Chemical Engineering Science, 1969, Vol 24, pp 1337–1349 Pergamon Press Printed in Great Britain

Solution of parabolic partial differential equations by a double collocation method

JOHN VILLADSEN and JAN P SØRENSEN
Institutet for Kemiteknik, Danmarks tekniske højskole, Danmark

(*Received* 28 *January* 1969)

Abstract—A method for solution of parabolic PDE by means of interpolation of the differential operators in two dimensions is described The method is developed on the basis of previously published methods for solution of ordinary differential equations by orthogonal collocation It is shown to be highly economical and very stable in comparison with the conventional Crank–Nicolson or with explicit methods such as the Runge–Kutta 4th order method The linear heat equation is used to illustrate the principle of the method and to discuss its convergence properties

THE MOST commonly used method for the numerical solution of parabolic partial differential equations is the finite difference method The explicit methods, which are the simplest from a computational standpoint are usually avoided, due to instability problems, that make it necessary to use very small increments in the direction of the "time" variable t The two-level implicit methods approximate the second order spatial derivative θ_{xx} by a weighted average between column j and column $(j+1)$ in the difference scheme If the weight factors of column j and column $(j+1)$ are respectively $1-\lambda$ and λ the following approximation results

$$\theta_{xx}|_{i,j} = \frac{\lambda}{(\Delta x)^2}[\theta_{i-1,j+1} - 2\theta_{i,j+1} + \theta_{i+1,j+1}] + \frac{1-\lambda}{(\Delta x)^2}[\theta_{i-1,j} - 2\theta_{i,j} + \theta_{i+1,j}] \qquad (1)$$

$$t = j\Delta t \text{ and } x = i\Delta x = 1,2 \quad .M.$$

The stability of the method increases with increasing value of $\lambda \epsilon [0, 1]$ and for $\lambda \geq \frac{1}{2}$ it is stable for any value of the mesh ratio $\Delta x/\Delta t$ as shown by Young and Frank[1] In the popular Crank–Nicolson method λ is equal to $\frac{1}{2}$ and this method is thus stable even for large time increments, while the explicit method follows for $\lambda = 0$ and demands $M = \Delta t/(\Delta x)^2 < \frac{1}{2}$

Special examples such as the linear heat equation have been solved by several of these methods and their truncation error, stability and convergence are discussed by Richtmyer[2]

For linear partial differential equations the solution of the M algebraic difference equations (1) and the supplementary equations at the boundaries $x = 0$ and $x = (M+1)\Delta x$ is accomplished either by a matrix method or by an explicit method described by Ames[3, pp 338–343] in his survey of implicit second order procedures Non linear equations with linear derivatives are solved by a quasi-linearization procedure One may linearize by the Picard method, or one might use a Newton–Raphson technique to get quadratic convergence in the iteration procedure A variant of this last method has recently been described by Lee[4] for the solution of a reactor design problem

A second approach to the solution of partial differential equations is to represent the dependent variable θ by a finite sum of trial functions Φ

$$\theta \sim \theta^{(m)} = \sum_{i=1}^{m} a_i^{(m)}\Phi_i + \Phi_0 \qquad (2)$$

For a problem of dimension P, the $a_i^{(m)}$ are constants or functions of 1, 2 $P-1$ of the independent variables depending on the number of independent variables, that are included in the Φ_i For parabolic equations the usual choice is to let the $a_i^{(m)}$ be functions of t and the Φ_i be functions of the space variables. In the Galerkin method, which is one of the most well known of

these methods Φ_0 satisfies the boundary conditions on x, while Φ_1, Φ_2... satisfy homogeneous boundary conditions for all $t > 0$, but other choices of Φ_0 are also permissible.

The $a_i^{(m)}$ are determined in one of several ways

Schechter[5] surveys variational methods in which the $a_i^{(m)}$ are determined such that a certain functional derived from the original problem is rendered stationary Another approach is to make a weighted average of the residual of the differential equation vanish. These so called weighted residual methods, which include the Galerkin method, are compared in an important paper by Finlayson and Scriven[6] An essential feature of any of these global methods is that the function θ at each t is represented by an m'th degree approximation in the bounded variable instead of the M interconnected parabolic arcs as in the difference method (1) The convergence of these methods for $m \to \infty$ is difficult to assert, especially since the time dependence of the coefficients $a_i^{(m)}$ must be found by a subsequent numerical integration in the t direction. Green[15] proves the existence, uniqueness and convergence of the method of moments for solution of the linear heat equation, and Bethel[16] treats transient evaporation through a finite region by the Galerkin method and makes some comments as to the convergence of the approximation Both authors concede, that rigorous proofs of convergence can only be obtained for extremely simple problems, but many test examples show, that global methods are indeed superior to finite difference methods with respect to speed of convergence. The principles of the methods have been known for a long time in engineering literature[7, 8] and the last decade has seen an increasing number of applications especially in the field of transport phenomena. The main objection to their use is that they are very unwieldy for non-linear problems and rather complicated to program for automatic computation

The so called collocation method in which the residual $L^V(\theta)$ with $\theta^{(m)}$ inserted for θ is equated to zero at m predetermined points in the domain of θ and the $a_i^{(m)}$ determined from the resulting ordinary differential equations is much easier to use, especially for non-linear problems. For an arbitrary positioning of the collocation points the method may be divergent even for rather simple ordinary differential equations[9] and for partial differential equations this situation probably occurs even more frequently Interpolation of θ for fixed t by a polynomial of degree m is thus seen to be very questionable while interpolation by interconnected parabolic arcs is a slow, but more reliable method. This is probably the reason, why very little use has been made of the collocation method even though it could be implemented in many cases where the computational difficulties of other global methods are prohibital.

A series of publications by Stewart and Villadsen[10–12] has suggested, that a positioning of the collocation abscissae at the zeros of orthogonal polynomials leads to a rapidly convergent interpolation scheme even for functions, that are poorly represented by polynomials The method of interior orthogonal collocation has been shown[10] to be identical to Galerkins interior method if the residual $L^V(\theta^{(m)})$ is a polynomial of degree $d \leq m$ in x and similar analogies to other weighted residual methods can also be established. Based on a large number of experiments, which include non linear examples, it appears, that the convergence properties of the method are very similar to variational and weighted residual methods, and the computational difficulties of these methods are consequently circumvented without any loss of accuracy

Several examples given in [10] show the positioning of the collocation points in the domain of the space variables. One example is the transient diffusion with chemical reaction in flat catalyst plates:

$$\frac{\partial\theta}{\partial t} = \frac{\partial^2\theta}{\partial x^2} - \Lambda^2\theta \tag{3}$$

B C.1 and 2

$$\theta = 1 \text{ at } x = \pm 1 \text{ for } t > 0$$

I.C

$$\theta = 0 \text{ for } -1 < x < 1 \text{ at } t = 0$$

Λ is a Thiele modulus defined in [10].

The symmetrical function θ is approximated by

$$\theta \sim \theta^{(m)} = \theta(1) + (1-x^2) \sum_{i=0}^{m-1} a_i^{(m)} P_i(x^2) \quad (4)$$

where $P_i(x^2)$ are orthogonal polynomials of even degree which (except for a scale factor) are defined by

$$\int_0^1 (1-x^2) P_i(x^2) P_j(x^2)\, dx = 0 \text{ for } i \neq j$$

The following set of m simultaneous ordinary differential equations results after collocation at the m zeros of $P_m(x^2)$ and replacing the value of the spatial derivatives at the interior collocation points by $(m+1)$ point interpolation formulae in the interior collocation point ordinates and the ordinate $\theta_{m+1} = 1$ at $x = x_{m+1} = 1$

$$\frac{d\theta^{(m)}(x_i)}{dt} = \sum_{j=1}^{m+1} B_{ij}^{(m)} \theta^{(m)}(x_j) - \Lambda^2 \theta^{(m)}(x_i) \quad (5)$$

The coefficient $B_{ij}^{(m)}$ is the value of the Lagrangian interpolation coefficient for the ordinate j in the approximation of the Laplacian operator at the i'th collocation point Algorithms for computation of the $B_{ij}^{(m)}$ are found in [10].

The final result of an application of a global approximation principle in the space domain is a set of simultaneous differential equations like (5) which either express a set of ordinates or a set of expansion coefficients $a_i^{(m)}$ as time functions For nonlinear partial differential equations the equations are different for different methods such as Galerkins method and orthogonal collocation, but their solutions $\theta^{(m)}(t, x_i)$ or $a_i^{(m)}(t)$ would probably be almost the same for large m due to the close relationship of the methods. The difference between the methods lies entirely in the ease with which (5) is constructed by the method of orthogonal collocation.

In the integration of Eqs. (5) an explicit method either of the Runge–Kutta or the predictor corrector type seems to be used by all investigators. Bethel[16] used an Adams Moulton predictor corrector scheme, while Marcussen[13] in a study of non linear adsorption kinetics integrated the collocation equations by a 4'th order variable step Runge–Kutta technique Balslev[14] solved the set of five parabolic partial equations for the oxidation of naphtalene to phthalic anhydride by collocation in the radial direction of the catalyst tube followed by a Runge–Kutta integration of the ensuing ordinary differential equations.

The Runge–Kutta and predictor corrector algorithms are basically explicit methods, and although they are much less prone to instability than the simple explicit scheme of (1) with $\lambda = 0$, (which corresponds to the first order Runge–Kutta method–Eulers method) they might not be the optimal methods with respect to computational efficiency. This is especially true for large sets of simultaneous equations, if some components of $\theta(x_i)$ are increasing while others are decreasing, such as is the case in reactor design problems with several independent reactants. With three collocation points in the radial direction and a boundary condition of "the radiation type" Balslev[14] solved $4 \cdot 5 = 20$ simultaneous ordinary differential equations for the four independent reactant concentrations and the temperature at five radial positions. For some combinations of entrance temperature and composition a very stringent control of step size in the axial direction was necessary in order to avoid a blow up of the solution to the reactor model (which could be mistaken for an unstable physical situation), and this naturally caused a high computing expenditure

The best way to avoid these instability problems, which are further discussed in Appendix B, is to use an implicit integration method in the t-direction One might use a two-level implicit formula as (1). In this case the collocation formula for the x-direction should be used at $t = j\Delta t$ and at $t = (j+1)\Delta t$ to compute the $\theta^{(m)}(x_i)$ at $t = (j+1)\Delta t$ simultaneously from

$\theta(x_i)$ at $t = j\Delta t$ The Crank–Nicolson method is however only a special case of a general n'th order approximation of the differential equation in the time direction by which $\theta(x_i)$ is found simultaneously at $t+\Delta\tau$ and at several intermediate t-levels between t and $t+\Delta\tau$, where $\Delta\tau$ is a much larger time step than Δt This procedure implies a collocation procedure not only in the space variables, but also in the time variable. An n point equidistant collocation in the t-direction would often fail to converge, but a collocation at the zeros of an orthogonal polynomial within each $\Delta\tau$ might presumably give a highly stable combination of an implicit integration within $\Delta\tau$ and an explicit forward integration from one time step to the next

In [9] this technique was used to integrate large sets of first order equations with given initial conditions The choice of Legendre polynomials as perturbation functions in the t-direction was shown to be optimal in the sense that the vector $\boldsymbol{\theta}(\mathbf{x})$ at $t = t+\Delta\tau$, where in this case x_i is interpreted as the i'th component of the dependent variable $\boldsymbol{\theta}$, was very accurately determined even though large errors occured in the components of $\boldsymbol{\theta}$ at the intermediate time levels $t = [t_1, t_2 \ldots\ldots, t_n] \epsilon [t, t+\Delta\tau]$ This is naturally a very desirable property of the method since it implies, that the inherited error in the t direction will be very small. Any approximation of $\boldsymbol{\theta}$ in $[t, t+\Delta\tau]$ by an orthogonal family is preferable to the Taylor series representation implied by a conventional forward integration technique, in the sense that some norm of $\boldsymbol{\theta}$ is minimized. The use of Legendre polynomials as perturbation functions has the unique advantage, that the only important value, namely $\theta(x_i)$ at $t+\Delta\tau$, is well determined The self-stabilizing property of the method discussed in Appendix B allows large time steps $\Delta\tau$ to be used and the saving in computer time as compared to a standard Runge–Kutta routine is of the order of a factor 30–50

Whether these features are retained, when an expansion in Legendre polynomials in the t-direction is combined with an expansion in some other polynomial family in the space domain can only be "proved" experimentally by solving a number of problems, whose exact solution is already known. Such arguments are far from ideal, but since realistic error bounds and convergence theorems for realistic problems are unattainable at the present, this seems to be the only approach

The following simple example of the heat equation serves to illustrate how the method is used, and an error analysis shows, that the above mentioned small end point error after each time step is carried over from the solution of simultaneous ordinary differential equations to the solution of parabolic partial differential equations The design of a catalytic reactor for oxidation of o-xylene to phthalic anhydride will be treated in a following paper in order to demonstrate that even very complicated systems of coupled partial differential equations can be treated by the method.

The heat equation for plane-parallel symmetry with constant temperature at $x = \pm 1$

$$\frac{\partial\theta}{\partial t} = \frac{\partial^2\theta}{\partial x^2} \tag{6}$$

B C.1 and 2

$$\theta = 0 \text{ at } x = \pm 1 \text{ for all } t > 0$$

I.C.

$$\theta = 1 \text{ for } -1 < x < 1 \text{ at } t = 0$$

The exact solution to (6) is [17, p. 45]

$$\theta(x,t) = \frac{4}{\pi}\sum_{p=0}^{\infty}\frac{(-1)^p}{2p+1}\cos\left(\frac{(2p+1)\pi x}{2}\right) \exp(-(2p+1)^2\pi^2 t/4). \tag{7}$$

The average value of θ is

$$\theta_{av}(t) = \int_0^1 \theta(x,t)\,dx = \frac{8}{\pi^2}\sum_{p=0}^{\infty}\frac{1}{(2p+1)^2} \exp(-(2p+1)^2\pi^2 t/4)$$

Take two interior collocation points in the

x-direction ($M=2$) and three points in the t-direction ($N=2$). The value of θ at the mesh points of the first time step $\Delta\tau_1$ are read from the following matrix

$\rightarrow j$				
x_1	x_2	$x_3=x_{M+1}=1$		
θ_{11}	θ_{12}	θ_{13}	$t_1=0$	$\downarrow$
θ_{21}	θ_{22}	θ_{23}	t_2	i
θ_{31}	θ_{32}	θ_{33}	t_3	
θ_{41}	θ_{42}	θ_{43}	$t_4=t_{N+2}=\Delta\tau_1$	

$x_1=0{\cdot}2852\ldots$ and $x_2=0{\cdot}7650\ldots$ are taken from Table 2 of [10].

$$t_2=\left(\tfrac{1}{2}-\frac{\sqrt{3}}{6}\right)\Delta\tau_1,\ t_3=\left(\tfrac{1}{2}+\frac{\sqrt{3}}{6}\right)\Delta\tau_1$$

are the zeros of the second degree shifted Legendre polynomial $Le_2(t)=6t^2-6t+1$

θ_{1j} are known from the initial condition and θ_{i3} are known from the boundary conditions The discrepancy between the value of θ_{13} as an initial value ($\theta_{13}=1$) and as a boundary value ($\theta_{13}=0$) is unimportant, since θ_{13} does not occur in the following calculations

Define the following matrices

$$\boldsymbol{\theta I}=\begin{pmatrix}\theta_{21} & \theta_{22}\\ \theta_{31} & \theta_{32}\\ \theta_{41} & \theta_{42}\end{pmatrix},\ \boldsymbol{\theta 1}=\begin{pmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\\ \theta_{31} & \theta_{32}\\ \theta_{41} & \theta_{42}\end{pmatrix},\ \boldsymbol{\theta 2}=\begin{pmatrix}\theta_{21} & \theta_{22} & \theta_{23}\\ \theta_{31} & \theta_{32} & \theta_{33}\\ \theta_{41} & \theta_{42} & \theta_{43}\end{pmatrix} \tag{8}$$

$\boldsymbol{\theta I}$ consists of unknown quantities only, while $\boldsymbol{\theta 1}$ and $\boldsymbol{\theta 2}$ are matrices used to set up the method

θ_{xx} is interpolated by a 2 M degree polynomial at each meshpoint of $\boldsymbol{\theta I}$

$$\left(\frac{\partial^2\theta}{\partial x^2}\right)_{ij}=B_{j1}\theta_{i1}+B_{j2}\theta_{i2}+B_{j3}\theta_{i3}$$

for $j=1$ and 2 and $i=2$, 3 and 4

The value of the coefficients B_{ji} are tabulated in [10] Table 2 for $n_t=M+1=3$ sample points in the x-direction and $j=1$ and 2 The total set of 6 equations is

$$\ddot{\boldsymbol{\theta}}\mathbf{I}^T=\mathbf{B}^{(2)}_{\text{int}}\boldsymbol{\theta 2}^T \text{ or } \ddot{\boldsymbol{\theta}}\mathbf{I}=\boldsymbol{\theta 2}(\mathbf{B}^{(2)}_{\text{int}})^T \tag{9}$$

$\mathbf{B}^{(2)}_{\text{int}}$ is a (2 3) matrix of the first 2 rows of $\mathbf{B}^{(2)}$ in [10]. $\ddot{\boldsymbol{\theta}}\mathbf{I}$ is a matrix of $\partial^2\theta/\partial x^2$ taken at the meshpoints of $\boldsymbol{\theta I}$

The time derivative is interpolated by a polynomial of degree $N+1$ at each meshpoint of $\boldsymbol{\theta I}$

$$\Delta\tau_1\left(\frac{\partial\theta}{\partial t}\right)_{ij}=A_{i1}\theta_{1j}+A_{i2}\theta_{2j}+A_{i3}\theta_{3j}+A_{i4}\theta_{4j}$$

for $i=2,3$ and 4 and $j=1$ and 2

The total set of 6 equations reads

$$\Delta\tau_1\dot{\boldsymbol{\theta}}\mathbf{I}=\mathbf{A}^{(2)}_{\text{t}}\boldsymbol{\theta 1} \tag{10}$$

$\dot{\boldsymbol{\theta}}\mathbf{I}$ is a matrix of $\dfrac{\partial\theta}{\partial t}$ taken at the meshpoints of $\boldsymbol{\theta I}$.

The coefficients A_{ij} are found as described in [10] Appendix C For convenience the complete matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ are shown in Appendix A Table A1 in the same way as $\mathbf{B}$ is listed in [10]. The last three rows of $\mathbf{A}^{(2)}$ define matrix $\mathbf{A}^{(2)}_{\text{int}}$. (9) and (10) are inserted into (6) and

$$\Delta\tau_1\,.\,\boldsymbol{\theta 2}\,.\,(\mathbf{B}^{(2)}_{\text{int}})^T=\mathbf{A}^{(2)}_{\text{int}}\,.\,\boldsymbol{\theta 1} \tag{11}$$

or

$$\Delta\tau_1\begin{pmatrix}\theta_{21} & \theta_{22} & \theta_{23}\\ \theta_{31} & \theta_{32} & \theta_{33}\\ \theta_{41} & \theta_{42} & \theta_{43}\end{pmatrix}\begin{pmatrix}B_{11} & B_{21}\\ B_{12} & B_{22}\\ B_{13} & B_{23}\end{pmatrix}=\begin{pmatrix}A_{21} & A_{22} & A_{23} & A_{24}\\ A_{31} & A_{32} & A_{33} & A_{34}\\ A_{41} & A_{42} & A_{43} & A_{44}\end{pmatrix}\begin{pmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\\ \theta_{31} & \theta_{32}\\ \theta_{41} & \theta_{42}\end{pmatrix} \tag{11a}$$

Equation (11a) is reformulated into the following set of linear algebraic equations

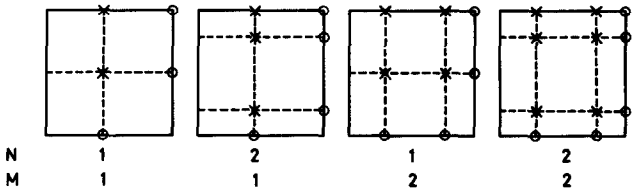

Fig 1 Meshpoints for the double collocation method with $N = 1$ and 2, $M = 1$ and 2 The time axis is vertical, the space axis horizontal Points marked ○ are known from initial– or boundary conditions

$$\begin{bmatrix} A_{22}-b_{11} & A_{23} & A_{24} & -b_{12} & 0 & 0 \\ A_{32} & A_{33}-b_{11} & A_{34} & 0 & -b_{12} & 0 \\ A_{42} & A_{43} & A_{44}-b_{11} & 0 & 0 & -b_{12} \\ -b_{21} & 0 & 0 & A_{22}-b_{22} & A_{23} & A_{24} \\ 0 & -b_{21} & 0 & A_{32} & A_{33}-b_{22} & A_{34} \\ 0 & 0 & -b_{21} & A_{42} & A_{43} & A_{44}-b_{22} \end{bmatrix} \cdot \begin{bmatrix} \theta_{21} \\ \theta_{31} \\ \theta_{41} \\ \theta_{22} \\ \theta_{32} \\ \theta_{42} \end{bmatrix} = \begin{bmatrix} \theta_{23} & b_{13} & -\theta_{11} \cdot A_{21} \\ \theta_{33} \cdot & b_{13} & -\theta_{11} \quad A_{31} \\ \theta_{43} \cdot & b_{13} & -\theta_{11} \quad A_{41} \\ \theta_{23} \cdot & b_{23} & -\theta_{12} \cdot A_{21} \\ \theta_{33} \cdot & b_{23} & -\theta_{12} \cdot A_{31} \\ \theta_{43} & b_{23} & -\theta_{12} \cdot A_{41} \end{bmatrix} \tag{11b}$$

where $b_{ij} = \Delta\tau \uparrow . B_{ij}$.
Equation (11b) can be written in condensed form

$$\mathbf{Q}_{\mathrm{NM}} \cdot \overline{\theta \mathbf{I}} = \bar{v}_{NM} \tag{12}$$

where $\overline{\theta \mathbf{I}}$ is a vector made up by joining together the columns of $\boldsymbol{\theta}\mathbf{I} \cdot \bar{v}_{NM}$ is the difference between two vectors $\bar{v}_1$ and $\bar{v}_2$, where $\bar{v}_1$ contains the boundary values θ_{i3} and v_2 the initial values θ_{1j} at the start of each timestep

Equations like (11b) can be set up by analogy for any M and N. Equation (11b) is a generally applicable interpolation formula for $[\nabla^2 - \partial/\partial t]$ and it might be used to construct discrete versions of parabolic differential equations just as The matrices **A** and **B** of[10] were used to discretizise ordinary differential equations For any parabolic differential equation with linear

$$\left[\nabla^2 - \frac{\partial}{\partial t}\right] \theta = F(x, t, \theta) \tag{13}$$

derivatives $\overline{\theta \mathbf{I}}$ can be computed by successive approximation as the solution to

$$\overline{\theta \mathbf{I}}_k = \mathbf{Q}_{\mathrm{NM}}^{-1} (\bar{v}_{NM} - \bar{F}(\bar{x}, \bar{t}, \overline{\theta \mathbf{I}}_{k-1})) \tag{14}$$

where $\overline{\theta \mathbf{I}}_k$ is the k'th approximation to $\overline{\theta \mathbf{I}}$ and $\overline{\theta \mathbf{I}}_0$ is equated to the accepted value of $\overline{\theta \mathbf{I}}$ at the end of the previous time interval A linearization of $\bar{F}$ from the values $\bar{F}_0$ at the end of the previous time interval might also be used to provide quadratic convergence of the iteration as described in[4]

The main point is, that once $\mathbf{Q}_{\mathrm{NM}}^{-1}$ is computed, it is the same all through the integration of the differential equation provided that $\Delta\tau_1 = \Delta\tau_2 \ldots$

It is noteworthy, that the use of a small N might be compensated by using smaller time steps $\Delta\tau$, while the choice of a certain M limits the accuracy of the whole integration process. Since most even functions occuring as solution to diffusion problems can be adequately represented by 4-degree polynomials in the space variables $M = 2$ is sufficient in a majority of practical applications

Figure 1 and Appendix A, Table A2 show the positioning of the meshpoints, the matrices $\mathbf{Q}_{\mathrm{NM}}$ and the right hand sides of (12) for $N = 1$ and 2 and $M = 1$ and 2 It is easily seen, how lower order or higher order approximations are constructed by deleting or adding certain rows or columns in (11b).

This completes the description of the numerical method A discussion of the expectable accuracy and speed of convergency is given in Appendix B for linear PDE

NOTATION

$a_i^{(m)}$ weight factor, defined under Eq (2)
$A_{ij}^{(m)}$ weight factor, element of matrix **A**
$\mathbf{A}_{\mathrm{int}}^{(m)}$ matrix, defined above Eq. (11)

b parameter, defined under Eq. (B10)
b_{ij} parameter, defined under Eq. (11b)
$B_{ij}^{(m)}$ weight factor, element of matrix $\mathbf{B}$
$\mathbf{B}_{\text{int}}^{(m)}$ matrix, defined under Eq (9)
$\mathbf{B}_{\text{intt}}^{(m)}$ matrix derived from $\mathbf{B}^{(m)}$ by deleting the last column
$L^v(\theta)$ differential-equation operator
$L^v(\theta^{(m)})$ differential-equation residual for approximate function $\theta^{(m)}$
M number of interior collocation points in the x-direction
n number of subintervals in $[0, 1]$ or $[0, \Delta\tau]$
N order of the Runge–Kutta formula or number of interior collocation points in the t-direction
$P_i(x^2)$ orthogonal polynomial of even degree, defined under Eq (4)
$\mathbf{Q}_{\text{NM}}$ matrix, defined under Eq (11b)
t time
$\bar{\nabla}_{NM}$ vector, defined under Eq. (12)
w_i weight factor, Eq. (B3)
x position co-ordinate
y dependent variable, defined in Eq. (B9)
ϵ relative error, defined by Eq (B2)
Φ trial function
λ weight factor, Eq. (1)
λ_i i-th eigenvalue
Λ Thiele modulus, Eq (3)
$\Delta\tau$ time step
$\theta^{(m)}$ m-th order approximation to θ
$\overline{\theta\mathbf{I}}$ matrix, defined in Eq. (8)
$\theta\mathbf{I}$ vector, defined under Eq (12)
$\theta\mathbf{1}$ matrix, defined in Eq (8)
$\theta\mathbf{2}$ matrix, defined in Eq (8)
$\theta\dot{\mathbf{I}}$ matrix, first order spatial derivative, defined under Eq. (10)
$\theta\ddot{\mathbf{I}}$ matrix, second order spatial derivative, defined under Eq. (9)

REFERENCES

[1] YOUNG D M and FRANK T G, *Int comput center Bull* 1963 **2**
[2] RICHTMYER R D, *Difference Methods for Initial Value Problems* Wiley 1957
[3] AMES W F, *Nonlinear Partial Differential Equations in Engineering* Academic Press 1965
[4] LEE E S, *Chem Engng Sci* 1966 **21** 143
[5] SCHECHTER R S, *The Variational Method in Engineering* McGraw-Hill 1967
[6] FINLAYSON B A and SCRIVEN L E, *Chem Engng Sci* 1965 **20** 395
[7] CRANDALL S H, *Engineering Analysis* McGraw-Hill 1956
[8] KANTOROVICH L V and KRYLOV V I, *Approximate Methods of Higher Analysis* Wiley 1958
[9] VILLADSEN J, *Nord DATA*, Vol 1, pp 138–175 Helsinki 1968
[10] VILLADSEN J and STEWART W E, *Chem Engng Sci* 1967 **22** 1483
[11] STEWART W E and VILLADSEN J, *A I Ch E Jl* 1969
[12] VILLADSEN J, *Selected Approximation Methods for Chemical Engineering Problems* Akademisk Forlag 1969
[13] MARCUSSEN L, Ph D Thesis, Instituttet for Kemiteknik, Danmarks tekniske Højskole 1968
[14] BALSLEV E, M Sc Thesis, Instituttet for Kemiteknik, Danmarks tekniske Højskole 1967
[15] GREEN J W, *J Res natn Bur Stand* 1953 **51** 2441
[16] BETHEL H E, *Int J Heat Mass Transfer* 1967 **10** 1509
[17] CRANK J, *Mathematics of Diffusion* Clarendon Press 1957
[18] FOX L, *Numerical Solution of Ordinary and Partial Differential Equations* Pergamon Press 1962

APPENDIX A

Tables of collocation constants for $\left(\nabla^2 - \frac{\partial}{\partial t}\right)$

Table A1 Collocation data for N interior collocation points in $[0, \Delta\tau]$

$N = 1$	$N = 2$
$A^{(1)} = \begin{pmatrix} -3 & 4 & -1 \\ -1 & 0 & 1 \\ 1 & -4 & 3 \end{pmatrix}$	$A^{(2)} = \begin{pmatrix} -7 & 3(a+1) & 3(1-a) & 1 \\ -a-1 & a & a & 1-a \\ a-1 & -a & -a & a+1 \\ -1 & 3(a-1) & -3(a+1) & 7 \end{pmatrix}$ $a = \sqrt{3}$
$t_i = \frac{1}{2}$ and 1	$t_i = \frac{1}{2}\left(1 - \frac{a}{3}\right), \frac{1}{2}\left(1 + \frac{a}{3}\right)$ and 1

Table A2 Interpolation matrices for $N = 1, 2$ and $M = 1, 2$

$N=1$ and $M=1$			$N=2$ and $M=1$			
$\mathbf{Q}_{11}$		$\bar{\nabla}_{11}$	$\mathbf{Q}_{21}$			$\bar{\nabla}_{21}$
$A_{22}-b_{11}$	A_{23}	$b_{12}\theta_{22}-A_{21}\theta_{11}$	$A_{22}-b_{11}$	A_{23}	A_{24}	$b_{12}\theta_{22}-A_{21}\theta_{11}$
A_{32}	$A_{33}-b_{11}$	$b_{12}\theta_{32}-A_{31}\theta_{11}$	A_{32}	$A_{33}-b_{11}$	A_{34}	$b_{12}\theta_{32}-A_{31}\theta_{11}$
			A_{42}	A_{43}	$A_{44}-b_{11}$	$b_{12}\theta_{42}-A_{41}\theta_{11}$

$N=1$ and $M=2$					$N=2, M=2$
$\mathbf{Q}_{12}$				$\bar{\nabla}_{12}$	
$A_{22}-b_{11}$	A_{23}	$-b_{12}$	0	$b_{13}\theta_{23}-A_{21}\theta_{11}$	See formula (11b)
A_{32}	$A_{33}-b_{11}$	0	$-b_{12}$	$b_{13}\theta_{33}-A_{31}\theta_{11}$	
$-b_{21}$	0	$A_{22}-b_{22}$	A_{23}	$b_{23}\theta_{23}-A_{21}\theta_{12}$	
0	$-b_{21}$	A_{32}	$A_{33}-b_{22}$	$b_{23}\theta_{33}-A_{31}\theta_{12}$	

APPENDIX B

Numerical analysis of the double collocation method

For any given M (= number of internal collocation points in the space domain) the partial differential equation is reformulated into M ordinary differential equations in t

$$\frac{\mathrm{d}\bar{\theta}}{\mathrm{d}t} = \mathbf{B}_{\mathrm{int}}^{(M)}\ \bar{\theta} + \bar{F}(\bar{x}, t, \bar{\theta}) \qquad \text{(B1)}$$

where $\mathbf{B}_{\mathrm{int}}^{(M)}$ is defined by (9) and $\bar{\theta}$ is the vector of ordinates at the collocation abscissae $\bar{x}$. In the following analysis the nonlinearities represented by $\bar{F}$ will be disregarded and consequently the analysis is only strictly applicable to the linear equation (6)

The time interval [0, 1] is divided into n subintervals of equal length $1/n$ and a collocation method with N internal t-abscissae is used in each time interval

The method follows as a discretization in two orthogonal directions and the error analysis is logically divided into three parts (1) What is the error of the discretization of (6) into M ordinary differential equations (B1) supposing that an ideal (i e analytical) solution of (B1) exists? (2) What further error is introduced by the discretization of each of the M equations (B1) in the t-direction? (3) What is the combined error of (1) and (2)?

Questions (1) and (2) can be answered rigorously In the first part of the analysis it will be shown that the solution of (B1) by an ideal method is rapidly convergent to the solution of (6) In the second part it will be shown that the numerical solution of (B1) using n subintervals with N interior collocation points in each subinterval is convergent to the exact solution of (B1) The relative error

$$\epsilon(t) = \frac{Ex - App}{Ex} \qquad \text{(B2)}$$

where Ex is the exact solution of (B1) will be found as a function of N and n, and a comparison with Runge–Kutta methods of different order N will be given

The answer to the third question can not be given unambiguously, since different n and N will give a small relative deviation $\epsilon(t)$ from the exact solution of (6) in different time intervals One might compare $\epsilon(n,N,M)$ at a fixed t (for example $t = 1$) It will be shown that $\epsilon(t)$ decreases with t until a very low error level with an almost constant, small slope $\mathrm{d}\epsilon/\mathrm{d}t$ is reached for $t > 1$ This "stationary" error level and $\mathrm{d}\epsilon/\mathrm{d}t$ $(t > 1)$ are used as sufficiently objective criteria for comparing the approximations at different M, N and n

There is also a choice as to which quantities should be compared The value of $\theta_{av}(t)$ is independent of x and has been preferred to $\theta(t)$ at one of the collocation points x_i Figures B1, B2, B3 show the relative error of $\theta_{av}(t)$ as a function of M and N at $n = 10$ For each M and n there will be a certain N value, above which $\epsilon(t)$ is independent of N This N value is clearly dependent on n For $n = 10$ and $(M,N) = (1,2)$, $(2,3)$ and $(3,4)$, $\epsilon(t)$ represents the relative error due to the discretization involved in (B1) The limit solutions for $M = 2$ and 3 rapidly converge to the correct solution of (6) when $t > 0\,4$–$0\,6$ The relative error of these limit solutions for various M are compared in Table B1 It follows from the figures, that the concept of a constant slope $\mathrm{d}\epsilon/\mathrm{d}t$ for $t > 1$ is sensible, and it is further discussed below The slopes for $t > 1$ are likewise shown in Table B1

Table B1 Relative accuracy $\epsilon(1)$ of θ_{av} at $t = 1$ and $\mathrm{d}\epsilon/\mathrm{d}t$ for $t > 1$ in the limit solution n or N large

	$M=1$	$M=2$	$M=3$
$\epsilon(1)$	$-4\,9\ \ 10^{-3}$	$-2\,6\ \ 10^{-5}$	$-1\,5\ \ 10^{-8}$
$\frac{\mathrm{d}\epsilon}{\mathrm{d}t}(t>1)$	$2\,8\ \ 10^{-2}$	$3\,75\ \ 10^{-5}$	$4\,5\ \ 10^{-9}$

For $M = 3$ $|\epsilon(\theta_{av}, t)|$ will be less than 10^{-7} for $1 < t < 27$, which means that for all practical purposes θ_{av}(exact, t) can be represented by the exact solution of

$$\frac{\mathrm{d}\bar{\theta}}{\mathrm{d}t} = \mathbf{B}_{\mathrm{int}}^{(3)}\ \bar{\theta},\ \bar{\theta}(0) = (1, 1, 1)$$

$$\theta_{av} = w_1\ \theta_1 + w_2\ \theta_2 + w_3\ \theta_3$$

where w_i are taken from Table 2 of [10]

Convergence of the limit solutions

For any M the solution of (B1) is a sum of exponentials with exponents equal to the eigenvalues of $\mathbf{B}_{\mathrm{intt}}^{(M)}$, a matrix

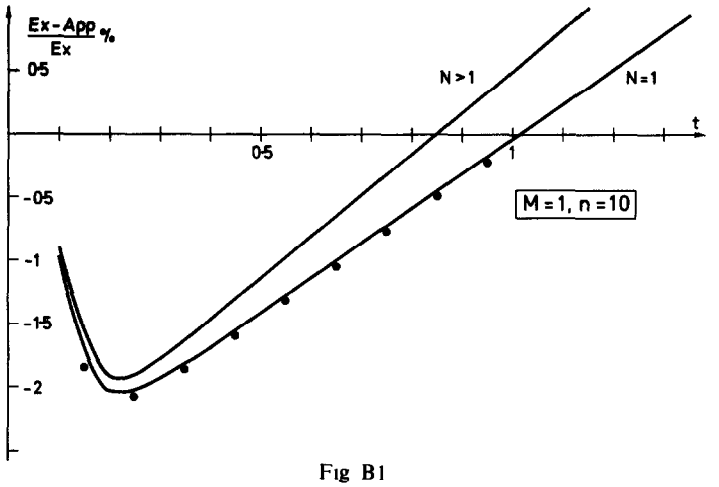

Fig B1

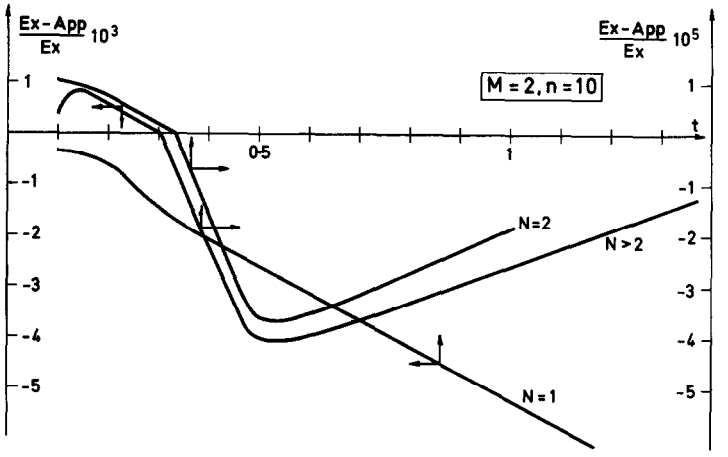

Fig B2

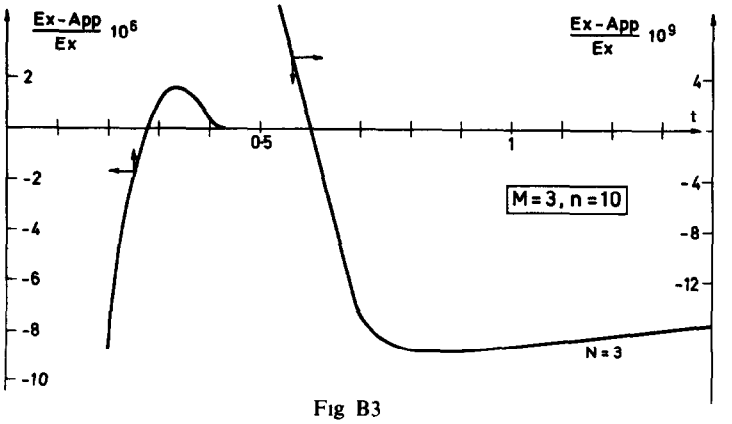

Fig B3

Figs B1–B3 Relative error ϵ in the average dimensionless temperature of a slab after contact with a medium at constant temperature for a time t ϵ (t) is shown for different M and N and constant $n = 10$

formed by deletion of the last column and row of matrix $\mathbf{B}^{(M)}$ in (10) Table 2 It is however shown in[10], that the eigenvalues of $\mathbf{B}_{\text{intt}}^{(M)}$ are approximations to the eigenvalues of

$$\frac{d^2y}{dx^2}+\lambda y=0 \quad y(1)=y'(0)=0 \tag{B4}$$

The eigenvalues of (B4) are $(2k-1)^2\pi^2/4$, $k=1,2$ ---- Since these eigenvalues are equal to the exponents in (7), the limit solutions (B1) converge to the correct functions if the eigenvalues of $\mathbf{B}_{\text{intt}}^{(M)}$ converge to the eigenvalues of (B4) for $M\to\infty$ The first eigenvalue is of primary interest, since the solution of (7) is dominated by the first term for $t>0\,6$ For $t>1$ the second term of $\theta_{av}(t)$ is less than 10^{-10} of the first term, and every term except the first may be neglected for $t>1$, if the working accuracy level of the computations is 10 D

An analysis of the characteristic equation for $B_{\text{intt}}^{(M)}$[12] shows, that the difference $\epsilon^{(M)}=\lambda_1\ (\text{true})-\lambda_1^{(M)}$, where $\lambda_1\ (\text{true})=\pi^2/4$ and $\lambda_1^{(M)}$ is the smallest eigenvalue of $B_{\text{intt}}^{(M)}$, is given by

$$\epsilon^{(M)}<-\frac{\lambda_1^{M+5/2}}{(4M+3)(2M+1)(M+1)2^{2M-2}\prod_{k=1}^{2M}(2k+1)\prod_{k=1}^{M-1}k(k+1)}$$

Values of $\epsilon^{(M)}$ are shown in Table B2

Table B2 Convergence of $\lambda_1^{(M)}$ to $\pi^2/4$ by orthogonal collocation and by Crank–Nicholson's method (C N)

$\epsilon^{(M)}=\frac{\pi^2}{4}-\lambda_1^{(M)}$	$M=1$	$M=2$	$M=3$
C N	1 2 10^{-1}	5 5 10^{-2}	3 2 10^{-2}
O C	−4 0 10^{-2}	−4 7 10^{-5}	−1 3 10^{-8}

The approximations $\lambda_1^{(M)}$ are indeed rapidly convergent to λ_1 since $\epsilon^{(M)}<10^{-3}\epsilon^{(M-1)}$ for all M

Any other selection of M collocation points will give a larger error $\epsilon^{(M)}$ for large M

The Crank–Nicolson method for solution of (B4) is also analyzed in [12] The convergence is shown to be linear

$$\log\epsilon^{(M)}=\log\left(\frac{\pi^2}{4}-\lambda_1^{(M)}\right)\sim+\log\frac{\pi^2}{192}-2\log M$$

The approximate eigenvalues are slowly convergent to λ_1 (true) as shown in the first row of Table B2

This means that the orthogonal collocation method is much preferable to the Crank–Nicholson method for $t>0\,6$, since the efficiency of the approximation is determined by the rate of convergence of $\lambda_1^{(M)}$ to λ_1 for these t values

Even if a considerable effort is devoted to the solution of the ordinary differential equations, which result after a Crank–Nicholson discretization in the x-direction no high accuracy will be obtained for large t, since the limit solution ($\Delta t\to 0$) of the resulting ordinary differential equations is different from the correct solution of (6) even for large M

The collocation solutions of (B1) for $M=1$ and $M=2$ are given below The parameters are taken from [10] Table 2

$$M=1\quad \theta(x=\sqrt{(\tfrac{1}{3})},t)=\exp(-2\,5t),$$
$$\theta_{av}(t)=\tfrac{5}{6}\exp(-2\,5t) \tag{B5}$$

$$M=2$$
$$\begin{pmatrix}\theta(x=x_1,t)\\ \theta(x=x_2,t)\end{pmatrix}=\frac{\exp((-14+\sqrt{133})t)}{2\sqrt{133}}\begin{pmatrix}(3\sqrt{7}+\sqrt{(133)}+7)-(3\sqrt{7}-\sqrt{(133)}+7)\exp(-2t\sqrt{133})\\ (-3\sqrt{7}+\sqrt{(133)}+7)-(-3\sqrt{7}-\sqrt{(133)}+7)\exp(-2t\sqrt{133})\end{pmatrix}$$
$$=0\,0433555\ \exp(-2\,46744\ t)\begin{pmatrix}26\,46981-3\,40469\ \exp(-23\,06512\ t)\\ 10\,59531+12\,46981\ \exp(-23\,06512\ t)\end{pmatrix} \tag{B6}$$

$$\theta_{av}(t)=w_1^{(2)}\ \theta(x=x_1,t)+w_2^{(2)}\ \theta(x=x_2,t)$$

Inserting $t=1$ in (B5) and in $8/\pi^2\exp(-\pi^2t/4)$

$$\theta_{av}(M=1,t=1)=\tfrac{5}{6}\ \exp(-2\,5)$$
$$\cong(1-4\,9\ \ 10^{-3})\ \frac{8}{\pi^2}\exp(-\pi^2t/4) \tag{B7}$$

The constant 4 9 10^{-3} is found by comparing the solutions at $t=1$

Since 4 9 $10^{-3}\ll 1$ it follows that

$$\epsilon(M=1,t)\sim-4\,9\ \ 10^{-3}+\left(\frac{d\epsilon}{dt}\right)_{t=1}\ t \tag{B8}$$

$(d\epsilon/dt)_{t=1}$ is calculated by comparison of (B5) and $8/\pi^2\exp(-\pi^2t/4)$ at $t=1\,2$ The result is shown in Table B1 By the same reasoning $(d\epsilon/dt)_{t=1}$ is computed for $M=2$ and 3 Increasing M by one always results in a reduction of $(d\epsilon/dt)_{t=1}$ by a factor $>10^3$, a result which is to be expected from the convergence of $\lambda_1^{(M)}$ of Table B2 to λ_1

Finite N and n

Having discussed the convergence of the solution of (B1) to the solution of (6) we shall now proceed to discuss the convergence of different forward integration schemes to the exact solution of (B1)

For $N\leq 4$ and N'th order Runge–Kutta method requires N M function evaluations per time step $1/n$ In the most general case the double collocation method requires the solution of M $(N+1)$ non-linear algebraic equations per time step If the derivatives $\nabla^2\theta$ and $\partial\theta/\partial t$ are multiplied with constants only, the much simpler iterative scheme (14) can be used If K is the necessary number of iterations per time step to obtain a stationary solution to (14), this corresponds to K M $(N+1)$ function evaluations $+K$ M $(N+1)$ multiplications per time step The double collocation method is thus at least K times slower than the Runge–Kutta method

This extra computing expenditure per time step is only justified, if the same accuracy can be obtained for a significantly smaller n

A Runge–Kutta method is in effect a reformulation of a truncated Taylor series expansion of the dependent variable from $t = t$ to $t = t + \Delta t$ Consider the simplest version of (B1) with $M = 1$ This follows from (B9) for $k = -B_{11}^{(1)} = 2\,5$

$$\frac{dy}{dt} + ky = 0 \quad y(t=0) = y_1 = 1 \tag{B9}$$

The approximate solution of (B9) by an Nth order Runge–Kutta procedure is

$$y(t) = \left(\left(1 - b + \frac{b^2}{2} - \frac{b^3}{6} + \quad (-1)^N \quad \frac{b^N}{N!}\right)^n\right)^t \tag{B10}$$

with $b = k/n$

For the collocation method the following results are obtained

$N = 1$

Let $y = (y_1, y_2, y_3)$ be the ordinates at $t = 0$, $t = 1/2n$ and $t = 1/n$, i e the ordinates at the collocation points of the first time step The collocation equations for y_2 and y_3 are

$$\begin{aligned} n(A_{21}y_1 + A_{22}y_2 + A_{23}y_3) + ky_2 = 0 \quad & by_2 + y_3 = y_1 = 1, \\ \text{or} \\ n(A_{31}y_1 + A_{32}y_2 + A_{33}y_3) + ky_3 = 0 \quad & -4y_2 + (3+b)y_3 \\ & = -y_1 = -1 \end{aligned}$$

Solving for y_3

$$y_3 = y\left(t = \frac{1}{n}\right) = \frac{(4-b)y_1}{b^2+3b+4} = \left(1 - b + \frac{b^2}{2} - \frac{b^3}{8} \quad \right) \quad 1 \tag{B11}$$

$$y_3(t) = \left(\frac{4-b}{b^2+3b+4}\right)^{nt} = \left(1 - b + \frac{b^2}{2} - \frac{b^3}{8} \quad \right)^{nt} \tag{B12}$$

$N = 2$

The collocation equations for $\bar{y} = (y_2, y_3, y_4)$ are

$$(\mathbf{A}_{\text{int}}^{(2)} - b\mathbf{I})\bar{y} = \mathbf{Q}\bar{y} =$$

$$\begin{pmatrix} a+b & a & 1-a \\ -a & -a+b & 1+a \\ 3(a-1) & -3(a+1) & 7+b \end{pmatrix} \begin{pmatrix} y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} a+1 \\ 1-a \\ 1 \end{pmatrix} y_1$$

with $a = \sqrt{3}$ and $b = k/n$

The last row of $\mathbf{Q}^{-1}$ is computed by an analytic inversion of $\mathbf{Q}$

$$\bar{\mathbf{Q}}_3 = \left(\frac{6a^2 - 3ab + 3b}{N}, \frac{6a^2 + 3ab + 3b}{N}, \frac{b^2}{N}\right)$$

$$N = 12a^2 + 6a^2b + b^3 + 7b^2 + 6b$$

y_4 is the scalar product of $\bar{\mathbf{Q}}_3$ and $(a+1, 1-a, 1)$

$$y_4 = \frac{12a^2 + 6b + b^2 - 6a^2b}{N}$$

$$= \frac{(b-6)^2}{(b+3)(b^2+4b+12)}$$

$$= 1 - b + \frac{b^2}{2} - \frac{b^3}{6} + \frac{b^4}{24} - \frac{b^5}{108} \tag{B13}$$

(B10), (B12) and (B13) are all convergent to the correct solution $\exp(-kt)$ for $n \to \infty$ (since $(1 - k/n)^n \to \exp(-k)$ for $n \to \infty$) Let the relative error at $t = 1$ be $\epsilon(1)$, then

$$\epsilon(t) = (1 + \epsilon(1))^t - 1 \sim \epsilon(1)t,$$

if t is not too large and n is sufficiently large to ensure that $\epsilon(t = 1) \ll 1$ It also follows, that the relative error at any fixed t is multiplied by a factor α, if k and n are both multiplied by α

For $M > 1$ each of the exponentials of the solution of (B1) is approximated by an expression like (B10), (B12) and (B13) This immediately follows for the Runge–Kutta formulae, but the derivation for the collocation method is somewhat complicated and will only be outlined

For $N = 1$ and $M = 2$ the following formula is obtained by direct application of the collocation principle

$$\begin{pmatrix} \theta_{31} \\ \theta_{32} \end{pmatrix} = \frac{\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \theta_{12} \\ \theta_{12} \end{pmatrix}}{1 - \frac{3}{4}\beta + \frac{1}{16}\alpha^2 + \frac{1}{4}\beta^2 + \frac{1}{16}\alpha - \frac{3}{16}\alpha\beta} \tag{B14}$$

$$a_{11} = 1 + \frac{1}{4}(b_{11} - 3b_{22}) + \frac{1}{4}b_{22}^2 + \frac{b_{22}\alpha}{16}$$

$$a_{22} = 1 + \frac{1}{4}(b_{22} - 3b_{11}) + \frac{1}{4}b_{11}^2 + \frac{b_{11}\alpha}{16}$$

$$a_{12} = b_{12}\left(1 - \frac{1}{4}\beta - \frac{1}{16}\alpha\right)$$

$$a_{21} = \frac{b_{21}}{b_{12}} a_{12}$$

$$\alpha = b_{11}b_{22} - b_{12}b_{21}, \quad \beta = b_{11} + b_{22}, \quad b_{ij} = \frac{B_{ij}^{(2)}}{n}$$

$$\theta_{11} = \theta_{12} = 1 \text{ at } t = 0$$

Inserting the values of b_{ij} from Table 2 of [10] the denominator of (B14) is reduced to

$$N = 1 + \frac{21}{n} + \frac{3199}{n^2} + \frac{21}{n^3} \quad \frac{63}{n^4} + \frac{63^2}{n^4}$$

$$= 16(b_2^2 + 3b_2 + 4)(b_1^2 + 3b_1 + 4)$$

where $b_1 = -(1/n)\lambda_1$ and $b_2 = -(1/n)\lambda_2$ λ_1 and λ_2 are the eigenvalues $-14 + \sqrt{133}$ and $-14 - \sqrt{133}$ of $\mathbf{B}_{\text{int}}^{(2)}$

Using the decomposition of the denominator suggested above one finally arrives at

$$\theta_{31}(t) = (3\sqrt{7} + \sqrt{(133)} + 7)\left(\frac{4-b_1}{b_1^2+3b_1+4}\right)^{nt}$$

$$+ (3\sqrt{7} - \sqrt{(133)} + 7)\left(\frac{4-b_2}{b_2^2+3b_2+4}\right)^{nt}$$

$$\theta_{32}(t) = (-3\sqrt{7} + \sqrt{(133)} + 7)\left(\frac{4-b_1}{b_1^2+3b_1+4}\right)^{nt}$$

$$+ (-3\sqrt{7} - \sqrt{(133)} + 7)\left(\frac{4-b_2}{b_2^2+3b_2+4}\right)^{nt} \tag{B15}$$

(B15) is seen to be the correct discrete version of (B6) with $N = 1$

Formulae for higher N or M can now be constructed by analogy with (B15)

The approximate solutions suggested by (B10) and (B12), (B13), (B15) are different in two respects

(1) The functions $F(b)$ in (B11) and (B13) can be shown to be numerically less than 1 for all positive b (i e all decreasing exponentials) The function $F(b)$ in (B10) may be numerically greater than 1 for all N

$|F(b)|$ in (B10) > 1 for $b > b_{crit}$ where $(N, b_{crit}) = (1,2)$, (2,2), (3,2 531), (4,2 785) etc

This means that the approximate solution will increase exponentially if $n < n_{crit}$, where n_{crit} depends on k (partial instability in the nomenclature of Fox[18, pp 49, 241] This phenomenon is not very critical for $M = 1$ ($k = 2\,5$) but for $M \gg 1$ very large negative eigenvalues occur ($\approx -2\,46, -25, -70$), and the correct solution is rapidly swamped if some of the strongly decreasing exponentials are inadvertently converted into increasing exponentials by the approximation The collocation formulae (B12) and (B13) on the other hand tend to underestimate the influence of the strongly decreasing exponentials at very small t (since both expressions rapidly become small for $b > 1$–2) and they will seem to converge rapidly to the correct solution for increasing t This self-stabilizing property was noted in [9] when the

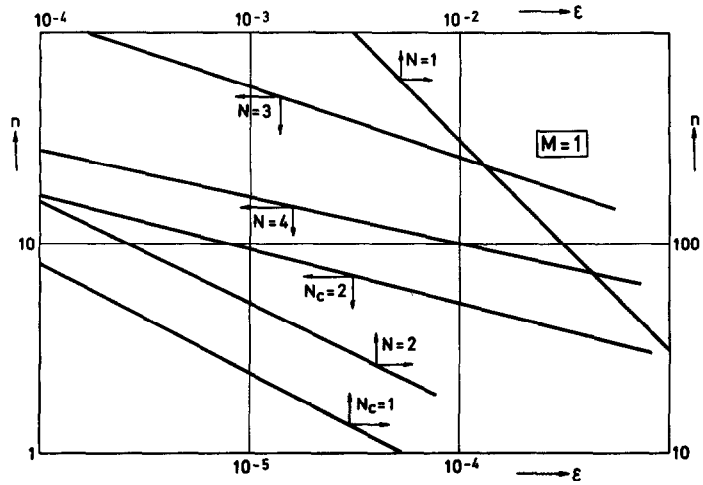

Fig B4

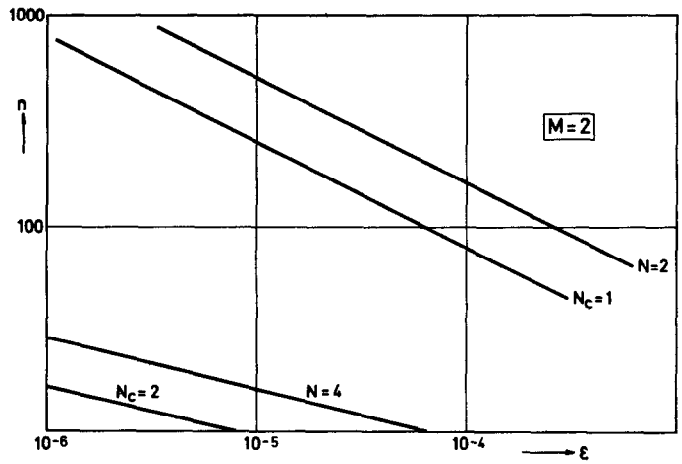

Fig B5

Figs B4–B5 Relative error in the solution of the collocation equations at $t = 1$ as a function of the number n of subintervals in [0, 1] The parameter N_c is the number of internal collocation points and N the order of a Runge–Kutta method used in the step by step integration from 0 to 1

collocation method was used to integrate kinetic equations with extremely small time constants

The lowest order collocation formula (B12) is seen to give an erroneous, damped oscillation for $b > 4$ This phenomenon is also observed in the uneven order Runge–Kutta formula It is much less dangerous than the instability phenomenon discussed above As long as the eigenfunction with the smallest eigenvalue is correctly reproduced, the solution will be accurate even for small t

(2) Another difference between the two sets of formula (B10) and (B12), (B13) lies in their relative speed of convergence to the solution of (B1)

If the last correct term in the power series is b^l then the relative error is approximately

$$\epsilon(n, \text{fixed } t) \sim n^{-l} \tag{B16}$$

In the Runge–Kutta formulae $\epsilon(n) \sim n^{-N}$, where N is the order of the formula, but the collocation method with N internal collocation points has $\epsilon(n) \sim n^{-2N}$, which corresponds to a significant reduction in computing time, especially for $M > 1$, where the overall obtainable accuracy is very high and a high accuracy (i e a large n) in the integration of (B1) is justified

Figure B4 and B5 show the numerical value of the relative error in the approximations (B10), (B12) and (B15) to the solution of (B1) for $M = 1$ and 2 N_c is the number of internal collocation points in each time interval and N the order of the Runge–Kutta formula

The collocation method with $N_c = 2$ is comparable with a 4'th order Runge–Kutta method Since however the term b^5 in the power series is almost correct in the collocation method while it is truncated in the 4'th order Runge–Kutta method

n(collocation) ~ 0 5 n(4'th order R K) for the same ϵ This difference is attenuated, if the eigenvalues of $\mathbf{B}_{\text{intt}}^{(M)}$ are further separated as observed in [9]

The figures clearly show, that $N_c = 2$ is much better than $N_c = 1$ (n is reduced by a factor 30) The "constant" error level for $t > 1$ is $\sim 10^{-5}$ for $M = 2$ and $\sim 10^{-2}$ for $M = 1$ Clearly a much higher n or a higher N_c is required for $M = 2$, if the approximation in the t-direction is to be of comparable accuracy

It should be remembered, that since the limiting solution deviates from the correct solution of (6) for small M a small n value might give a better solution to (6) than the high n values needed to obtain a good solution to (B1)

Résumé—Une méthode permettant de résoudre la parabole PDE au moyen de l'interpolation des opérateurs différentiels dans les deux dimensions est décrite dans cet ouvrage La méthode est basée sur le principe des méthodes précédemment publiées et relatives à la solution des équations différentielles ordinaires par une collocation orthogonale On démontre que cette méthode est particulièrement économique et très stable, par comparaison à la méthode Crank–Nicholson ou aux méthodes explicites, telles que la méthode Runge–Kutta—une méthode de 4 ème ordre L'équation de chaleur linéaire est utilisée pour illustrer le principe de la méthode et discuter ses propriétés de convergence

Zusammenfassung—Es wird eine Methode für die Losung parabolischer, partieller Differentialgleichungen durch Interpolierung der Differentialoperatoren in zwei Dimensionen beschrieben Die Methode wurde auf Grund fruher beschriebener Methoden für die Losung gewohnlicher Differentialgleichungen durch orthogonale Anordnung entwickelt Es wird gezeigt, dass diese Methode im Vergleich mit der klassischen Crank–Nicholson oder mit expliziten Methoden wie der Runge–Kutta Methode 4 Ordnung ausserst wirtschaftlich und sehr stabil ist Die lineare Warmegleichung wird verwendet um das Prinzip der Methode zu erlautern und ihre Konvergenzeigenschaften zu behandeln