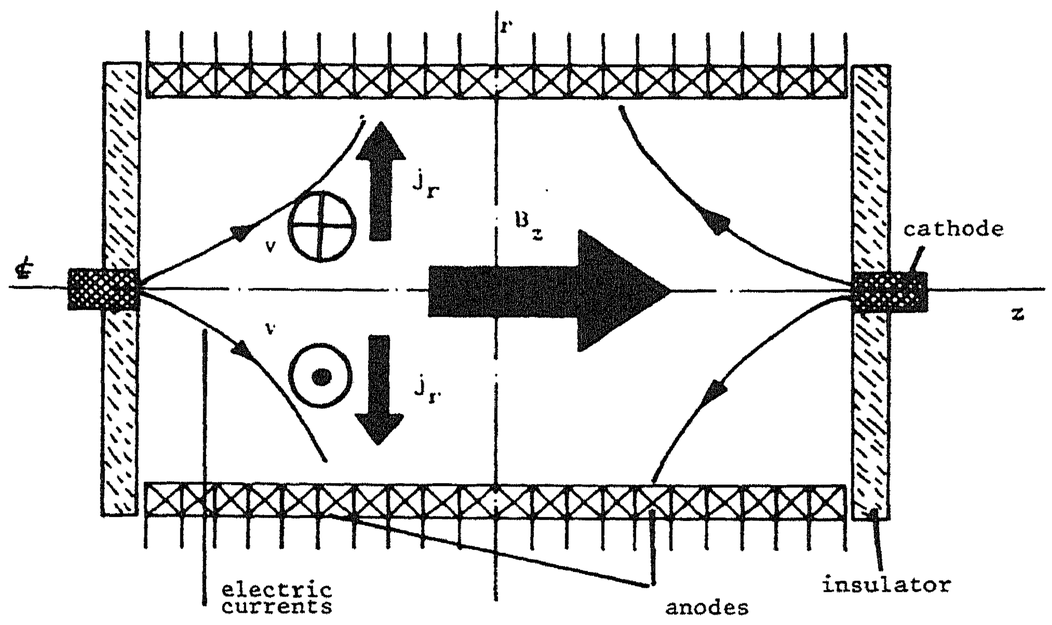


Fig. 1. The physical model of the rotating plasma problems.

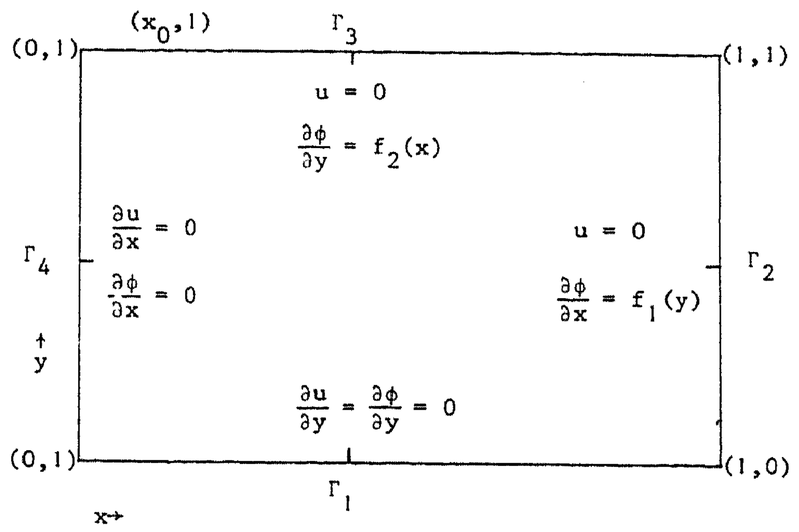


Fig. 2. Boundary conditions on Γ.

or u and ϕ, the following system of partial differential equations can be derived [1]

$$+\frac{3}{x}\frac{\partial u}{\partial x}+\gamma^2\frac{\partial^2 u}{\partial y^2}-\frac{\mathrm{Ha}^2}{1+\beta^2}\left(u-\frac{1}{x}\frac{\partial \phi}{\partial x}\right)=0, \tag{2a}$$

$$+\frac{1}{x}\frac{\partial \phi}{\partial x}+\gamma^2(1+\beta^2)\frac{\partial^2 \phi}{\partial y^2}-xu_x-2u=0, \qquad 0\leqslant x, y\leqslant 1, \tag{2b}$$

boundary conditions (see fig. 2)

$$x=0:\ \frac{\partial u}{\partial x}=\frac{\partial \phi}{\partial x}=0, \qquad \Gamma_2:\ x=1:\ u=0,\ \frac{\partial \phi}{\partial x}=f_1(y),$$

$$y=0:\ \frac{\partial u}{\partial y}=\frac{\partial \phi}{\partial y}=0, \qquad \Gamma_3:\ y=1:\ u=0,\ \frac{\partial \phi}{\partial y}=f_2(x). \tag{3}$$

rk. As a matter of fact, Van den Berg [1] derives a system of PDEs for v and ϕ. 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 $f_1(y)$ $_2(x)$ are not independent functions but satisfy the integral relation

$$(y)\,dy+\gamma^2(1+\beta^2)\int_0^1 xf_2(x)\,dx=0. \tag{4}$$

sponding to the homogeneity of the current from the cathodes, $f_2(x)$ is a piecewise constant function, ed by

$$=\begin{cases} C_2>0, & 0\leqslant x<x_0,\\ 0, & \text{elsewhere.}\end{cases} \tag{5}$$

riational formulation of the problem

e define the function spaces

$$)=\{f|\,f \text{ continuous on } R;\ \partial f/\partial x \text{ and } \partial f/\partial y \text{ continuous almost everywhere on } R\},$$

$$)=\{f|f\in C^0(R),\ f=0, \text{ on } \Gamma_2 \text{ and } \Gamma_3\} \tag{6}$$

and the (linear and bilinear) functionals

$$I(p, q) = \iint_R x^3 [p_x^2 + \gamma^2 p_y^2] \, dx \, dy + \frac{Ha^2}{1+\beta^2} \iint_R x [(xp - q_x)^2 + \gamma^2 (1+\beta^2) q_y^2] \, dx \, dy, \quad p \in C_0^0(R), \; q \in C^0(R), \tag{7}$$

$$b(q) = \frac{Ha^2}{1+\beta^2} \int_\Gamma G(\sigma) q(\sigma) \, d\sigma, \quad q \in C^0(R), \tag{8}$$

$$G(\sigma) = \begin{cases} 0, & \sigma \in \Gamma_1 \cup \Gamma_4, \\ \gamma^2 x f_2(x)(1+\beta^2), & \sigma \in \Gamma_3, \\ f_1(y), & \sigma \in \Gamma_2, \end{cases} \tag{9}$$

$$a_1(p, q) = \iint_R x^3 \left(p_x q_x + \gamma^2 p_y q_y + \frac{Ha^2}{1+\beta^2} pq \right) dx \, dy, \tag{10}$$

$$a_2(p, q) = \frac{Ha^2}{1+\beta^2} \iint_R x^2 p q_x \, dx \, dy, \tag{11}$$

$$a_3(p, q) = \gamma^2 \frac{Ha^2}{1+\beta^2} \iint_R x (p_x q_x + \gamma^2 (1+\beta^2) p_y q_y) \, dx \, dy. \tag{12}$$

Theorem 1. *If (u, ϕ) is a solution of (2)–(3), it minimizes*

$$E(p, q) = I(p, q) - 2b(q), \quad (p, q) \in C_0^0(R) \times C^0(R), \tag{13}$$

and satisfies the relations

$$a_1(u, p) - a_2(p, \phi) = 0, \quad p \in C_0^0(R), \tag{14a}$$

$$-a_2(u, q) + a_3(\phi, q) = b(q), \quad q \in C^0(R). \tag{14b}$$

Moreover, u is unique and ϕ is unique up to an additive constant.

Proof. If we multiply (2a) with $x^3 p$ and (2b) with $[xHa^2/(1+\beta^2)]q$, then after partial integration we obtain (14). By using (14), we also find that

$$E(p, q) - E(u, \phi) = I(u - p, \phi - q) \geqslant 0, \quad p \in C_0^0(R), \; q \in C_0^0(R),$$

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 $\phi - q \equiv$ constant.

Remark. One can make ϕ 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, ϕ) by minimizing $E(p, q)$ over a finite-dimensional subspace $S_0 \times S$ of $C_0^0(R) \times C^0(R)$. 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

y-interval on the input record (see section 7). For S_0, we select the subspace of S satisfying the zero boundary conditions on Γ_2 and Γ_3.

The minimization of E over the finite element space results in the system of linear equations

$$\begin{pmatrix} A_1 & -A_2 \\ -A_2^{\mathrm{T}} & A_3 \end{pmatrix}\begin{pmatrix} U \\ \Phi \end{pmatrix}=\begin{pmatrix} 0 \\ c \end{pmatrix} \tag{15}$$

with

$$A_1 = (a_1(B_i^0, B_j^0)), \quad A_2 = (a_2(B_i^0, B_j)), \quad A_3 = (a_3(B_i, B_j)), \quad c = (b(B_i)), \tag{16}$$

where $\{B_i^0\}$ and $\{B_i\}$ are the basis functions of S_0 and S, respectively, and where a_1, a_2, a_3 and b are given by (8)–(12). 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

$$\int_{x_{i-1}}^{x_i}\int_{y_{j-1}}^{y_j} x^m F(x, y)\,\mathrm{d}x\,\mathrm{d}y \approx \omega_1 F(x_{i-1}, y_{j-1}) + \omega_2 F(x_{i-1}, y_j) + \omega_3 F(x_i, y_{j-1}) + \omega_4 F(x_i, y_j), \tag{17}$$

where the weights $\omega_1, \ldots, \omega_4$ are chosen such that (17) is exact, if 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

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

From the value $\int_0^1 f_1(y)\,\mathrm{d}y$, 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 ϕ (see figs. 3 and 4), the library DISSPLA 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 ϕ 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.

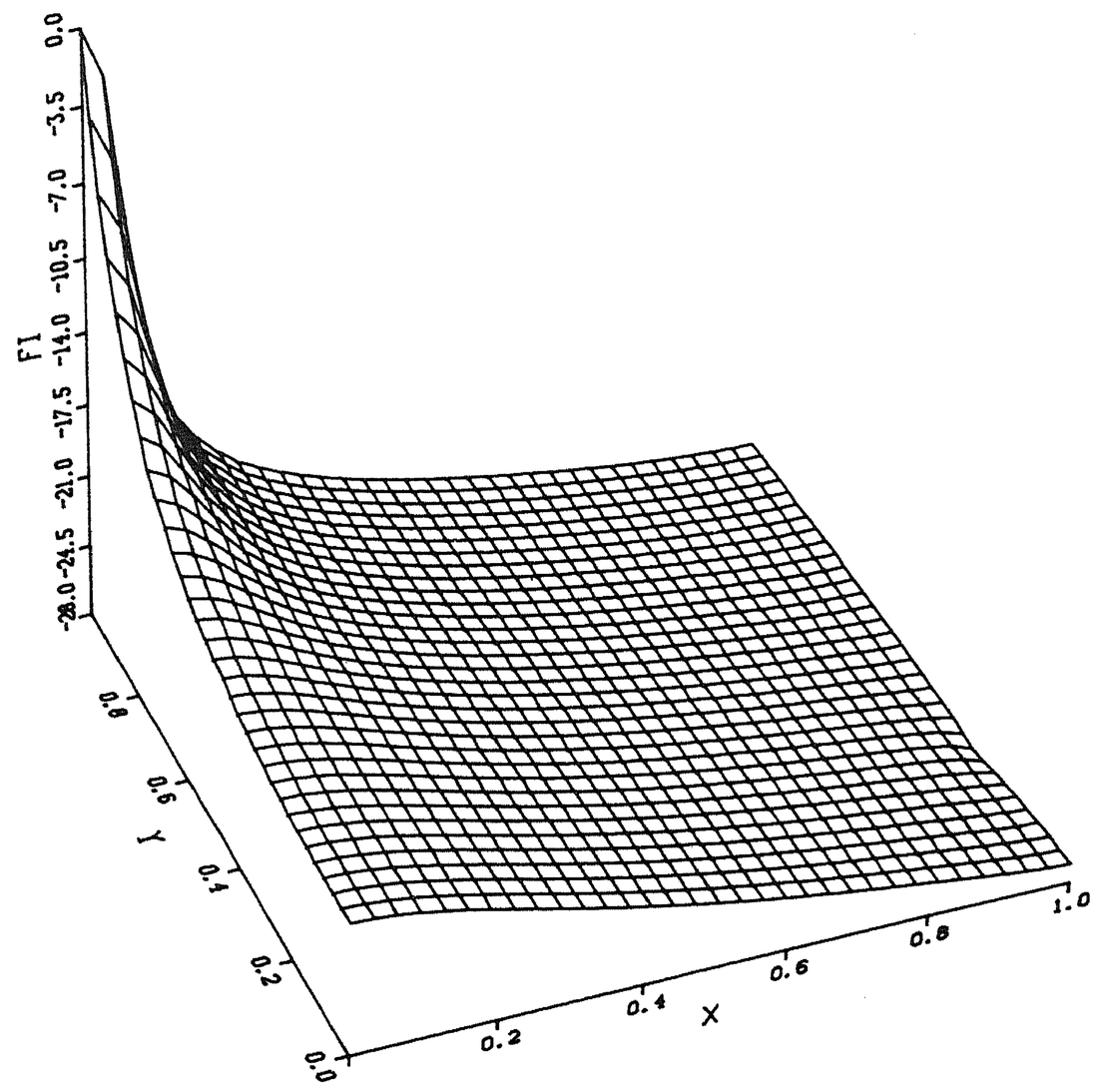


Fig. 3. Graph of $\phi(x, y)$.

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

$$x = Dy, \qquad Sy = DADy = Db, \tag{18}$$

where D is a diagonal matrix whose entries are given by

$d_i = a_{ii}^{-1/2}$;

CONGR – this subroutine solves (18) iteratively by the CG method.
MATVEC – this subroutine computes the matrix-vector product Sv, given a vector v.
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 assignated its value. Only the profile is to be given, i.e. the function value, up to a

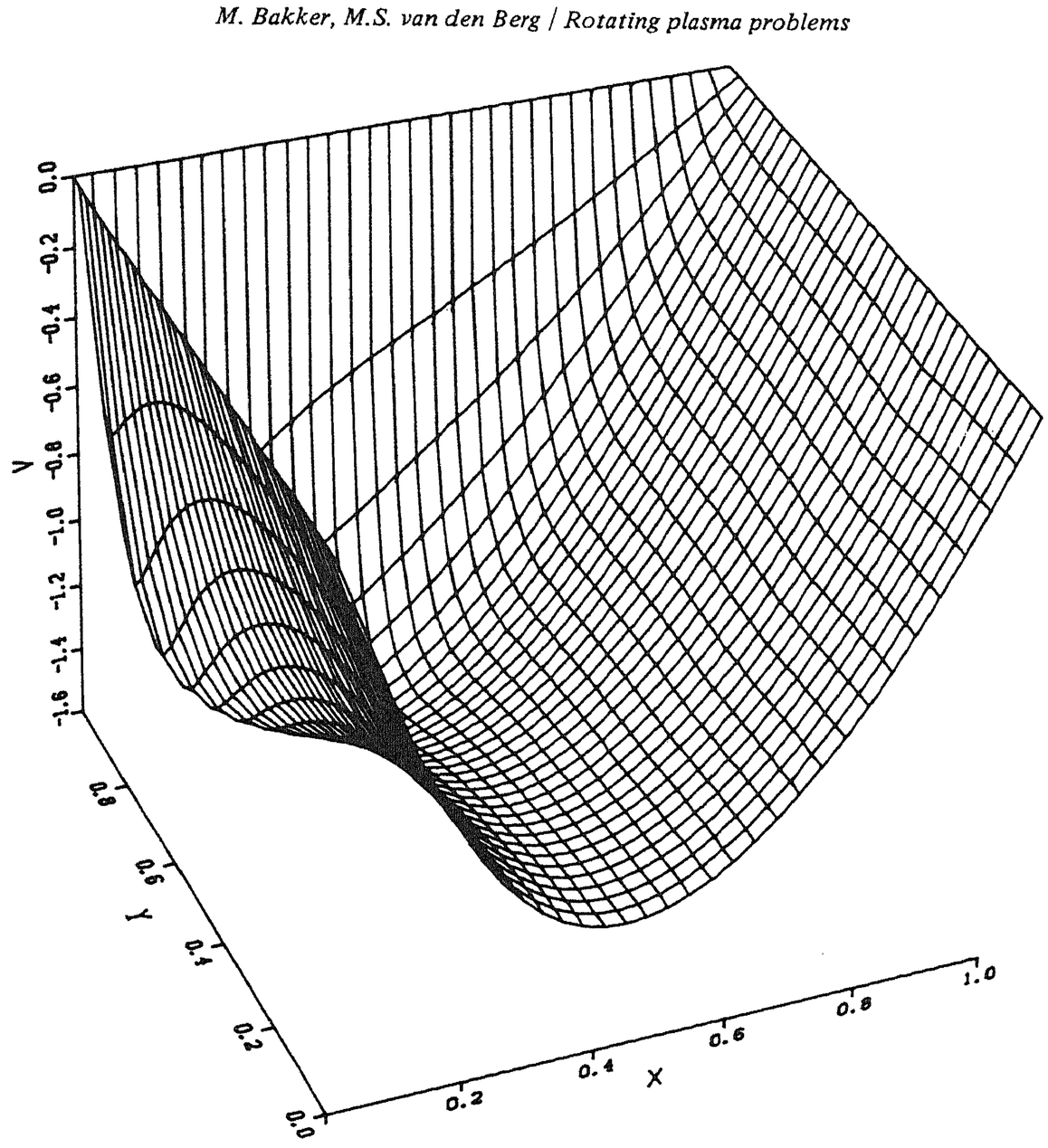


Fig. 4. Graph of $v(x, y)$.

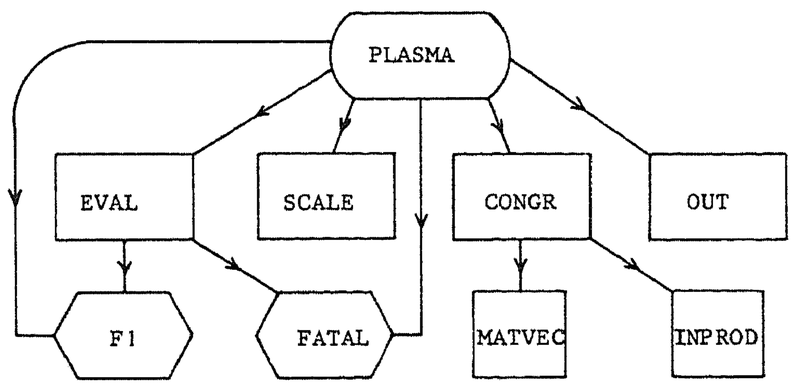


Fig. 5. Hierarchical structure of subprograms.

constant factor. This factor is computed from the value of

$$Q = \int_0^1 f_1(y)\,\mathrm{d}y \tag{19}$$

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) $Q, \beta, \gamma, \mathrm{Ha}, x_0$ in FORMAT (F12.4),
2) M in FORMAT (I4),
3) $x_1, \ldots, x_M$ in FORMAT (F12.4),
4) N, in FORMAT (I4),
5) $y_1, \ldots, y_N$ in FORMAT (F12.4),
6) ND in FORMAT (I4),
7) $yd_1, \ldots, yd_{ND}$ in FORMAT (F12.4),

where Q, β, γ, Ha 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 of $f_1(y)$, 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

```
-0.0146                    1033
13.3                       1034
 0.25                      1035
23.0                       1036
 0.0625                    1037
15                         1038
                           1039
 0.0                       1040
 0.02                      1041
 0.04                      1042
 0.0625                    1043
 0.08                      1044
 0.10                      1045
 0.15                      1046
 0.20                      1047
 0.25                      1048
 0.375                     1049
 0.500                     1050
 0.625                     1051
 0.750                     1052
 0.875                     1053
 1.000
15                                    1054
                           0.0        1055
                           0.10       1056
                           0.15       1057
                           0.20       1058
                           0.25       1059
                           0.28       1060
                           0.3125     1061
                           0.34       1062
                           0.42       1063
                           0.5        1064
                           0.6        1065
                           0.7        1066
                           0.8        1067
                           0.9        1068
                           1.000      1069
                           2          1070
                           0.25       1071
                           0.3125     1072
```

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 *yd* should contain at least one point, even if $f_1(y)$ 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 $f_1(y)$ and $f_2(x)$, the user is advised to refine the grid. This especially holds for x_0, since there the jump of $\partial\phi/\partial n$ is generally larger than near the discontinuity points of $f_1(y)$. 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/

GAMSQ γ^2,
GAMSQB $\gamma^2(1+\beta^2)$,
HASQ $\mathrm{Ha}^2/(1+\beta^2)$,
NXO grid number of x_0: X(NXO) = x_0,
AA value of $(1+\beta^2)\int_0^1 f_1(y)\,\mathrm{d}y$,
XO x_0,
FAC scaling factor of $f_1(y)$.

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;
NY – locates the y-grid
  WORK(NY + L) – Lth y-point, L = 1, ..., N;
NFYD – locates the jumping points of $f_1(y)$
  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 + I) – approximation of v in $(x(I), y(J))$,
  WORK(NU + MN + (J – 1)*M + I) – approximation of ϕ in $(x(I), y(J))$;
NEL – three times the number of nonzero entries in the strict upperdiagonal part;
NBOUND – 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

[1] M.S. van den Berg, Theory on a partially ionized gas centrifuge (FOM, Amsterdam, 1979).
[2] R. Courant and D. Hilbert, Methods of mathematical physics, vol. I (Interscience, New York, 1953).
[3] A.R. Mitchell and R. Wait, The finite element method in partial differential equations (Wiley, Chichester, 1977).
[4] Large sparse sets of linear equations, ed. J.K. Reid (Academic Press, London, 1971) p. 231.
[5] G. Strang and G.J. Fix, An analysis of the finite element method (Prentice Hall, Englewood Cliffs, NJ, 1973).

INTEGRAL VALUE -.0146
HALL PARAMETER 13.3000
GAMMA .2500
HARTMANN CONSTANT 23.0000
XO .0625
JUMPING POINTS OF F1(Y) .2500 .3125

V

X \ Y	0.0000	.1000	.1500	.2000	.2500	.2800	.3125	.3400	.4200	.5000	.6000	.7000	.8000	.9000	1.000
0.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
.0200	-.10	-.10	-.11	-.11	-.11	-.11	-.12	-.12	-.13	-.14	-.17	-.21	-.28	-.38	0.0
.0400	-.20	-.21	-.21	-.21	-.22	-.22	-.23	-.24	-.26	-.28	-.34	-.42	-.54	-.71	0.0
.0625	-.31	-.32	-.32	-.33	-.34	-.35	-.35	-.36	-.39	-.43	-.51	-.62	-.79	-1.00	0.0
.0800	-.40	-.40	-.41	-.42	-.43	-.44	-.45	-.46	-.49	-.54	-.63	-.76	-.94	-1.14	0.0
.1000	-.49	-.50	-.50	-.51	-.53	-.54	-.55	-.56	-.60	-.66	-.75	-.89	-1.08	-1.24	0.0
.1500	-.70	-.70	-.71	-.73	-.74	-.75	-.77	-.78	-.83	-.89	-1.00	-1.14	-1.30	-1.37	0.0
.2000	-.87	-.88	-.89	-.90	-.91	-.93	-.94	-.95	-1.00	-1.06	-1.16	-1.28	-1.41	-1.38	0.0
.2500	-1.00	-1.01	-1.02	-1.03	-1.05	-1.06	-1.07	-1.08	-1.12	-1.18	-1.26	-1.36	-1.44	-1.34	0.0
.3750	-1.19	-1.20	-1.20	-1.21	-1.22	-1.23	-1.24	-1.25	-1.27	-1.31	-1.36	-1.42	-1.44	-1.25	0.0
.5000	-1.20	-1.20	-1.21	-1.21	-1.22	-1.22	-1.22	-1.22	-1.24	-1.25	-1.28	-1.31	-1.30	-1.09	0.0
.6250	-1.06	-1.06	-1.07	-1.07	-1.07	-1.07	-1.07	-1.07	-1.07	-1.07	-1.08	-1.09	-1.06	-.87	0.0
.7500	-.81	-.82	-.82	-.82	-.82	-.82	-.81	-.81	-.79	-.79	-.78	-.78	-.75	-.61	0.0
.8750	-.46	-.47	-.47	-.48	-.47	-.47	-.46	-.45	-.44	-.42	-.41	-.41	-.39	-.31	0.0
1.0000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0

Φ

X \ Y	0.0000	.1000	.1500	.2000	.2500	.2800	.3125	.3400	.4200	.5000	.6000	.7000	.8000	.9000	1.000
0.0000	-20.55	-20.50	-20.43	-20.34	-20.21	-20.11	-20.00	-19.88	-19.47	-18.90	-17.86	-16.23	-13.52	-8.75	0.0
.0200	-20.57	-20.52	-20.45	-20.36	-20.23	-20.14	-20.02	-19.91	-19.51	-18.95	-17.95	-16.39	-13.83	-9.34	-.9
.0400	-20.61	-20.56	-20.50	-20.40	-20.28	-20.19	-10.08	-19.98	-19.60	-19.07	-18.15	-16.75	-14.52	-10.70	-3.3
.0625	-20.68	-20.63	-20.57	-20.49	-20.38	-20.30	-20.19	-20.10	-19.75	-19.29	-18.49	-17.34	-15.65	-13.09	-8.9
.0800	-20.75	-20.71	-20.66	-20.58	-20.47	-20.40	-20.31	-20.22	-19.90	-19.49	-18.80	-17.86	-16.59	-14.95	-13.3
.1000	-20.86	-20.82	-20.77	-20.70	-20.60	-20.54	-20.45	-20.38	-20.10	-19.75	-19.18	-18.45	-17.57	-16.65	-16.0
.1500	-21.19	-21.16	-21.13	-21.08	-21.01	-20.97	-20.91	-20.86	-20.69	-20.48	-20.17	-19.81	-19.45	-19.15	-19.0
.2000	-21.57	-21.56	-21.53	-21.50	-21.46	-21.44	-21.40	-21.37	-21.27	-21.15	-20.90	-20.82	-20.66	-20.55	-20.5
.2500	-21.98	-21.97	-21.95	-21.93	-21.91	-21.90	-21.88	-21.86	-21.81	-21.74	-21.66	-21.58	-21.52	-21.47	-21.4
.3750	-22.98	-22.98	-22.97	-22.97	-22.96	-22.96	-22.95	-22.95	-22.94	-22.92	-22.90	-22.88	-22.87	-22.86	-22.8
.5000	-23.83	-23.83	-23.83	-23.83	-23.83	-23.82	-23.82	-23.82	-23.82	-23.81	-23.81	-23.80	-23.80	-23.80	-23.8
.6250	-24.54	-24.54	-24.54	-24.54	-24.54	-24.54	-24.54	-24.53	-24.53	-24.53	-24.52	-24.52	-24.52	-24.52	-24.5
.7500	-25.14	-25.14	-25.14	-25.14	-25.13	-25.13	-25.13	-25.13	-25.12	-25.11	-25.10	-25.09	-25.08	-25.08	-25.0
.8750	-25.69	-25.69	-25.68	-25.67	-25.66	-25.65	-25.64	-25.64	-25.61	-25.58	-25.54	-25.52	-25.49	-25.48	-25.4
1.0000	-26.22	-26.25	-26.28	-26.33	-26.31	-26.24	-26.18	-26.12	-26.00	-25.90	-25.80	-25.73	-25.69	-25.66	-25.6

Computer Physics Communications 22 (1981) 439–450
North-Holland Publishing Company

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 Molecuulfysica, 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, of 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 Baeri 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 cm^2, 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 ΔR_p. Due to the laser-pulse, a melt front moves back from $x = a$ to $x = 0$ at a velocity v which is assumed to be constant. Thus the position of the front is given by $g(t) = a - vt$. 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

$$\partial C/\partial t = D\,\partial^2 C/\partial x^2 , \qquad 0 < x < g(t) , \tag{1.1}$$

with

$$g(t) = a - vt . \tag{1.2}$$

The absence of diffusion in the solid phase implies

$$\partial C/\partial t = 0 ; \qquad g(t) < x < a . \tag{1.3}$$

The boundary, initial and interface conditions are the following:

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

$$C(x, 0) = \exp(-[(x - R_p)/\Delta R_p]^2/2) ; \tag{1.4}$$

b) no mass transport across the vacuum–liquid interface is allowed:

$$\partial C/\partial x = 0 \qquad x = 0\,,\ t > 0\,; \tag{1.5}$$

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

$$\lim_{x \downarrow g(t)} C(x, t) = k \lim_{x \uparrow g(t)} C(x, t)\,, \qquad t > 0\,. \tag{1.6}$$

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

$$F(s) = \int_0^s C(x, t)\,\mathrm{d}x \Big/ \int_0^a C(x, t)\,\mathrm{d}x\,, \tag{1.7}$$

$$s \ll a\,, \qquad t \geqslant (a - s)/v\,, \tag{1.8}$$

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

$$s = 240\ \text{Å}\,, \qquad a = 3000\ \text{Å}\,, \qquad v = 4\ \text{m/s}\,, \tag{1.9}$$

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

$$\rho = \frac{x}{a}\,, \quad \alpha = \frac{D}{av}\,, \quad \tau = \frac{D}{a^2}t\,, \quad \mu = \frac{R_\mathrm{p}}{a}\,, \quad \sigma = \frac{\Delta R_\mathrm{p}}{a}\,. \tag{2.1}$$

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

$$\frac{\partial c}{\partial \tau} = \begin{cases} 0\,, & 1 - \alpha\tau < \rho \leqslant 1\,, \quad 0 \leqslant \tau \leqslant \alpha^{-1}\,, \\ \dfrac{\partial^2 c}{\partial \rho^2}\,, & 0 \leqslant \rho < 1 - \alpha\tau\,, \end{cases} \tag{2.2a}$$

with boundary conditions

$$\frac{\partial c}{\partial \rho} = 0\,, \qquad \rho = 0; \quad \lim_{\rho \downarrow 1-\alpha\tau} c = k \lim_{\rho \uparrow 1-\alpha\tau} c\,, \qquad \tau > 0\,, \tag{2.2b}$$

and initial conditions

$$c(\rho, \tau) = \exp(-\tfrac{1}{2}[(\rho - \mu)/\sigma]^2\,. \tag{2.2c}$$

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

$$\int_0^1 c(\rho, \tau)\,\mathrm{d}\rho = \int_0^1 c(\rho, 0)\,\mathrm{d}\rho\,, \qquad \tau \geqslant 0\,, \tag{2.3}$$

which can physically be interpreted as the conservation of mass during the resolidification process. This means that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\int_0^1 c(\rho,\tau)\,\mathrm{d}\rho = 0\,, \qquad \tau \geqslant 0\,. \tag{2.4}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left[\int_0^{g(\tau)} c(\rho,\tau)\,\mathrm{d}\rho + \int_{g(\tau)}^1 c(\rho,\tau)\,\mathrm{d}\rho\right]$$

$$= \frac{\mathrm{d}g}{\mathrm{d}\tau}\lim_{\rho\uparrow g(\tau)} c(\rho,\tau) + \int_0^{g(\tau)} \frac{\partial c}{\partial\tau}(\rho,\tau)\,\mathrm{d}\rho - \frac{\mathrm{d}g}{\mathrm{d}\tau}\lim_{\rho\downarrow g(\tau)} c(\rho,\tau) + \int_{g(\tau)}^1 \frac{\partial c}{\partial\tau}(\rho,\tau)\,\mathrm{d}\rho$$

$$= \frac{\mathrm{d}g}{\mathrm{d}\tau}(1-k)\lim_{\rho\uparrow g(\tau)} c + \int_0^{g(\tau)} \frac{\partial^2 c}{\partial\rho^2}(\rho,\tau)\,\mathrm{d}\rho = \frac{\mathrm{d}g}{\mathrm{d}\tau}(1-k)\lim_{\rho\uparrow g(\tau)} c + \lim_{\rho\uparrow g(\tau)}\frac{\partial c}{\partial\rho} = 0\,. \tag{2.5}$$

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

$$\lim_{\rho\uparrow 1-\alpha\tau}\frac{\partial c}{\partial\rho} = \alpha(1-k)\lim_{\rho\uparrow 1-\alpha\tau} c\,, \tag{2.2d}$$

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 τ, the interface is situated in the interval $[\rho_M, \rho_{M+1}]$ with

$$\rho_i = h(i-1), \qquad i = 1, \ldots, N+1 \qquad h = 1/N\,, \qquad M = [1 + N(1-\alpha\tau)]\,. \tag{2.6}$$

Let $c(\rho_i, \tau) \approx c_i(\tau), i = 1, \ldots, N+1$.
For $i = 1, \ldots, M-1$, we use the standard finite difference approximation [2] for (2.2a)

$$\frac{\mathrm{d}c_i}{\mathrm{d}\tau} = \frac{c_{i-1} - 2c_i + c_{i+1}}{h^2}\,, \quad i = 2, \ldots, M-1\,, \qquad \frac{\mathrm{d}c_1}{\mathrm{d}\tau} = \frac{2}{h^2}(c_2 - c_1)\,. \tag{2.7a}$$

For $i > M$, it is also simple:

$$\mathrm{d}c_i/\mathrm{d}\tau = 0\,, \quad i > M\,. \tag{2.7b}$$

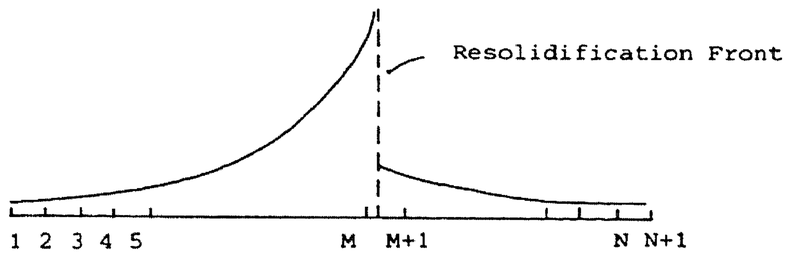


Fig. 1. Graph of $c(\rho, \tau)$.

For $i = M$, approximation of $dc_M/d\tau$ is more problematic, since c is discontinuous on $[\rho_M, \rho_{M+1}]$, see fig. 2.
In order to get a suitable approximation of $\partial^2 c/\partial\rho^2$ at $\rho = \rho_M$, we put

$$\delta = 1 - \alpha\tau - \rho_M\,, \qquad \frac{\partial^2 c}{\partial\rho^2}(\rho_M, \tau) \simeq \omega_1\, c(\rho_{M-1}, \tau) + \omega_2 c(\rho_M, \tau) + \omega_3 c(\rho_M + \delta, \tau)\,, \tag{2.7c}$$

where ω_1, ω_2 and ω_3 are chosen in such a way that (2.7c) has maximum accuracy, hence we find

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

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

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

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

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

Elaboration of (2.9) yields

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

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

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

Numerical experiments have shown that (2.10) should be applied if $\beta < 4/9$ and (2.11) otherwise.
Summarizing, we found for $dc_M/d\tau$ the formula

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

where c_{int} is defined by (2.10) or (2.11).
Remark. In earlier versions of this program, c_{int} was approximated by linear interpolation between c_M and c_{M+1}/k.

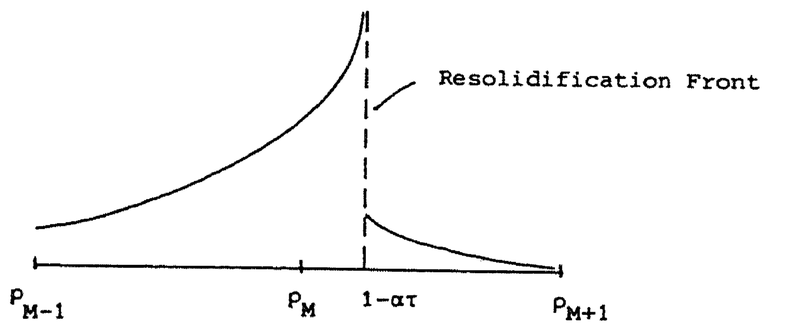


Fig. 2. Graph of $c(\rho_M, \tau)$.

This approach was rather unsatisfactory, since the behaviour of $c(\rho, \tau)$ sometimes differs left and right of the interface (monotonicly increasing to the left and monotonicly decreasing to the right, see fig. 2) which resulted in a systematical underestimation of c_{int}. Because of this underestimation, relation (2.3) refused to remain valid, especially as α was large and k was 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 α 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.

$$N > \alpha(1 - k)\,\beta/(1 - \beta^2)\,, \qquad 0 \leqslant \beta \leqslant \tfrac{1}{2}\,. \tag{2.12}$$

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

$$N \approx \alpha/5k\,. \tag{2.13}$$

2.4. Integration of semidiscretized problem

In this section, we discuss the organizational 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

$$M(\tau) = \text{entier}((1 - \alpha\tau) * N) + 1\,, \qquad 0 \leqslant \tau < \alpha^{-1}\,. \tag{2.14}$$

If we treat (2.7) as a system of fixed length by putting $\mathrm{d}c_i/\mathrm{d}\tau = 0, i > M$ 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

$$\tau_i = (i - 1)/\alpha N = \Delta\tau(i - 1)\,, \qquad i = 1, \ldots, N\,.$$

If $\tau_i \leqslant \tau < \tau_{i+1}$, the de facto dimension of (2.7) is $N + 1 - i$. If τ passes τ_{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 τ_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 τ_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 τ, $\tau_i \leqslant \tau < \tau_{i+1}$

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

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 $\Delta\tau = (m\alpha N)^{-1}$, where $m = 1 + 2N/\alpha$; in formula

$$c_j(\rho, \tau + \Delta\tau) = c_j(\rho, \tau) + \Delta\tau \frac{\mathrm{d}}{\mathrm{d}\tau} c_j(\rho, \tau) \,. \tag{2.15}$$

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

$$Q(\tau) = \int_0^1 c(\rho, \tau)\, \mathrm{d}\rho \quad \int_0^1 c(\rho, 0)\, \mathrm{d}\rho \tag{2.16}$$

was computed. According to (2.3) Q should be 1, but as $\tau \to \alpha^{-1}$, Q took unacceptable values, especially if α was large and k 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 $C(x, t)$ 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 $\mathrm{d}\boldsymbol{c}/\mathrm{d}\tau$, given an input vector $\boldsymbol{c}$ at time τ according to (2.7). It is called by RKC and its auxiliary routines;

SPECTR — This function is assigned the value of the spectral radius of the Jacobian matrix $(\partial \dot{c}_i/\partial c_j)$. Knowledge of this value is essential to numerically stable time-integrations [5];

UINIT — A function subprogram to be implemented by the user, which gives $C(x, 0)$ its value for given x.

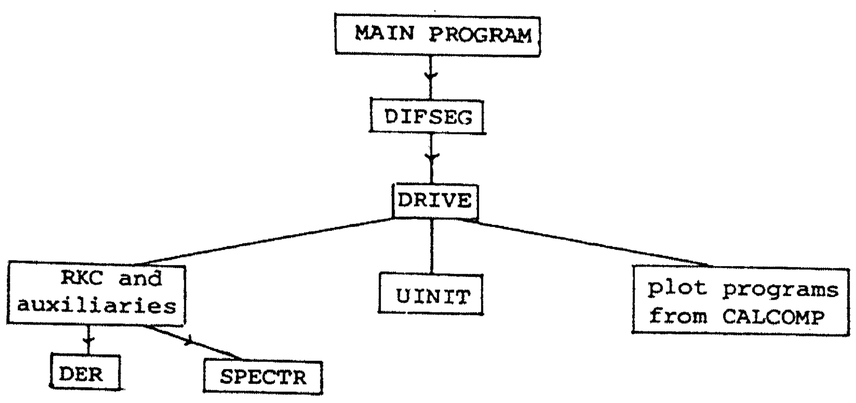


Fig. 3. Organization scheme of the program.

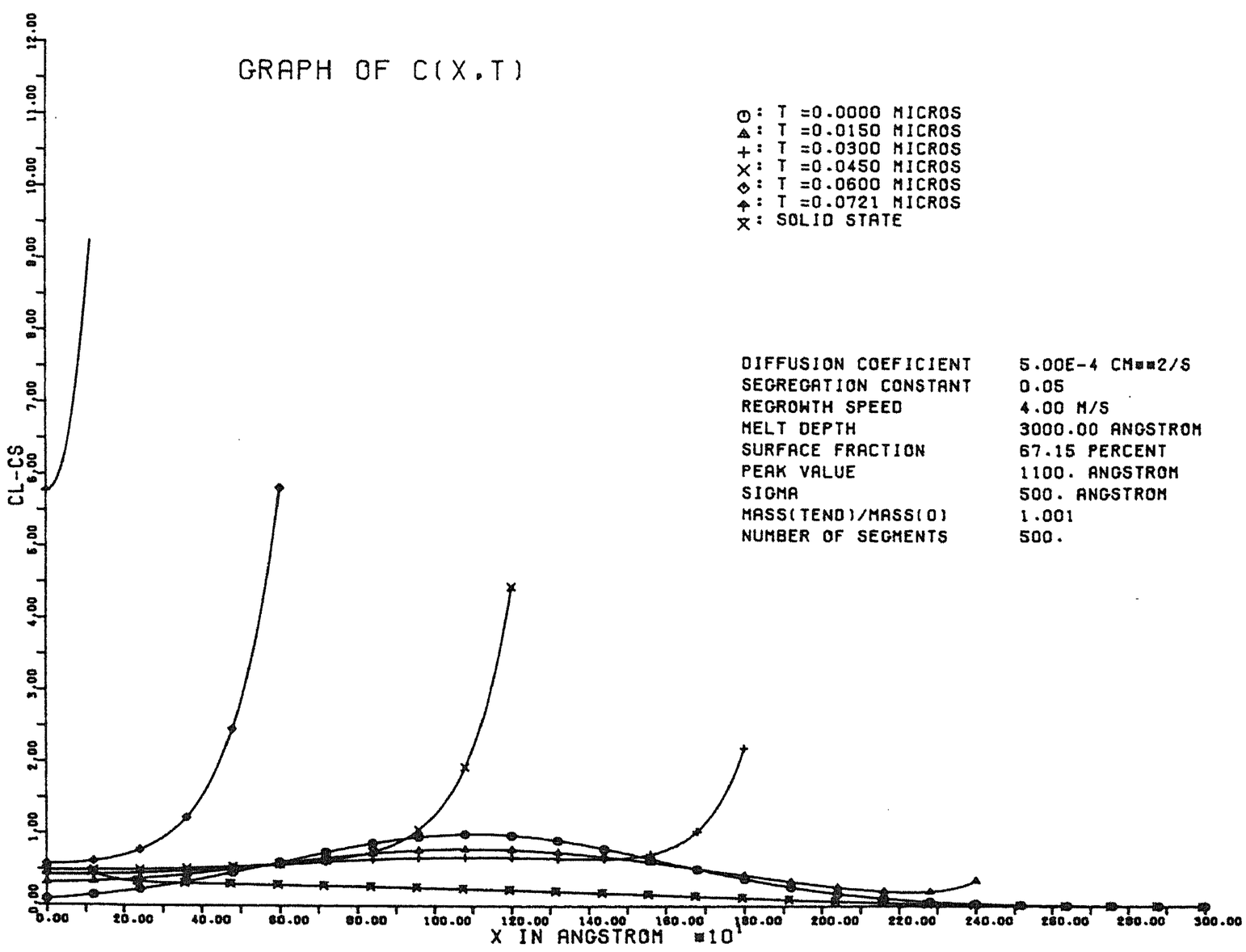


Fig. 4. Graph of $C(x, t)$.

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.

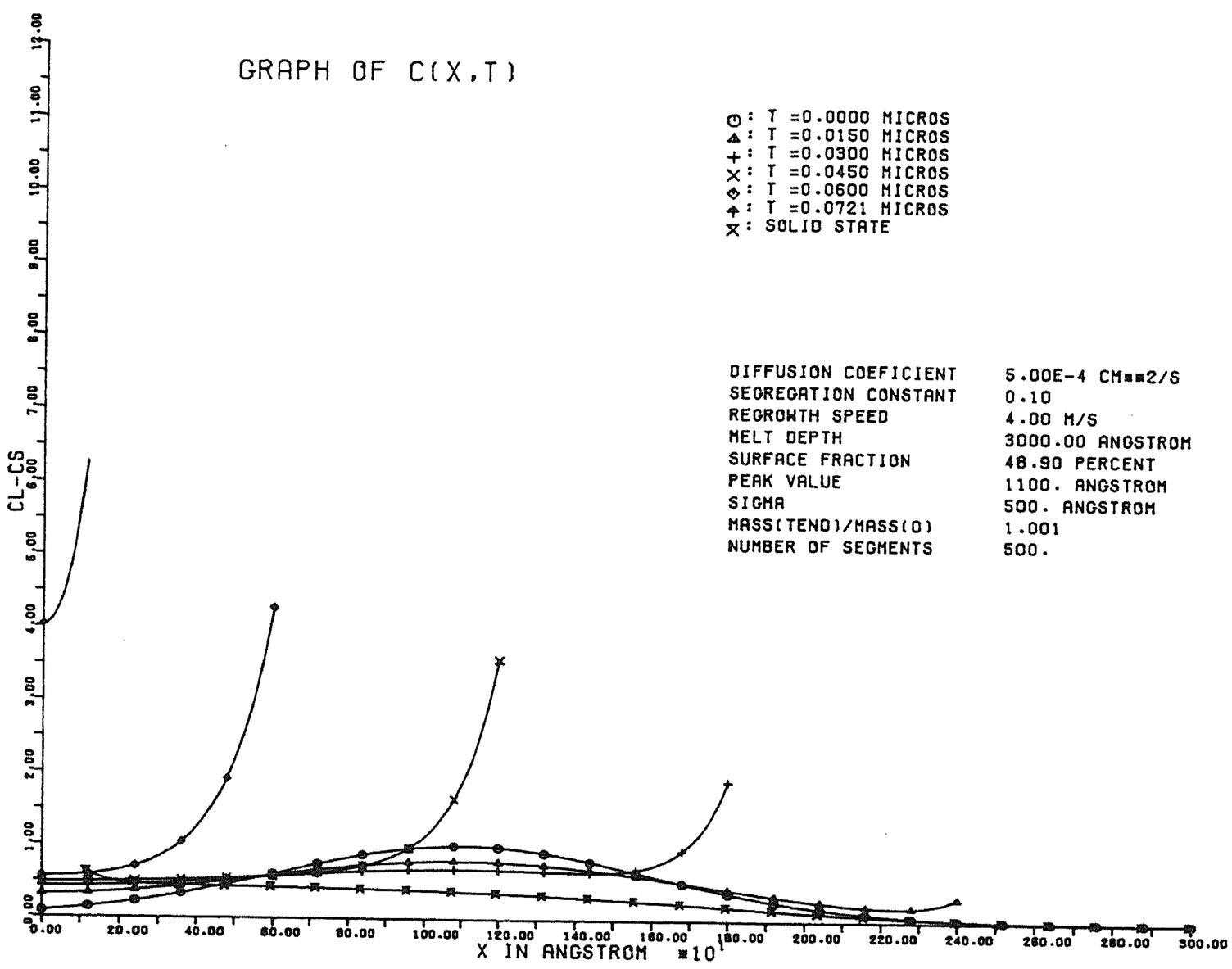


Fig. 5. Graph of $C(x, t)$.

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.

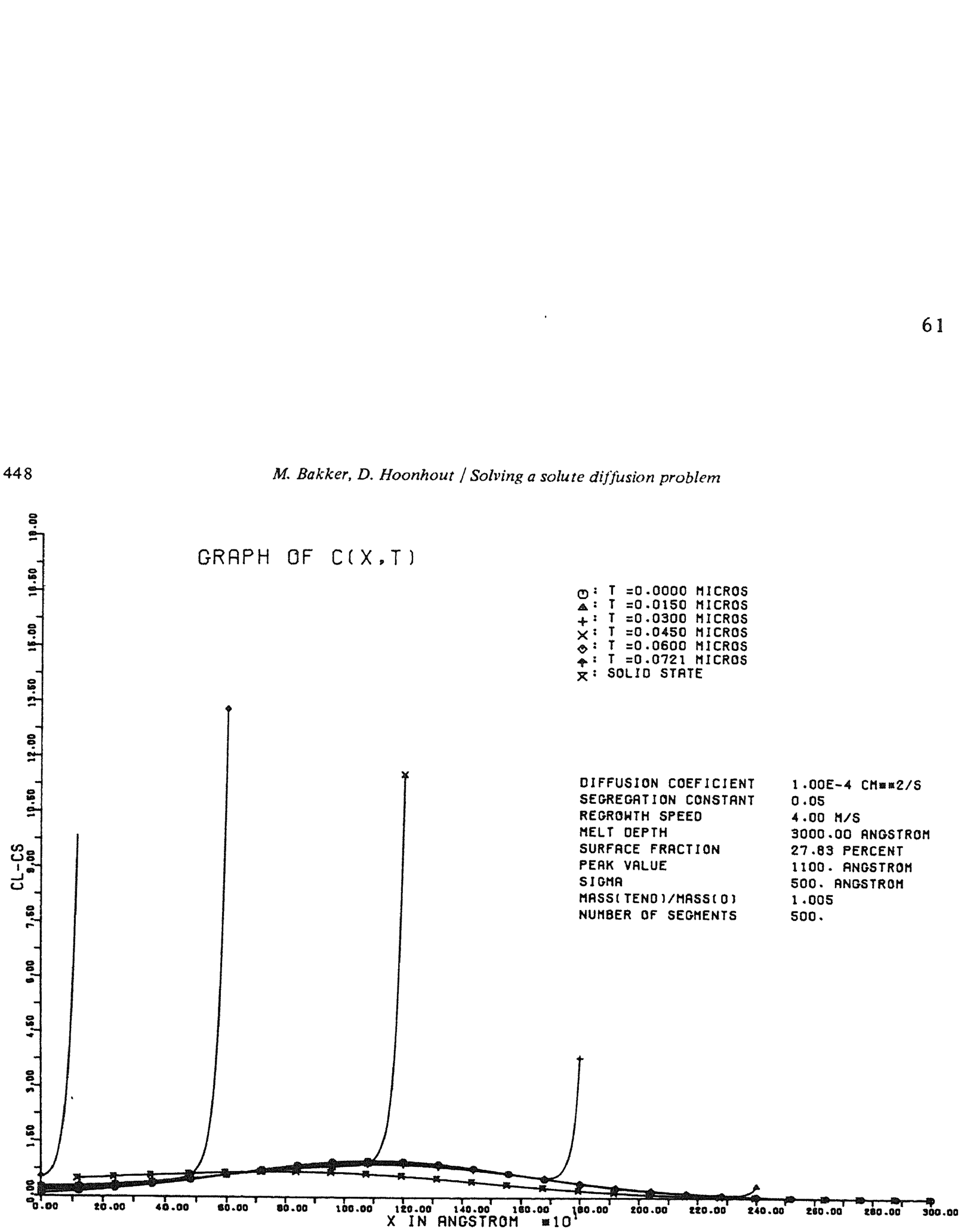


Fig. 6. Graph of $C(x, t)$.

Table 1
Input records for test examples

s	240 Å	PLOT?	yes	
D	1×10^{-4} cm^2/s 5×10^{-4} cm^2/s	k	0.05	0.1
a	3000 Å	R_p	1100 Å	
v	4 m/s	ΔR_p	500 Å	
N	500			

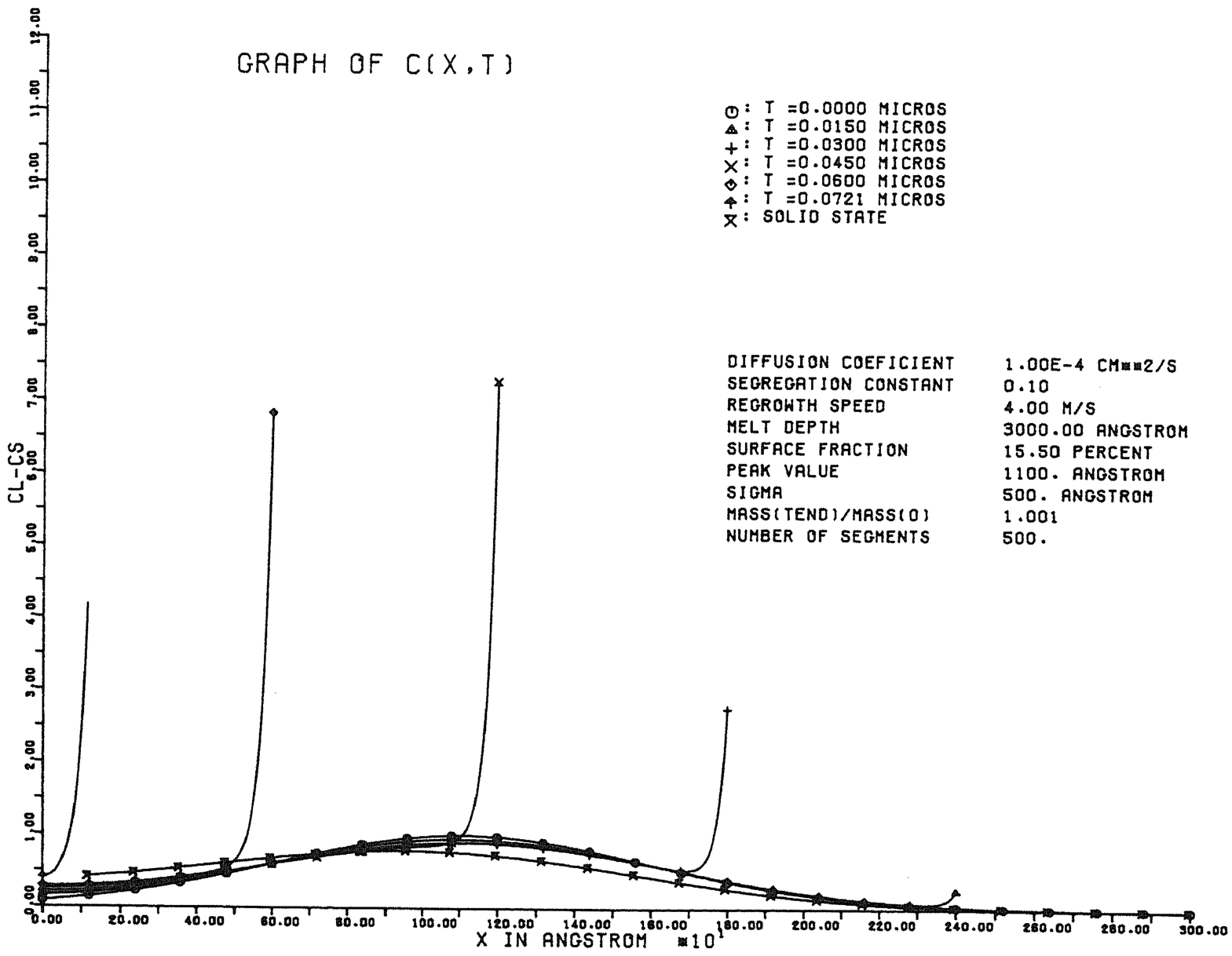


Fig. 7. Graph of $C(x, t)$.

Acknowledgements

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

References

[1] P. Baeri, S.U. Campisano, G. Foti and E. Rimini, Phys. Rev. Lett. 41 (1978) 2346.
[2] G.E. Forsythe and W.R. Wasow, Finite difference methods for partial differential equations (Wiley, New York, 1960).
[3] D. Hoonhout, Y. Tamminga, R. Garrett and F.W. Saris, Proc. Conf. on Laser effects in ion-implanted semiconductors, University of Catania (Italy) (1978).
[4] D. Hoonhout and F.W. Saris, J. Appl. Phys., submitted.
[5] B.P. Sommeijer and P.J. van der Houwen, Z. Angew. Math. Mech., to appear.
[6] A survey of numerical analysis, ed. J. Todd (McGraw-Hill, New York, 1962).
[7] C.W. White, S.R. Wilson, B.R. Appleton and F.W. Yound, J. Appl. Phys. 51 (1980) 738.

TEST RUN OUTPUT

```
SEGREGATION CONSTANT                                    .05
DIFFUSION COEFFICIENT IN CM**2/SEC                       .50E-03
SURFACE DEPTH IN ANGSTROM                               240.
MELTING DEPTH IN ANGSTROM                              3000.
NUMBER OF SEGMENTS IN WHICH [0,1] IS DIVIDED            500
SPEED OF RESOLIDIFICATION FRONT IN M/SEC               4.00
MEAN VALUE IN ANGSTROM                                 1100.
SIGMA IN ANGSTROM                                       500.
SURFACE FRACTION                                         .6715
MASS(T)/MASS(0)                                         1.0012



SEGREGATION CONSTANT                                    .10
DIFFUSION COEFFICIENT IN CM**2/SEC                       .50E-03
SURFACE DEPTH IN ANGSTROM                               240.
MELTING DEPTH IN ANGSTROM                              3000.
NUMBER OF SEGMENTS IN WHICH [0,1] IS DIVIDED            500
SPEED OF RESOLIDIFICATION FRONT IN M/SEC               4.00
MEAN VALUE IN ANGSTROM                                 1100.
SIGMA IN ANGSTROM                                       500.
SURFACE FRACTION                                         .4890
MASS(T)/MASS(0)                                         1.0010



SEGREGATION CONSTANT                                    .05
DIFFUSION COEFFICIENT IN CM**2/SEC                       .10E-03
SURFACE DEPTH IN ANGSTROM                               240.
MELTING DEPTH IN ANGSTROM                              3000.
NUMBER OF SEGMENTS IN WHICH [0,1] IS DIVIDED            500
SPEED OF RESOLIDIFICATION FRONT IN M/SEC               4.00
MEAN VALUE IN ANGSTROM                                 1100.
SIGMA IN ANGSTROM                                       500.
SURFACE FRACTION                                         .2783
MASS(T)/MASS(0)                                         1.0049



SEGREGATION CONSTANT                                    .10
DIFFUSION COEFFICIENT IN CM**2/SEC                       .10E-03
SURFACE DEPTH IN ANGSTROM                               240.
MELTING DEPTH IN ANGSTROM                              3000.
NUMBER OF SEGMENTS IN WHICH [0,1] IS DIVIDED            500
SPEED OF RESOLIDIFICATION FRONT IN M/SEC               4.00
MEAN VALUE IN ANGSTROM                                 1100.
SIGMA IN ANGSTROM                                       500.
SURFACE FRACTION                                         .1550
MASS(T)/MASS(0)                                         1.0009
```

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 funktie wordt opgedeeld in kleine partjes, ook wel genaamd segmenten; voorbeelden zijn intervallen, driehoeken, viervlakken, etc.;
- op ieder afzonderlijk segment wordt de funktie 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 algebraïsche vergelijkingen of (impliciete) stelsels gewone differentiaalvergelijkingen, waarvan de bijbehorende matrices *ijle* 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 extra grote 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 cylindersymetrisch 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 gesemidiskretiseerd door middel van de EEM. De hieruit ontstane gewone differentiaalvergelijking wordt met een aangepaste methode van Runge-Kutta opgelost.

MC NR

35220