ADVANCES IN COMPUTER METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS - II

Proceedings of the Second IMACS (AICA) International Symposium
on Computer Methods for Partial Differential Equations
held at Lehigh University - Bethlehem, Pennsylvania, U.S.A.
June 22-24, 1977

Edited by
R. VICHNEVETSKY
Rutgers University
New Brunswick, New Jersey (U.S.A.)

Published by IMACS (AICA)
1977

IMACS (AICA)
Dept. of Computer Science
Rutgers University
New Brunswick, N.J. 08903 U.S.A.

IMACS (AICA
E.R.M.
Avenue de la Renaissance 30
B-1040 Brussels, BELGIUM

DVANCES IN COMPUTER METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS - II
. VICHNEVETSKY (EDITOR)
UBL. IMACS(AICA) 1977

ORTHOGONAL COLLOCATION ON FINITE ELEMENTS FOR ELLIPTIC EQUATIONS

P. W. Chang and B. A. Finlayson

Department of Chemical Engineering
University of Washington
Seattle, Washington 98195

Abstract

The method of orthogonal collocation on finite elements (OCFE) combines the features of orthogonal collocation with those of the finite element method. The method is illustrated for a Poisson equation (heat conduction with source term) in a rectangular domain. Two different basis functions are employed: either Hermite or Lagrange polynomials (with first derivative continuity imposed to ensure equivalence to the Hermite basis). Cubic or higher degree polynomials are used. The equations are solved using an LU-decomposition for the Hermite basis and an alternating direction implicit (ADI) method for the Lagrange basis.

Summary

The ADI method with Lagrange basis functions converges in a few iterations using the least computer time and storage when the element sizes are uniform. The optimal iteration parameters are given. When elements have widely different sizes the Hermite basis with a direct LU decomposition requires the least computer time.

For one problem treated here, it is more advantageous to increase the degree of polynomial of the basis functions than to use more elements. For problems with few elements, the collocation method with Hermite basis functions uses fewer multiplications to do an LU-decomposition than the Galerkin method with Hermite basis functions. As the number of elements increases the Galerkin method is preferred. The Ritz method is always preferred if it is applicable.

Introduction

We solve the problem

$$\nabla^2 T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = f(x,y) \quad \text{in A} \tag{1}$$

$$T = g(x,y) \quad \text{on C} \tag{2}$$

for a rectangular domain $x\varepsilon(0,1)$, $y\varepsilon(0,1)$. Problem I uses $f = -4$, $g = 0$ and corresponds to fully developed flow of a Newtonian fluid in a rectangular duct, or heat transfer with zero wall temperature and uniform heat generation. Problem II uses $f = 0$, $g = 0$, except on $y = 0$ where $g = 1$. This problem corresponds to heat transfer with one surface maintained at unit temperature, and the solution has a singularity at the corners $(x,y) = (0,0), (1,0)$. This problem is linear and easily solved using separation of variables, fast Fourier transforms, etc. However, we use it as a prototype problem to test methods that can be applied to nonlinear problems. The problem has a variational principle, so that the Ritz method is applicable, but we concentrate on the collocation and Galerkin methods, which are applicable to all problems.

Lagrange Interpolation and ADI

In each element the unknown T is approximated by

$$T^{k\ell}(x,y) = \sum_{i=1}^{NPX} \sum_{j=1}^{NPY} L_i(u)\, L_j(v)\, C_{ij}^{k\ell} \tag{3}$$

$$u = (x-x_k)/\Delta x_k \quad , \quad v = (y-y_\ell)/\Delta y_\ell \tag{4}$$

Here $C_{ij}^{k\ell}$ is the value of $T^{k\ell}$ at the collocation point (u_i, v_j) in the $k\ell$-th element. The collocation points u_i are the zeroes of the (NPX-1)-th degree shifted Legendre polynomial on $0 \le u \le 1$. Similar definitions apply in the y(v) direction. The element sizes are $\Delta x_k = x_{k+1} - x_k$, $\Delta y_k = y_{k+1} - y_k$, where $\{x_k\}$ denote the x positions of the element sides. Thus a rectangular array of elements is used.

Orthogonal collocation[1] is applied at each interior collocation point of each element $(k\ell)$.

$$\frac{1}{\Delta x_k^2} \sum_{n=1}^{NPX} B_{in}\, T_{nj}^{k\ell} + \frac{1}{\Delta y_\ell^2} \sum_{n=1}^{NPY} B_{jn}\, T_{in}^{k\ell} = f(x_i^{k\ell}, y_j^{k\ell}) \tag{5}$$

$i = 2, \ldots, NPX-1$; $j = 2, \ldots, NPY-1$, $k = 1, \ldots, NEX$; $\ell = 1, \ldots, NEY$.

NEX and NEY are the number of elements in the x and y directions, respectively, and NPX and NPY are the degree + 1 of the polynomial in the x and y direction. Cubic polynomials have NP = 4. In addition to these equations we require the solution and first derivative be continuous across element boundaries[2]. The boundary conditions are satisfied at the collocation points on the boundary, e.g. at $(x,y) = (0, v_j \Delta y_k + y_k)$, $k=1,\ldots,NEY$;

j=2,..., NPY-1. In addition the function and first derivative are made continuous on the boundary, and the boundary conditions are satisfied at the four corners of the domain. The approximation is in C^2 at least on each element, and is in C^1 globally.

The algebraic equations (5) plus the boundary and continuity conditions are solved using ADI:

$$\omega T_{ij}^{k\ell,s+1/2} - \frac{1}{\Delta x_k^2} \sum_{n=1}^{NPX} B_{in}\, T_{nj}^{k\ell,s+1/2} = \quad (6)$$

$$= \omega T_{ij}^{k\ell,s} + \frac{1}{\Delta y_\ell^2} \sum_{n=1}^{NPY} B_{jn}\, T_{in}^{k\ell,s} - f_{ij}^{k\ell}$$

$$\omega T_{ij}^{k\ell,s+1} - \frac{1}{\Delta y_\ell^2} \sum_{n=1}^{NPY} B_{jn}\, T_{in}^{k\ell,s+1} = \quad (7)$$

$$= \omega T_{ij}^{k\ell,s+1/2} + \frac{1}{\Delta x_k^2} \sum_{n=1}^{NPX} B_{in}\, T_{nj}^{k\ell,s+1/2} - f_{ij}^{k\ell}$$

The iteration parameter, ω, strongly affects the rate of convergence with iteration number, s. Eq. (6) is solved line by line at constant y (i.e. for j=2,...,NPY-1, ℓ=1,...,NEY). Since the matrix is the same for all j, ℓ an LU decomposition is performed only once per problem. The matrix is block diagonal and the LU decomposition does not require pivoting. Each block is of size NPX x NPX, with one line overlap between elements. After one half iteration the $T_{ij}^{k\ell,s+1/2}$ is known for j=2,...,NPY-1 but not for j=1 or NPY. These are obtained by smoothing the solution in the y-direction at each x value. The other half iteration is performed similarly, line by line at constant x, followed by smoothing in the x-direction.

The iteration error is governed by

$$(\underline{\underline{E}}^s - \underline{\underline{EX}}^s) = \underline{\underline{M}}^s (\underline{\underline{E}}^\circ - \underline{\underline{EX}}^\circ) \underline{\underline{N}}^{Ts} \quad (8)$$

where $\underline{\underline{E}}^s = \underline{\underline{T}}^s - \underline{\underline{T}}$, the difference between the solution at the s-th iteration and the exact solution to the algebraic equations. $\underline{\underline{EX}}$ is used to eliminate the values on element boundaries (which do not affect the iteration error) and $\underline{\underline{M}}$ and $\underline{\underline{N}}$ are matrices derived from those in Eqs.(6,7). (See Ref.3 for details.) We thus have

$$\| \underline{\underline{E}}^{s+1} - \underline{\underline{EX}}^{s+1} \| \le \| \underline{\underline{M}} \| \; \| \underline{\underline{E}}^s - \underline{\underline{EX}}^s \| \; \| \underline{\underline{N}}^T \| \quad (9)$$

The reduction in error from one iteration to another is thus dependent on the spectral radii of $\underline{\underline{M}}$ and $\underline{\underline{N}}^T$ (actually $\underline{\underline{MR}}$, a portion of $\underline{\underline{M}}$ and $\underline{\underline{NR}}$, a postion of $\underline{\underline{N}}^T$.

$$\| \underline{\underline{E}}^{s+1} - \underline{\underline{EX}}^{s+1} \| \le \| \underline{\underline{MR}} \|^s \; \| \underline{\underline{NR}} \|^s \; \| \underline{\underline{E}}^\circ - \underline{\underline{EX}}^\circ \| \quad (10)$$

$$\lim_{s \to \infty} \| \underline{\underline{MR}} \|^s \; \| \underline{\underline{NR}} \|^s = 0 \text{ if } \rho(MR)\rho(NR) < 1 \quad (11)$$

Clearly the error decreases faster each iteration if ρ is amaller. $\rho < 1$ is necessary for convergence, and the value of ρ depends on the iteration parameter, ω. If ω is specified, ρ is known, and $\underline{\underline{M}}=\underline{\underline{N}}$, the maximum errors at the s-th iteration decrease as

$$\text{Error} = \rho^{2s} \quad (12)$$

Thus knowledge of ρ permits calculation of the number of iterations s needed to make the iterat error less than a specified value. Sequences o iteration parameters, ω^s, are not considered he It is essential to use a ρ which is as small a possible. There is no point, of course, in reducing the iterate error (to the solution of the algebraic equations) much below the approximatio error (to the solution of the differential equations.)

The dependence of $\rho(\underline{\underline{MR}})$ on ω is shown in Figure 1 as a function of NP (the degree of polynomial is NP-1) and NE(the number of elements The optimal parameters (minimum ρ) are shown in Fig. 2. Note that ρ increases (requiring more iterations) as NP is increases, as NE is increased, or as the total number of points in one direction, NT = (NP-1)NE+1, is increased.

The spectral radius depends also on the element sizes, Δx_k and Δy_ℓ. We consider a given number of elements, say NEX, and specify an element distribution given by

$$H = \Delta x_{k+1} / \Delta x_k \quad (13$$

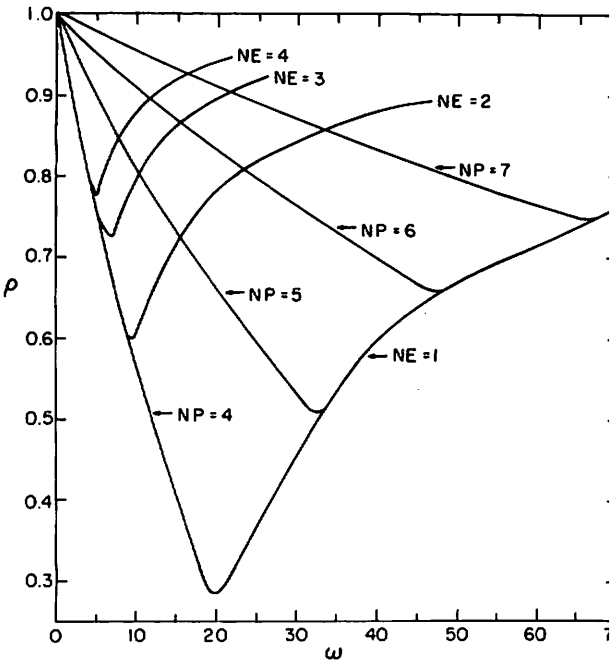


FIGURE 1. SPECTRAL RADIUS VERSUS ITERATION PARAMETER.

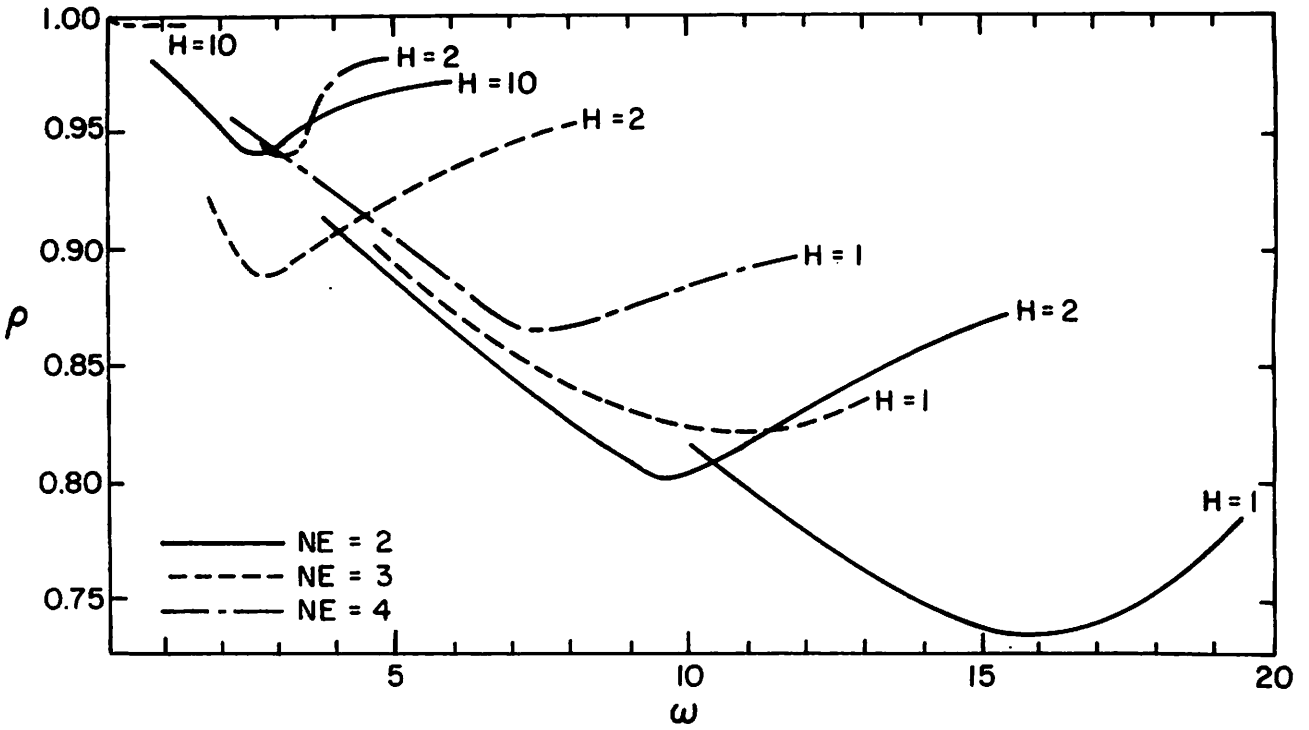


FIGURE 3. EFFECT OF NON-UNIFORM ELEMENTS.

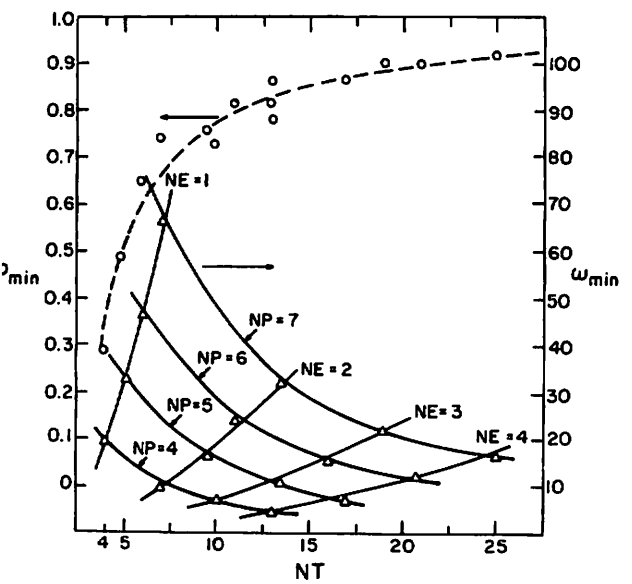


FIGURE 2. OPTIMUM PARAMETERS.

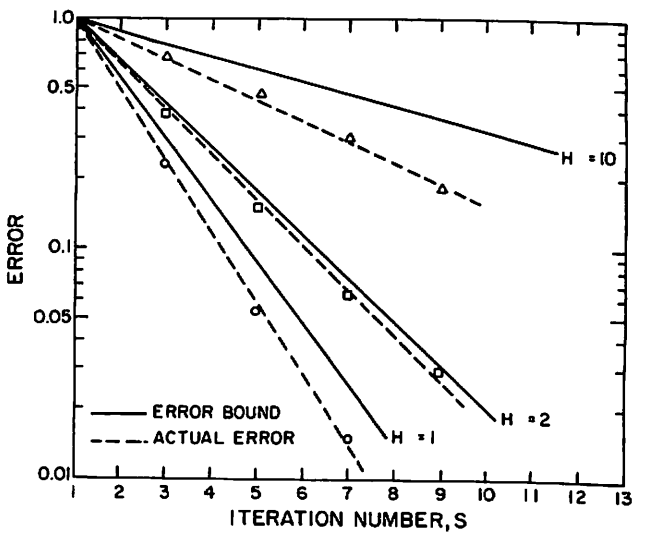
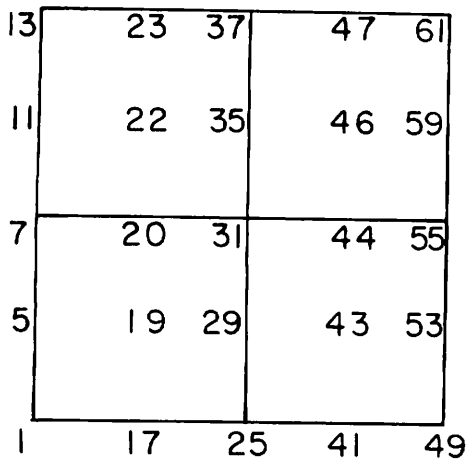
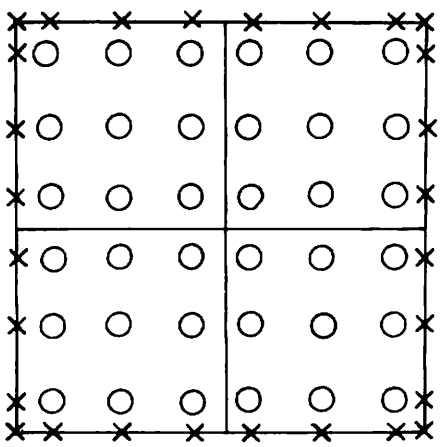


FIGURE 4. RATE OF CONVERGENCE.



(a) NUMBERING OF UNKNOWNS, NP = 5



(b) COLLATION POINTS, × BC, ○ DE

FIGURE 5.

Thus the elements are not all the same size. Fig 3 shows the dramatic effect on ρ. For example, for NE=3, NP=5, changing H from 1 to 2 to 10 changes ρ from 0.817 to 0.88 to 0.992. If the same element distribution is used in the x and y directions and the iterate error is 0.001, this increases the number of iterations from 17 to 27 to 430. Clearly the ADI method is not as suitable for very non-uniform element sizes.

The iterate error is plotted versus iteration number in Fig. 4 for problem II and the error decreases at a rate slightly faster than the theoretical upper bound given by Eq.(12).

For problems which require widely different element sizes, the ADI method is less suitable. Then we must go to a direct solution of all the equations (6,7) together. Unfortunately, the bandwidth of the matrix is too large, being roughly twice as wide as the bandwidth for Hermite polynomials. This causes a fourfold increase in decomposition time, so the Hermitian interpolation is used in these cases.

Hermite Interpolation and Direct Solution

The Hermite interpolation gives

$$T^{k\ell} = \sum_{i=1}^{NPX} \sum_{j=1}^{NPY} H_i(u)\, H_j(v)\, C_{ij}^{k\ell} \qquad (13)$$

where the Hermite polynomials are defined on $u \in [0,1]$ and the parameters $C_{ij}^{k\ell}$ are values of T, $\partial T/\partial x$, $\partial T/\partial y$, or $\partial^2 T/\partial x \partial y$ at the element corners (for NPX=NPY=4, cubic polynomials).

$$\begin{aligned} H_1(u) &= (1 + 2u)(1 - u)^2 \\ H_2(u) &= u(1 - u)^2 \\ H_3(u) &= (3 - 2u)u^2 \\ H_4(u) &= (u - 1)u^2 \end{aligned} \qquad (14)$$

For polynomials higher than cubics we add the functions

$$H_{i+4}(u) = u^{i+1}(1 - u)^2 \qquad (15)$$

Within an element $k\ell$ the derivatives of T can be related to the derivatives of H, e.g.

$$\left.\frac{\partial T^{k\ell}}{\partial x}\right|_{mn} = \frac{1}{\Delta x^k}\left.\frac{\partial T}{\partial u}\right|_{mn} = \frac{1}{\Delta x^k}\sum_{i=1}^{NPX}\sum_{j=1}^{NPY}\frac{dH_i}{du}(u_m)H_j(v_n)C_{ij}^{k\ell}$$

For cubic polynomials the x-derivatives at the four collocation points are given by a 4x16 matrix multiplying a 16x1 vector $(C_{ij}^{k\ell})$.

The differential equation (1) is satisfied at the collocation points interior to each element (see Figure 5b) and the boundary conditions are satisfied on the boundary. The value for $T^{k\ell}$ at an element edge is shared by $C_{ij}^{k\ell}$ in one element being the same variable as the corresponding $C_{ij}^{k\ell}$ in the adjacent element. Thus the approximation is automatically in C^1. The resulting matrix (of size NTX·NTY x NTX·NTY) can be

decomposed (LU) using the same block diagonal version used for ADI, except that pivoting is necessary. Here NTX = NEX(NPX-2)+2 and NTY= NEY(NPY-2)+2. We perform partial pivoting within a block. The numbering scheme for the variables is given in Fig. 5a. The extraneous zeros introduced by using the block diagonal decomposition cause additional, unnecessary, multiplications, but there are minimized by checking for zeros in the decomposition. A banded decomposition would be faster, and one with a variable band width would be even faster, but both would have to pivot.

Approximation Error

For one-dimensional problems deBoor and Swartz[4] showed that the error of the collocation method depends on the number of elements as follows:

$$\text{error} \quad \alpha \quad (1/NE)^{NP}, \quad NE \to \infty, \quad NP \quad \text{fixed} \qquad (17)$$

For global polynomials (NE=1) in one direction Ciarlet, Schultz and Varga[5] show that the error of a Galerkin method follows:

$$\text{error} \quad \alpha \quad (1/(NP-2))^{NP-2} \qquad (18)$$

A similar dependence was found empirically for the collocation method.[3] For two-dimensional problems Prenter and Russell[6] have shown that

$$\text{error} \quad \alpha \quad (1/NE)^{3-\varepsilon} \qquad (19)$$

where the 3 characterizes the continuity of the exact solution. We find for problems I and II the approximation errors follow

$$ER = C\left(\frac{1}{NP}\right)^{\alpha NP}\left(\frac{1}{NE}\right)^{\min(NP,k)} \qquad (20)$$

where NEX=NEY=NE and NPX=NPY=NP. The dependence on NP rather than NP-2 is empirical; it gives a correlation of the results at low NP (3 to 7). parameter k characterizes the continuity of solution and C and α are independent of NP and NE.

For Problem I Fig. 6 shows the error as a function of NP and NE. Here[6] k=3. and the curves for different NP $\geq$ 4 all have the same slope, since min (NP,k)=3. Fig. 7 shows the same data plotted versus NP log NP , and the straight lines are evident. For this problem the constants in Eq.(20) are C=0.07, α=0.5, k=3.

For Problem II the k=1 since the solution has a singularity. The results are shown in Fig. 8, and here C=0.038, α=0.2,k=1. This problem has discontinuous first derivatives at the corners, yet we are approximating it with a basis function in C^1. Adequate results can be achieved, however, and improvement results from using a graded mesh. The boundary conditions must be carefully applied if the correspondence between Lagrange and Hermite basis functions is to hold.

Computation time

We next compare computation time of OCFE with that of Galerkin & Ritz method. The calculation time is due to formulation of the equations, decomposition of the matrix, and fore and aft sweeps of the right-hand sides. For OCFE and Galerkin methods, both using Hermite polynomials the formulation time is similar for linear problems. For nonlinear problems the Galerkin method requires more time since more quadrature points are generally used than collocation points. We concentrate here on the number of multiplications in the decomposition and fore and aft sweeps.

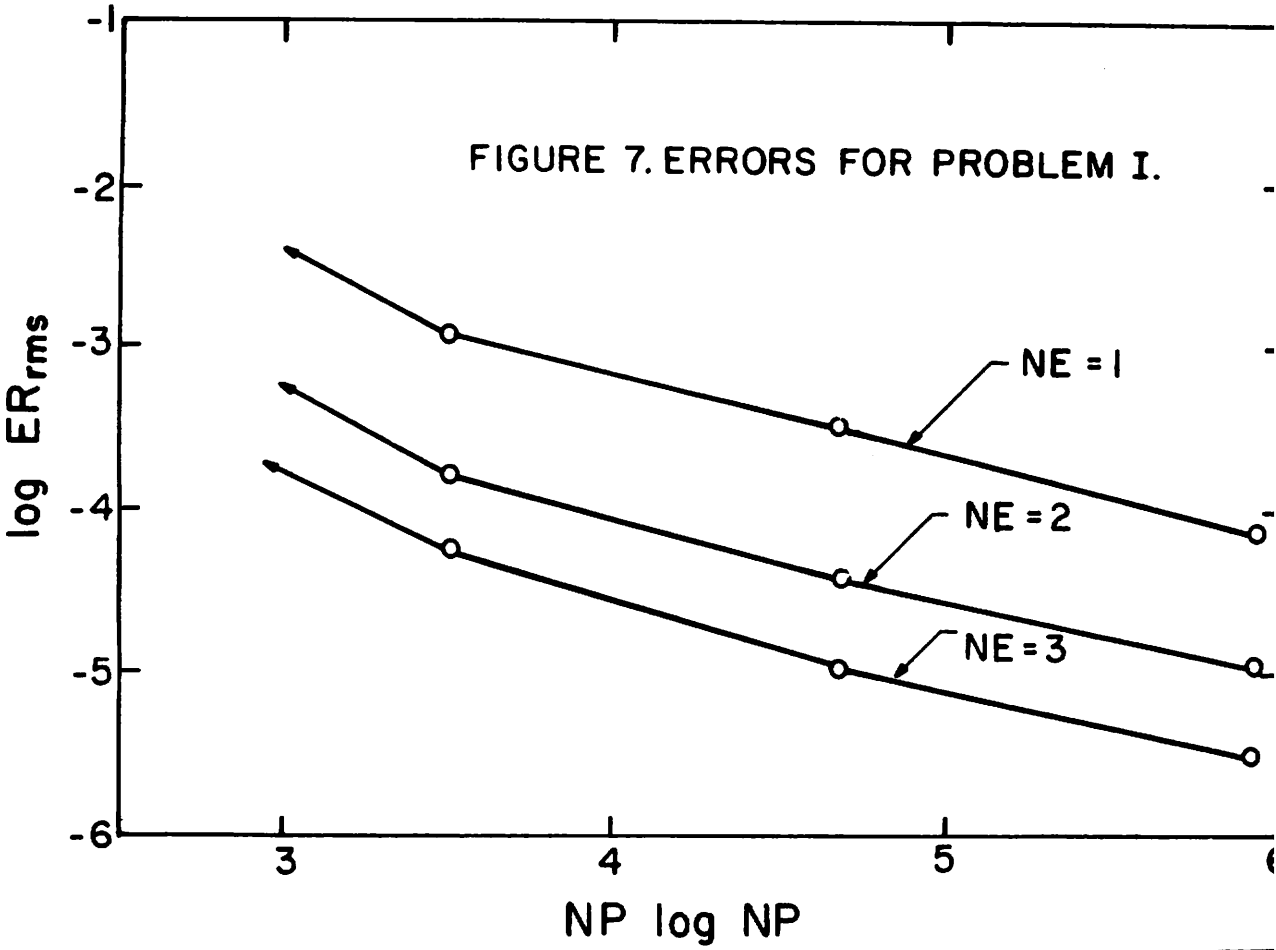
The decomposition cost depends on the method used. For elements not on the boundary and the same numbering scheme the bandwidth of collocation is significantly less than Galerkin (about 2 times smaller). However, the boundary elements cause the bandwidth of the collocation to increase. We compare here the decomposition cost for Galerkin using a banded matrix decomposition with the cost for collocation using a block diagonal matrix decomposition. This provides a penalty for the collocation method, and the penalty increases with the number of elements. Both methods would be faster using a frontal approach. Table I gives the formulas used to approximate the number of multiplications for the different methods, while Table II gives the actual multiplication count.

Comparison of Methods

For few elements, OCFE-Hermite has fewer multiplications than Galerkin-Hermite, while the reverse holds for more elements. For NP=4 the crossover point is NEX=NEP=3, or 9 elements with 64 unknowns. For NP=5, however, it is NEX=NEY=5 or 25 elements with 289 unknowns. Clearly OCFE-Lagrange with direct decomposition of all variables is not competitive. The Ritz method, with a symmetric matrix is of course preferred when applicable. The collocation method leads to an unsymmetric matrix even for self-adjacent problems with variational principles. For most problems, however, the Ritz method is not possible because no variational principle exists and the collocation method should be compared to the Galerkin method.

It is more difficult to compare the OCFE-Lagrange-ADI with OCFE-Hermite-direct because the ADI uses iteration to solve the equations and the direct method solves them in one iteration. However, let us use the data presented to make a comparison. For ADI choose an iterate error of 10% of the approximation error for that discretization. Then use Eq.(12) to find how many iterations are necessary. The results in Table III show that the Lagrange-ADI method uses only about 1/4 the time of the Hermite-direct when the elements are uniform. When the elements are non-uniform, however, ρ_{min} approaches one, s increases and the direct methods are more competitive.

Another interesting comparison is whether it is better to increase the number of elements or the degree of polynomial. Using the actual approximation error for Problem I from Fig. 6 and the multiplication count in Table 2, we can conclude that for equivalent accuracy (ER=3x10^{-4}), cubics



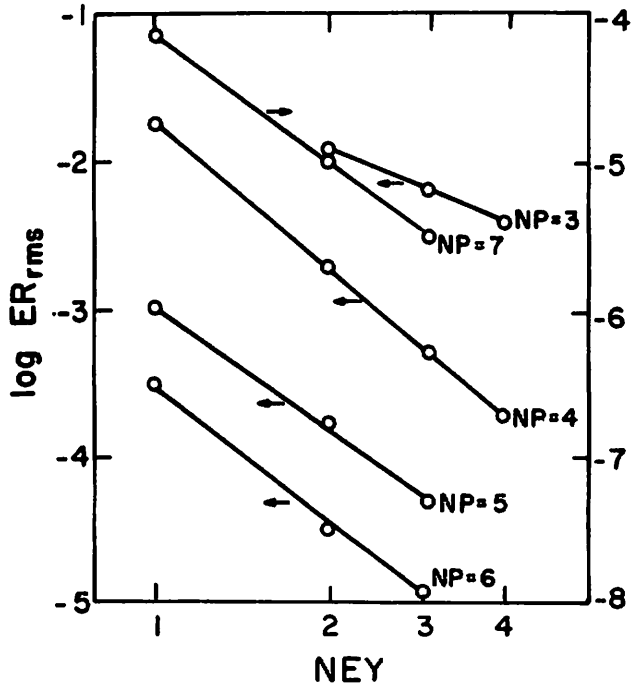


FIGURE 6. ERRORS FOR PROBLEM I.

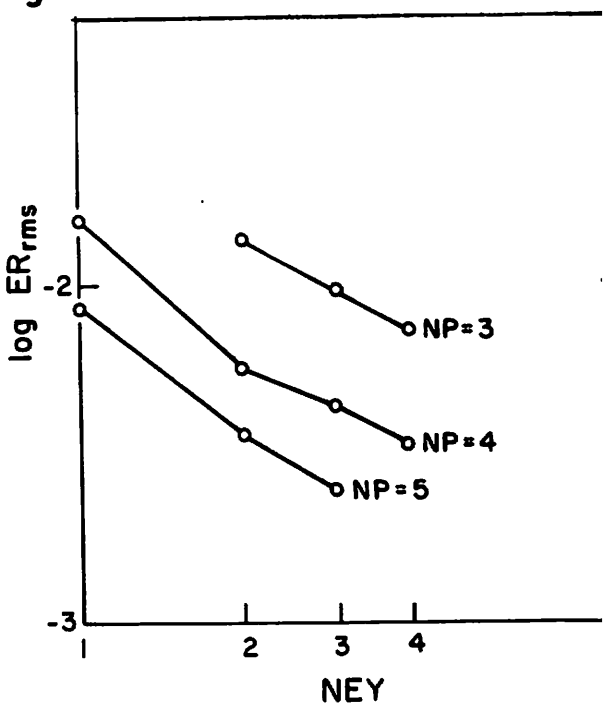


FIGURE 8. ERRORS FOR PROBLEM II.

(NP-4) require 150% more multiplications than quartics (NP=5), and quintics (NP=6) require 18% less. The difference between NP=5,6,7 is fairly small, but NP=5 is a significant improvement over NP=4. For Problem II (which has the singularity) there is almost no difference between cubics (NP=4) and quartics (NP=5).

Table 1 Decomposition Cost

Method	N.T.	Half-Bandwidth or Block Size	No. of Blocks	Approximate Cost
Galerkin	NTX · NTY	NBW = (NPY-2)·NTX+2·NPY-1		$(NT-NEW+1)\ (NBW+1)^2 + \sum_{i=1}^{NBW} i^2$
Ritz				1/2 of the above
OCFE-Hermite direct		NPY · NTX	NEY	$NEY\ NTX^3(NPY^3-8)/3$
OCFE-Lagrange direct	((NPX-1) NEX+1)· ((NPY-1) NEY+1)	NPY((NPX-1)·NEX+1)	NEY	$NEY\ ((NPX-1)\ NEX+1)^3\ (NPY^3-1)/3$

NTX = (NPX-2) NEX + 2

Table 2 - Operation Counts for Direct Methods*

NP	4	4	4	4	5	5	5	5
NE	2	3	4	5	2	3	4	5
Galerkin, FE	9.3	27.9	64.2	126	48.4	172	474	926
Ritz, FE	4.6	14.0	32.1	62.9	24.2	85.8	236	463
OCFE, Lagrange	14.4	63.0	185	1300	60.3	272	812	5700
OCFE,Hermite	7.6	27.5	72.3	157	38.6	152	426	943

*In thousands of multiplications

Table 3 - Operation Counts for OCFE-Lagrange ADI for Prob. II

$DC = (NEX\ NPX^3 + NEY\ NPY^3)/3$

$SB = (NPY-2)NEY\ NEX\ NPX^2 + (NPX-2)\ NEX\ NEY\ NPY^2$

Total Counts = (DC) + s (SB)

NP	4	4	4	5	5	4	4
NE	2	3	4	2	3	3	4
$H=\Delta x_{k+1}/\Delta x_k$	1	1	1	1	1	2	2
Approximation Error Bound	6(-3)*	4(-3)	3(-3)	4(-3)	3(-3)	3(-3)	3(-3)
ρ_{min}	0.60	0.73	0.78	0.73	0.82	0.90**	0.95**
s_{max}	8	13	17	13	21	39	77
Total Counts†	2.1	7.6	17.6	8.0	28.6	22.6	79.0
(ADI) OCFE-Hermite	.28	.28	.24	.21	.19	.82	1.09
Galerkin OCFE-Hermite	1.27	1.01	.89	1.25	1.13	1.01	.89

*$6(-3) = 6 \times 10^{-3}$ Relative iteration error is taken 10% of the approximation error bound

**Estimated

†Thousands of multiplications

References

1. Finlayson, B. A., The Method of Weighted Residuals and Variational Principles, Chap. 5, Academic Press, 1972.

2. Carey, G. F. and B. A. Finlayson, Chem. Eng. Sci. 30 587 (1975)

3. Chang, P. W., M.S. thesis, University of Washington (1975).

4. deBoor, C. and B. Swartz, SIAM J. Numer. Anal. 10 582 (1973).

5. Ciarlet, P. G., Schultz, M. H., and Varga, R. S., Num. Math. 9 394 (1967).

6. Prenter, P. M. and R. D. Russell, SIAM J. Numer. Anal. 13 923 (1976).