RICE UNIVERSITY

ROUNDING ERRORS IN THE SOLUTION OF THE ONE DIMENSIONAL HEAT EQUATION USING A GALERKIN TECHNIQUE

by

Catherine Gardner

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

Thesis Director's signature:

Houston, Texas

May, 1970

ABSTRACT

ROUNDING ERRORS IN THE SOLUTION OF
THE ONE DIMENSIONAL HEAT EQUATION
USING A GALERKIN TECHNIQUE

Catherine Gardner

In solving the one dimensional heat equation, the solution is approximated using a Galerkin technique. Then, if the approximate solution U is required to lie in a Hermite space of piecewise polynomials of degree 3 based on a rectangular grid of mesh size h and if $\Delta t=c_1 h$ or $\Delta t=c_2 h^2$ where c_1, c_2 are constants and if as $\Delta t,h\to 0$ $\frac{\nu}{\Delta t h}$ is kept constant then, in either case, it can be shown that the computed solution $\hat{U}$ satisfies $||\hat{U}-U||_{L^2} \leq$ a constant. The solution is computed using floating point base N-arithmetic with a τ digit mantissa and $\nu=N^{1-\tau}$.

ACKNOWLEDGMENTS

Of the many people who deserve my gratitude, I would especially like to thank Professor Henry H. Rachford, Jr., my thesis director, whose understanding guidance encouraged me in this undertaking. I am also indebted to Professors Graeme Fairweather and Howard L. Resnikoff, who served on my thesis committee, and to Mrs. Shirley Payne, who patiently typed the very difficult manuscript. I would also like to thank the United States Air Force and Rice University whose generous support in the form of an AFOSR grant and a Rice Fellowship enabled me to continue my education on the graduate level.

ROUNDING ERRORS IN THE SOLUTION OF THE ONE DIMENSIONAL HEAT EQUATION USING A GALERKIN TECHNIQUE

Introduction

We shall be concerned with solutions of the following boundary value problem

$$u_{xx} = u_t \qquad \text{on } (0,1) \times (0,T] \tag{1a}$$

$$u(x,0) = u_0(x) \tag{1b}$$

$$u(1,t) = u(0,t) = 0 \tag{1c}$$

and we shall approximate the solution of (1) by a function

$$U(x,t) = \sum_{\ell=1}^{N} \alpha_\ell(t)\ v_\ell(x)$$

where $\{v_\ell(x)\}$ is a suitable collection of functions such that $v_\ell(x) \in C_0^1(0,1)$ and the collection $\{v_\ell(x)\}$ is linearly independent. If M is the span of the $\{v_\ell(x)\}$ then U is defined by

$$\langle U_{xx}, V\rangle = \langle U_t, V\rangle \qquad \text{for all } V \in M \tag{2a}$$

$$\langle U_0, V\rangle = \langle u_0, V\rangle \qquad t=0 \quad V \in M \tag{2b}$$

where $\langle W,V\rangle = \int_0^1 WV\,dx$

It can be shown [1] that under reasonable hypotheses the approximation U(x,t) and its first order space derivatives tend to the solution of (1a) with the error bounded by a

multiple of $h^{2m-1} + \Delta t^2$ where the approximate solution is required to lie in a Hermite space of piecewise polynomials of degree 2m-1 based on a rectangular grid of mesh size h.

The system is actually an initial value problem for the set of linear ordinary differential equations

$$A\ \alpha'(t) + B\ \alpha(t) = 0 \tag{3a}$$

$$A\ \alpha(0) = \left(\langle u_0 v_1\rangle, \ldots\ldots \langle u_0 v_N\rangle\right)^T \tag{3b}$$

where $\alpha = (\alpha_1, \ldots\ldots \alpha_N)^T$

and A and B are the positive definite matrices given by

$$A = a_{ij} = \langle v_i, v_j\rangle$$

$$B = b_{ij} = \langle v_{ix}, v_{jx}\rangle$$

We shall analyze the following difference scheme

$$A\ \frac{\left(\alpha_{n+1} - \alpha_n\right)}{\Delta t} = -B\ \frac{\left(\alpha_{n+1} + \alpha_n\right)}{2} \tag{4a}$$

$$\alpha_{n+1} = \left(A + \frac{\Delta t}{2} B\right)^{-1} \left(A - \frac{\Delta t}{2} B\right)\ \alpha_n \tag{4b}$$

The direct solution of the linear algebraic system (4) can be carried out by factoring $\left(A + \frac{\Delta t}{2} B\right) = LU$ where L is unit lower triangular and U is upper triangular then solving

$$L\delta = \left(A - \frac{\Delta t}{2} B\right) \alpha_n$$

$$U\alpha_{n+1} = \delta.$$

We shall be interested in the bound on the norm of the rounding error when the above scheme is carried out in floating point computation. Throughout we shall use a symbol with a caret, e.g. $\hat{U}_n$ to denote the computed value of the exact element U_n. Thus, we seek a bound on $||\hat{U}_n - U_n||_{L^2}$

where $$U_n = \sum_{\ell=1}^{N} \alpha_\ell(t_n)\ v_\ell(x)$$

and $$\hat{U}_n = \sum_{\ell=1}^{N} \hat{\alpha}_\ell(t_n)\ v_\ell(x)$$

Then $$||\hat{U}_n - U_n||_{L^2}^2 = \int_0^1 \left(\sum_\ell \hat{\alpha}_\ell\ (t_n)\ v_\ell(x) - \sum \alpha_\ell(t_n)\ v_\ell(x) \right)^2 dx$$

$$= \int_0^1 \left(\sum (\hat{\alpha}_\ell(t_n) - \alpha_\ell(t_n)) v_\ell(x) \right)^2 dx$$

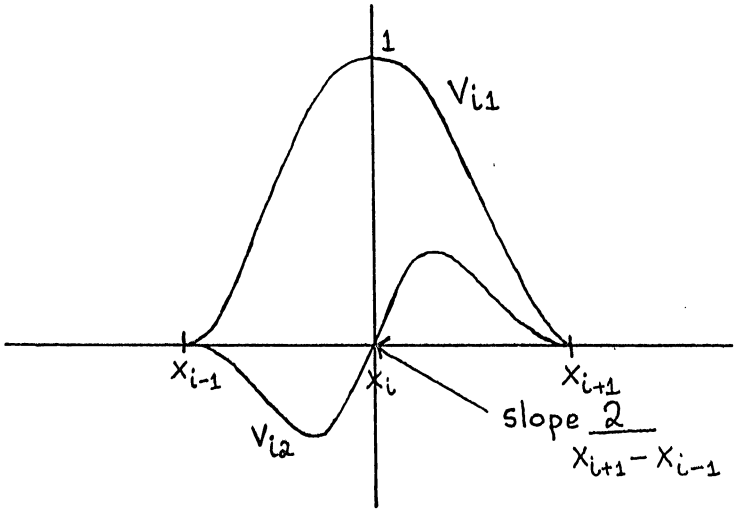
$$= (\hat{\alpha}_n - \alpha_n)^T\ A\ (\hat{\alpha}_n - \alpha_n)$$

$$= ||\ \hat{\alpha}_n - \alpha_n\ ||_A^2.$$

And thus the bound we seek is on the A norm of $\hat{\alpha}_n - \alpha_n$. We shall be interested in two cases, i.e. that where we assume $\Delta t = ch$ and that where $\Delta t = ch^2$ where c is a constant, and will compare the results so obtained with those obtained in [2].

Choice of Basis Functions and Calculation of Entries for the Matrices A & B

For our basis functions $\{v_\ell(x)\}$ we shall choose Hermite cubics of the following form: ——



And if we order the basis functions as follows

$$v_{02}\ v_{11}\ v_{12}\ v_{21}\ v_{22}\ \cdots\cdots\cdots\ v_{\frac{N}{2}-1,1}\ v_{\frac{N}{2}-1,2}\ v_{\frac{N}{2},2}$$

then the boundary conditions (1c) are satisfied and the matrix A is an N×N matrix with the following entries.

$$A = a_{jk} = \begin{cases} \left\langle v_{\frac{j-1}{2},2}\ ,\ v_{\frac{k-1}{2},2} \right\rangle & j, k \text{ odd} \\ \left\langle v_{\frac{j}{2},1}\ ,\ v_{\frac{k}{2},1} \right\rangle & j, k \text{ even} \\ \left\langle v_{\frac{j-1}{2},2}\ ,\ v_{\frac{k}{2},1} \right\rangle & j \text{ odd, } k \text{ even} \\ \left\langle v_{\frac{j}{2},1}\ ,\ v_{\frac{k-1}{2},2} \right\rangle & j \text{ even, } k \text{ odd} \end{cases}$$

v_{i1} and v_{i2} are given by the equations

$$v_{i1} = -2\frac{(x-x_{i-1})^3}{(x_i-x_{i-1})^3} + 3\frac{(x-x_{i-1})^2}{(x_i-x_{i-1})^2} \qquad x\varepsilon\left[x_{i-1},x_i\right]$$

$$= -2\frac{(x_{i+1}-x)^3}{(x_{i+1}-x_i)^3} + 3\frac{(x_{i+1}-x)^2}{(x_{i+1}-x_i)^2} \qquad x\varepsilon\left[x_i,x_{i+1}\right]$$

$$v_{i2} = \frac{2}{x_{i+1}-x_{i-1}}\left[\frac{(x-x_{i-1})^3}{(x_i-x_{i-1})^2} - \frac{(x-x_{i-1})^2}{(x_i-x_{i-1})}\right] \qquad x\varepsilon\left[x_{i-1},x_i\right]$$

$$= \frac{-2}{x_{i+1}-x_{i-1}}\left[\frac{(x_{i+1}-x)^3}{(x_{i+1}-x_i)^2} - \frac{(x_{i+1}-x)^2}{(x_{i+1}-x_i)}\right] \qquad x\varepsilon\left[x_i,x_{i+1}\right]$$

Since the support of v_{i1}, v_{i2} is contained in $[x_{i-1},x_{i+1}]$ we note immediately that $\langle v_{ik},v_{jk}\rangle = 0 \quad |i-j|>1 \quad k = 1,2.$

And that $\langle v_{ik},v_{jk}\rangle = \langle v_{jk},v_{ik}\rangle \quad k = 1,2.$

If we further assume that $x_{i+1}-x_i = h$ $\forall i$ then since v_{i1} is even about x_i and v_{i2} is odd about x_i then $\langle v_{i1},v_{i2}\rangle = 0$ and $\langle v_{i1},v_{j2}\rangle = -\langle v_{i2},v_{j1}\rangle$.

And the matrix A has the simplified form

$$A = \begin{bmatrix}
\frac{b}{2} & -e & g & & & & & & & & \\
-e & a & 0 & f & e & & & & & & \\
g & 0 & b & -e & g & & & & & & \\
 & f & -e & a & 0 & f & e & & & & \\
 & e & g & 0 & b & -e & g & & & & \\
 & & & f & -e & a & 0 & f & e & & \\
 & & & e & g & 0 & b & -e & g & & \\
 & & & & & & & \cdot & \cdot & \cdot & \\
 & & & & & & & & \cdot & \cdot & \cdot \\
 & & & & & & & f & -e & a & 0 & +e \\
 & & & & & & & e & g & 0 & b & g \\
 & & & & & & & & & +e & g & \frac{b}{2}
\end{bmatrix}$$

Where $a = \langle v_{11}, v_{11} \rangle$

$b = \langle v_{12}, v_{12} \rangle$

$e = \langle v_{11}, v_{22} \rangle$

$f = \langle v_{11}, v_{21} \rangle$

$g = \langle v_{12}, v_{22} \rangle$

and calculation gives the following values for entries in A.

$$a = \frac{26}{35} h \qquad b = \frac{2}{105} h \qquad e = \frac{-13}{420} h$$

$$f = \frac{9}{70} h \qquad g = \frac{-1}{140} h .$$

An exactly analogous argument shows that the matrix B has the form

$$B = \begin{bmatrix}
\frac{b'}{2} & -e' & g' & & & & & & & \\
-e' & a' & 0 & f' & e' & & & & & \\
g' & 0 & b' & -e' & g' & & & & & \\
 & f' & -e' & a' & 0 & f' & e' & & & \\
 & e' & g' & 0 & b' & -e' & g' & & & \\
 & & & & & \cdot & & \cdot & & \\
 & & & & & & \cdot & & \cdot & \\
 & & & & & & & f' & -e' & a' & 0 & e' \\
 & & & & & & & e' & g' & 0 & b' & g' \\
 & & & & & & & & & e' & g' & \frac{b'}{2}
\end{bmatrix}$$

Where $a' = \langle v_{11x'}, v_{11x} \rangle = \frac{12}{5h}$

$b' = \langle v_{12x'}, v_{12x} \rangle = \frac{4}{5h}$

$e' = \langle v_{11x'}, v_{22x} \rangle = \frac{1}{10h}$

$f' = \langle v_{11x'}, v_{21x} \rangle = \frac{-6}{5h}$

$g' = \langle v_{12x'}, v_{22x} \rangle = \frac{-1}{30h}$

Thus $||A||_2 \leq \max_i \sum_j |a_{ij}| = |a| + 2|f| + 2|e|$

$\leq \frac{223}{210} h.$

$||A + \frac{\Delta t}{2} B||_2 \leq |a + \frac{\Delta t}{2} a'| + 2|f + \frac{\Delta t}{2} f'| + 2|e + \frac{\Delta t}{2} e'|$

$\leq \frac{5\Delta t}{2h} + \frac{223}{210} h.$

We also need to calculate $\| |A - \frac{\Delta t}{2} B| \|_2 \leq |a - \frac{\Delta t}{2} a'|$

$$+ 2|f - \frac{\Delta t}{2} f'| + 2|e - \frac{\Delta t}{2} e'|$$

$$\leq \frac{5\Delta t}{2h} + \frac{223}{210} h.$$

Calculation of a Bound for $|| \hat{\alpha}_n - \alpha_n ||_A$.

Recall that we are trying to solve equation (4b)

$$\left(A + \frac{\Delta t}{2} B\right) \alpha_{n+1} = \left(A - \frac{\Delta t}{2} B\right) \alpha_n$$

if $\Delta t = k$ and $\left(A + \frac{k}{2} B\right)^{-1} = Q$, $\left(A - \frac{k}{2} B\right) = M$

then (4b) reduces to $\alpha_{n+1} = QM\, \alpha_n$.

We seek a uniform bound for $||\hat{\alpha}_n - \alpha_n||_A$ where $\{\hat{\alpha}_n\}$ is the computed sequence.

The recursion for $\{\hat{\alpha}_n\}$ is given by

$$\hat{\alpha}_{n+1} = \hat{Q}\, \hat{d}_n \text{ where } \hat{d}_n = f\ell\ (M\hat{\alpha}_n) \text{ and } \hat{Q} = (1+R)\ Q$$

for some R which depends on the method of solving $Q^{-1}\alpha = \hat{d}$.

Now $\hat{d}_n = f\ell\ (M\hat{\alpha}_n) = M\hat{\alpha}_n + e_n$

$$\hat{\alpha}_{n+1} = (1+R)\ Q\ (M\hat{\alpha}_n + e_n)$$

$$\alpha_{n+1} = QM\alpha_n$$

Let $w_n = \hat{\alpha}_n - \alpha_n$ and by subtracting the recursions for $\hat{\alpha}_n$ and α_n we obtain

$$\hat{\alpha}_{n+1} - \alpha_{n+1} = (1+R)\ Q\ (M\hat{\alpha}_n + e_n) - QM\alpha_n$$

and simplifying we obtain

$$w_{n+1} = R\alpha_{n+1} + (1+R)\ Q\ (Mw_n + e_n) \qquad (5)$$

In using $f\ell(\cdot)$ we have extended Wilkinson's notation [4] to include matrix and vector operations. Briefly, if a and b are scalars $f\ell(a \square b)$ means the floating point realization of the operation $\square$ between them, where $\square$ is +, -, ×, or ÷ . The relevant facts are $f\ell(a \square b) = (a \square b)\ (1+\mu)$ for some μ satisfying $|\mu| \leq \nu = N^{1-\tau}$, where τ is the number of floating base N-digits in the mantissa, and

$$f\ell\left(\sum_{i=1}^{p} a_i b_i\right) = \sum_{i=1}^{p} a_i b_i\ (1+\mu_i)$$

where $|\mu_i| \leq 1.06p\nu$.

We also assume that we can ignore any terms of order ν^2.

Now $|e_n| = 1{\cdot}06 \times 6\nu \times |M||\hat{\alpha}_n|$

$$\leq 6.36\ \nu|M|\ |w_n + \alpha_n|$$

since at most six products suffice for each element of $M\hat{\alpha}_n$.

$\therefore\quad e_n = 6.36\ \nu\ \check{M}\ (w_n + \alpha_n)$ where $\check{m}_{ij} \leq \theta_{ij}\ m_{ij}$.

$$|\theta_{ij}| \leq 1$$

$$\text{and } ||\check{M}||_2 \leq ||\ |M|\ ||_2$$

Therefore substituting into (5) we have

$$w_{n+1} = R\alpha_{n+1} + (1+R)\ Q\ (Mw_n + 6.36\nu\ \check{M}\ (w_n + \alpha_n))$$

$$= (1+R)\ (QM + 6.36\ \nu\ Q\check{M})w_n + R\alpha_{n+1}$$

$$+ (1+R)\ Q\ 6.36\nu\ \check{M}\alpha_n$$

Thus

$$||w_{n+1}||_A \leq (1+||R||_A)(||QM||_A+6\cdot36\nu||\check{QM}||_A)||w_n||_A \qquad (6)$$

$$+ ||R||_A\,||\alpha_{n+1}||_A + (1+||R||_A)\,6\cdot36\nu||\check{QM}||_A\,||\alpha_n||_A$$

We shall now show that $||QM||_A \leq 1$ and that $||\alpha_n||_A \leq ||\alpha_o||_A \leq$ a constant. Now $||QM||_A = ||A^{\frac{1}{2}}QMA^{-\frac{1}{2}}||_2$ and if we let $\frac{k}{2}B = C$ then

$$A^{\frac{1}{2}}QMA^{-\frac{1}{2}} = A^{\frac{1}{2}}\,(1+A^{-1}C)^{-1}\,(1-A^{-1}C)\,A^{-\frac{1}{2}}$$

$$= A^{\frac{1}{2}}\,(1+A^{-1}C)^{-1}A^{-\frac{1}{2}}A^{\frac{1}{2}}\,(1-A^{-1}C)\,A^{-\frac{1}{2}}$$

$$= (A^{\frac{1}{2}}\,(1+A^{-1}C)A^{-\frac{1}{2}})^{-1}\,(A^{\frac{1}{2}}(1-A^{-1}C)A^{-\frac{1}{2}})$$

$$= (1+A^{-\frac{1}{2}}CA^{-\frac{1}{2}})^{-1}\,(1-A^{-\frac{1}{2}}CA^{-\frac{1}{2}})$$

Thus $A^{\frac{1}{2}}QMA^{-\frac{1}{2}}$ is symmetric and therefore if λ_i is an eigenvalue of $A^{-\frac{1}{2}}CA^{-\frac{1}{2}}$ then since $A^{-\frac{1}{2}}CA^{-\frac{1}{2}}$ is similar to $A^{-1}C$, λ_i is both real and positive.

$$\therefore \qquad ||A^{\frac{1}{2}}QMA^{-\frac{1}{2}}||_2 = \frac{1-\lambda_i}{1+\lambda_i} \leq 1$$

$$||QM||_A \leq 1$$

Now we note that if in equation (3a), $V = U_{n+1} + U_n$ and the same difference scheme is used, then

$$-\int_0^1 \frac{\partial}{\partial x}\left[\left(\frac{U_{n+1}+U_n}{2}\right)\right]^2 = \frac{1}{\Delta t}\int_0^1 \left(U_{n+1}-U_n\right)\left(U_{n+1}+U_n\right)$$

or

$$\left|\left|U_{n+1}\right|\right|^2_{L^2} - \left|\left|U_n\right|\right|^2_{L^2} + \frac{\Delta t}{2}\left|\left| U_{n+1} + U_n\right|\right|^2_{H_0^1} = 0$$

thus

$$\left|\left|U_{n+1}\right|\right|^2_{L^2} \leq \left|\left|U_n\right|\right|^2_{L^2}$$

Now

$$U_{n+1} = \sum_{\ell=1}^{N} \alpha_\ell\left(t_{n+1}\right)\, v_\ell(x)$$

$$\left|\left|U_{n+1}\right|\right|_{L^2} = \int_0^1 \left(\sum_\ell \alpha_\ell\left(t_{n+1}\right)\, v_\ell(x)\right)^2$$

$$= \alpha^T_{n+1}\, A\alpha_{n+1}$$

and similarily

$$\left|\left|U_{n+1} + U_n\right|\right|^2_{H_0^1} = \left(\alpha_{n+1} + \alpha_n\right)^T B\left(\alpha_{n+1} + \alpha_n\right)$$

$$\left|\left|\alpha_{n+1}\right|\right|^2_A - \left|\left|\alpha_n\right|\right|^2_A + \frac{\Delta t}{2}\left|\left|\alpha_{n+1} + \alpha_n\right|\right|^2_B = 0$$

$$\left|\left|\alpha_{n+1}\right|\right|_A \leq \left|\left|\alpha_n\right|\right|_A$$

and thus $\left|\left|\alpha_n\right|\right|_A \leq \left|\left|\alpha_o\right|\right|_A \qquad \forall_n$

Now utilizing equation (2b) we can write

$$\left\langle U_o, U_o\right\rangle = \left\langle u_o, U_o\right\rangle$$

$$||U_o||^2_{L_2} = \int_0^1 \left(\sum_\ell \alpha_\ell(t_o) v_\ell(x) \right)^2 = \left(\alpha_o^T A \alpha_o \right) = ||\alpha_o||^2_A$$

$$= \langle u_o, U_o \rangle \leq ||u_o||_{L^2} \; ||U_o||_{L^2}$$

$$||\alpha_o||_A \leq ||u_o||_{L^2} = \text{a constant which we shall}$$

denote β.

Thus, if P is a bound for $||R||_A$ then substituting into equation (6)

$$||w_{n+1}||_A \leq (1+P)\ (1+6\cdot 36\nu\ ||\check{QM}||_A)\ ||w_{n+1}||_A$$

$$+ P\beta + (1+P)\ 6\cdot 36\nu\ ||\check{QM}||_A \beta$$

$$\leq ||w_n||_A\ \phi_1 + \beta\phi_2$$

where $\phi_1 = (1+P)\ (1+6\cdot 36\nu\ ||\check{QM}||_A$

$$\phi_2 = P + (1+P)\ 6\cdot 36\nu\ ||\check{QM}||_A$$

$$\therefore\ ||w_{n+1}||_A \leq \beta\ \phi_2\ \frac{\left(\phi_1^n - 1\right)}{\left(\phi_1 - 1\right)}$$

Since $||\check{QM}||_A = ||A^{\frac{1}{2}} \check{QM} A^{-\frac{1}{2}}||_2$

$$\leq ||A^{\frac{1}{2}}||_2\ ||A^{-\frac{1}{2}}||_2\ ||Q||_2\ ||\check{M}||_2$$

$$\leq ||A^{\frac{1}{2}}||_2\ ||A^{-\frac{1}{2}}||_2\ ||Q||_2\ ||\ |M|\ ||_2$$

it merely remains to obtain bounds on $||A^{-\frac{1}{2}}||_2$, $||Q||_2$ and the bound P on $||R||_A$

Calculation of $||Q||_2$ and $||A^{-1}||_2$

Recall that $Q = \left(A + \frac{k}{2} B\right)^{-1}$

A^{-1} and Q^{-1} are positive definite and therefore

$$||Q||_2 = \frac{1}{\lambda_{\min} Q^{-1}} \quad \text{and} \quad ||A^{-1}||_2 = \frac{1}{\lambda_{\min} A}$$

Consider first the matrix A which has the form discussed earlier. We shall show that if λ_1 is the minimum eigenvalue of A then

$$\lambda_1 \geq \frac{h}{5x420}.$$

We shall define a series of matrices

$$A^{(i)} = a_{jk}^{(i)} = \begin{cases} \left\langle v_{\frac{j-1}{2},2}, v_{\frac{k-1}{2},2} \right\rangle^{(i)} & j,k \text{ odd} \\ \left\langle v_{\frac{j}{2},1}, v_{\frac{k}{2},1} \right\rangle^{(i)} & j,k \text{ even} \\ \left\langle v_{\frac{j-1}{2},2}, v_{\frac{k}{2},1} \right\rangle^{(i)} & j \text{ odd, } k \text{ even} \\ \left\langle v_{\frac{j}{2},1}, v_{\frac{k-1}{2},2} \right\rangle^{(i)} & j \text{ even, } k \text{ odd} \end{cases}$$

Where $\left\langle v_{eg}, v_{mg} \right\rangle^{(i)} = \int_{(i-1)h}^{ih} v_{eg}, v_{mg}.$

Then clearly $\sum_{i=1}^{N/2} A^{(i)} = A.$

Now for $i \neq 1, N/2$ each $A^{(i)}$ is an NxN matrix with non-zero entries only in the (i,i+3 : i,i+3) principal submatrix,

and for v_i the entries in the non-zero principal submatrix are identical. Thus, we can denote by D the (i,i+3 : i,i+3) principal submatrix of each $A^{(i)}$. It is clear that D is positive definite, since it is a grammian and the basis functions are linearly independent over [(i-1)h,ih]. For boundary conditions satisfying (1c), $A^{(1)}$ is an N×N matrix with non-zero entries only in the (1,3 : 1,3) principal submatrix, we shall denote this principal submatrix D', and note that it is positive definite and also a principal submatrix of D. $A^{N/2}$ is also an NxN matrix with non-zero entries only in the $\left(\frac{N}{2} - 2, \frac{N}{2} : \frac{N}{2} - 2, \frac{N}{2}\right)$ principal submatrix, we shall denote this principal submatrix D" and note that it also is positive definite and a principal submatrix of D.

Now $$x^T A x = x^T \left(\sum_{i=1}^{N/2} A^{(i)} \right) x$$

and if we choose x to be the eigenvector corresponding to the minimum eigenvalue of A and partition x so that

$$x^T = \left(x_1, x_2, \ldots . x_{\frac{N}{2}-1}, x_{\frac{N}{2}} \right)$$

where x_i, $i \neq 1, \frac{N}{2}$ is the vector such that $x_i^T = \left(x_i^{(1)} x_i^{(2)} \right)$ and x_1, $x_{\frac{N}{2}}$ are vectors consisting of a single element. Then provided $x^T x = 1$

$$\lambda_{\min A} = (Ax,x) = \left(D' \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) + \left(D \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} \right) + \ldots$$

$$+ \left(D'' \begin{bmatrix} X_{\frac{N}{2}-1} \\ X_{\frac{N}{2}} \end{bmatrix}, \begin{bmatrix} X_{\frac{N}{2}-1} \\ X_{\frac{N}{2}} \end{bmatrix} \right)$$

$$\geq \lambda_{\min} D' \left|\left| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right|\right|^2 + \lambda_{\min} D \left|\left| \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} \right|\right|^2 + \cdots$$

$$\cdots + \lambda_{\min} D'' \left|\left| \begin{bmatrix} x_{\frac{N}{2}-1} \\ x_{\frac{N}{2}} \end{bmatrix} \right|\right|^2$$

but $\lambda_{\min} D' \geq \lambda_{\min} D$ and $\lambda_{\min} D'' \geq \lambda_{\min} D$ since both are principal submatrices of D.

$$\lambda_{\min} A \geq \lambda_{\min} D \left(\left|\left| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right|\right|^2 + \left|\left| \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} \right|\right|^2 \cdots + \left|\left| \begin{bmatrix} x_{\frac{N}{2}-1} \\ x_{\frac{N}{2}} \end{bmatrix} \right|\right|^2 \right)$$

$$\geq \lambda_{\min} D \, ||x||^2 + \lambda_{\min} D \left(x_2^2 + x_3^2 \cdots + x_{\frac{N}{2}-1}^2 \right)$$

$$\geq \lambda_{\min} D$$

Now D is the following 4×4 matrix

$$D = \begin{bmatrix} \frac{a}{2} & -d & f & +e \\ -d & \frac{b}{2} & -e & g \\ f & -e & \frac{a}{2} & +d \\ +e & g & d & \frac{b}{2} \end{bmatrix}$$

where a, f, e, b, g have the values given previously and

$$d = \langle v_{11}, v_{12} \rangle^1 = \int_0^h v_{11}\, v_{12} = -\frac{11h}{210}$$

$$\therefore \quad D = \frac{h}{420} \begin{bmatrix} 156 & 22 & 54 & -13 \\ 22 & 4 & 13 & -3 \\ 54 & 13 & 156 & -22 \\ -13 & -3 & -22 & 4 \end{bmatrix}$$

Calculation shows that the characteristic polynomial of 420D is

$$\lambda^4 - 320\lambda^3 h + 22617\lambda^2 h^2 - 29792\lambda h^3 + 5145h^4$$

which, by calculations based on the change in sign of the characteristic polynomial gives $\lambda_{\min D} \geq \frac{h}{5\times 420}$

and $\lambda_{\min A} \geq \frac{h}{5\times 420}$. Therefore, $\left|\left|A^{-1}\right|\right|_2 \leq \frac{2100}{h}$

We also seek a bound for the minimum eigenvalue of $A + \frac{k}{2}B$. We could look at

$$A + \frac{k}{2}B = \sum_{i=1}^{N/2} A^{(i)} + \frac{k}{2} \sum_{i=1}^{N/2} B^{(i)}$$

However the non-zero principal submatrix of $B^{(i)}$ $i \neq 1, N/2$, which we shall denote E is singular since the basis functions when differentiated are linearly dependent over $[(i-1)h, ih]$. Thus the bound we would obtain in this fashion is no better than that we obtained for A.

We can, however, look at

$$A + \frac{k}{2} B = \sum_{i=1}^{N/2} \left(A^{(i)} + \frac{k}{2} B^{(i)} \right)$$

where D+E is the following 4×4 matrix

$$D+E = \begin{bmatrix} \frac{a}{2} + \frac{ka'}{4} & d + \frac{k}{2} d' & f + \frac{k}{2} f' & e + \frac{k}{2} e' \\ d + \frac{k}{2} d' & \frac{b}{2} + \frac{kb'}{4} & -\left(e + \frac{k}{2} e'\right) & g + \frac{k}{2} g' \\ f + \frac{k}{2} f' & -\left(e + \frac{k}{2} e'\right) & \frac{a}{2} + \frac{ka'}{4} & -\left(d + \frac{k}{2} d'\right) \\ e + \frac{k}{2} e' & g + \frac{k}{2} g' & -\left(d + \frac{k}{2} d'\right) & \frac{b}{2} + \frac{kb'}{4} \end{bmatrix}$$

where a, a', f, f', e, e', b, b', g, g' and d have the same values as before and d' = -e'. Calculation of the characteristic polynomial of 420 (D+E) gives

$$\lambda^4-\lambda^3\left(320h+560\frac{k}{h}\right)+\lambda^2\left(22,617h^2+126,770k+27,195\frac{k^2}{h^2}\right)$$

$$-\lambda\left(29,792h^2+1,296,540kh+5,938,840\frac{k^2}{h}+308,700\frac{k^3}{h^3}\right)$$

$$+5145h^4+143,724kh^2+16,053,200k^2+61,593,058\frac{k^3}{h^2}$$

Now if $k=c_1 h$ then provided $h<1$ and $c_1 \geq 1$

then $\lambda_{min(D+E)} \geq \frac{198}{420} h$

$\lambda_{min(A+\frac{k}{2}B)} \geq \frac{198}{420} h$ and $||Q||_2 \leq \frac{420}{198h}$

But if $k = c_2 h^2$ then, for $1 \leq c_2 < 10$, $A + \frac{k}{2} B$ is a strictly diagonally dominant symmetric matrix and a bound for the minimum eigenvalue is given by Gesgorin.

Minimum eigenvalue $\left(A + \frac{k}{2} B\right) \geq \frac{28h}{420}$

$\therefore \quad ||Q||_2 \leq \frac{420}{28h}$.

Calculation of P

Recall that P is a bound for $||R||_A$

where $\hat{Q} = (1+R)Q$ $\qquad$ $Q = (A+\frac{k}{2}B)^{-1}$

Now suppose Gw = d and invert G by Gaussian elimination. The Gaussian algorithm expresses the set of equations Gw = d in the form

$$LUw = d \tag{8}$$

where L is unit lower triangular and U is upper triangular. The computed $L(\hat{L})$ and $U(\hat{U})$ satisfy

$$\hat{L}\hat{U} = G + E$$

hence if we solve equation (8) without further rounding error we obtain the solution of the system

$$(G+E)w = d.$$

Now equation (8) can be solved in two stages

$$\hat{L}y = d$$

$$\hat{U}w = y$$

and in practice we obtain

$$(\hat{L} + d\hat{L})\hat{y} = d$$

$$(\hat{U} + d\hat{U})\hat{w} = \hat{y}$$

or $(G + E + \hat{L}d\hat{U} + d\hat{L}\hat{U} + d\hat{L}d\hat{U})\hat{w} = d$

or $(G + K)\hat{w} = d$

where $K = E + d\hat{L}\hat{U} + \hat{L}d\hat{U} + d\hat{L}d\hat{U}$

and $||K||_\infty \leq ||E||_\infty + ||d\hat{L}\hat{U}||_\infty + ||\hat{L}d\hat{U}||_\infty + ||d\hat{L}d\hat{U}||_\infty$

Now $\hat{w} - w = \left((G+K)^{-1} - G^{-1}\right)d$

$= \left((G+K)^{-1}G - I\right)w$

but $G + K = G(I + G^{-1}K)$

and $(G+K)^{-1} = (I + G^{-1}K)^{-1}\, G^{-1}$

$\left((G+K)^{-1}G-I\right) = \left(I+G^{-1}K\right)^{-1} - I$

and $\left(I+G^{-1}K\right)$ is non-singular if $||G^{-1}K||_2 \leq 1.$

let $G^{-1}K = F$

then $$\frac{||\hat{w}-w||_2}{||w||_2} = ||(I+F)^{-1} - I||_2 \leq \frac{||F||_2}{1-||F||_2}$$

$$\leq \frac{||G^{-1}||_2||K||_2}{1-||G^{-1}||_2||K||_2}$$

but $R = \hat{Q}Q^{-1} - I$

substituting $Q^{-1} = G$ and $\hat{Q} = (G+K)^{-1}$

then $R = (G+K)^{-1}G - I$

$$\therefore \quad \frac{||\hat{w}-w||_2}{||w||_2} = ||R||_2 \leq \frac{||Q||_2 \, ||K||_2}{1-||Q||_2||K||_2}$$

now $\quad ||K||_2 \leq \sqrt{||K||_\infty \, ||K||_1}$

and $\quad ||R||_2 \leq \dfrac{||Q||_2\sqrt{||K||_\infty||K||_1}}{1-||Q||_2\sqrt{||K||_\infty||K||_1}}$

provided $\quad ||Q||_2\sqrt{||K||_\infty||K||_1} < 1$

$$||R||_A \leq ||A^{-\frac{1}{2}}||_2 \, ||R||_2 \, ||A^{\frac{1}{2}}||_2$$

$$\leq \frac{||A^{-\frac{1}{2}}||_2 \, ||Q||_2\sqrt{||K||_\infty||K||_1} \, ||A^{\frac{1}{2}}||_2}{1 - ||Q||_2\sqrt{||K||_\infty||K||_1}} \qquad (9)$$

Error Analysis of Gaussian Elimination for a Positive Definite Matrix $A = a_{ij}$ where

for $i \leq j$	$a_{ij} = 0$	i, even	$j-i \geq 4$
		i, odd	$j-i \geq 3$.
$i > j$	$a_{ij} = 0$	i, even	$i-j \geq 3$
		i, odd	$i-j \geq 4$

The solution of a set of equations and the inversion of a matrix by Gaussian elimination is based on the triangularization of the matrix. If we denote the original set of equations by $A^{(1)}x = b$

then N-1 equivalent sets

$$A^{(r)}x = b^{(r)} \qquad (r = 2 \ . \ . \ . \ N)$$

are produced the matrix $A^{(N)}$ of the final set being of upper triangular form. In general $A^{(r)}$ is already of triangular form as regards its first r rows and it has a square matrix (which for convenience we shall denote $B^{(r)}$) of non-zero elements in the bottom right-hand corner. This square matrix is of order N+1-r. The matrix $A^{(r+1)}$ is derived from $A^{(r)}$ by subtracting a multiple m_{ir} of the r^{th} row from the i^{th} row for values of i from r+1 to N. The multipliers m_{ir} are defined by

$$m_{ir} = \frac{a_{ir}^{(r)}}{a_{rr}^{(r)}}$$

Then we have

$$\hat{L}\hat{U} = A^{(1)} + E$$

where

$$\hat{L} = \begin{bmatrix} 1 & & & & & \\ \hat{m}_{21} & 1 & & & & \\ \hat{m}_{31} & \cdot & \cdot & & & \\ \cdot & & \cdot & \cdot & & \\ \cdot & & & \cdot & \cdot & \\ \cdot & & & & \cdot & \\ \hat{m}_{N1} & \cdot & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

and

$$\hat{U} = \begin{bmatrix} \hat{a}_{11}^{(1)} & \hat{a}_{12}^{(1)} & \cdot & \cdot & \cdot & \hat{a}_{1N}^{(1)} \\ & \hat{a}_{22} & \cdot & & & \cdot \\ & & \cdot & \cdot & & \cdot \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \\ & & & & & a_{NN}^{(N)} \end{bmatrix}$$

Wilkinson [4] shows that when a positive definite matrix is reduced by Gaussian elimination then no element in any $B^{(r)}$ exceeds the maximum element in the original matrix, and that each $B^{(r)}$ is also positive definite. We shall show that $\hat{B}^{(r)}$ is positive definite provided $1/||B^{(r)-1}||_2 > 5\nu$ and $\hat{a}_{ij}^{(r)} \leq 1$ $\forall r$ provided $a_{ij}^{(1)} \leq (1-3\nu)$.

If we denote the elements of E by e_{ij} and note that E always has non-zero entries where A has non-zero entries, then the history of the (i,j) element differs according as $i \leq j$ or $i > j$.

i) $i \leq j$.

The element is modified in each transformation until $A^{(i)}$ is obtained after which it remains constant. For r even

$$\hat{a}^{(r)}_{r,r} = a^{(1)}_{r,r} - \hat{m}_{r,r-1}\,\hat{a}^{(r-1)}_{r-1,r} - \hat{m}_{r,r-2}\,\hat{a}^{(r-2)}_{r-2,r} + e_{r,r}$$

$$\hat{a}^{(r)}_{r,r+1} = a^{(1)}_{r,r+1} - \hat{m}_{r,r-1}\,\hat{a}^{(r-1)}_{r-1,r+1} - \hat{m}_{r,r-2}\,\hat{a}^{(r-2)}_{r-2,r+1} + e_{r,r+1}$$

$$\hat{a}^{(r)}_{r,r+2} = a^{(1)}_{r,r+2} + e_{r,r+2}$$

$$\hat{a}^{(r)}_{r,r+3} = a^{(1)}_{r,r+3} + e_{r,r+3}$$

For r odd

$$\hat{a}^{(r)}_{r,r} = a^{(1)}_{r,r} - \hat{m}_{r,r-3}\,\hat{a}^{(r-3)}_{r-3,r} - \hat{m}_{r,r-2}\,\hat{a}^{(r-2)}_{r-2,r} - \hat{m}_{r,r-1}\,\hat{a}^{(r-1)}_{r-1,r} + e_{rr}$$

$$\hat{a}^{(r)}_{r,r+1} = a^{(1)}_{r,r+1} - \hat{m}_{r,r-1}\,\hat{a}^{(r-1)}_{r-1,r+1} + e_{r,r+1}$$

$$\hat{a}^{(r)}_{r,r+2} = a^{(1)}_{r,r+2} - \hat{m}_{r,r-1}\,\hat{a}^{(r-1)}_{r-1,r+2} + e_{r,r+2}$$

The computed multiplier $\hat{m}_{ij}$ satisfies

$$\hat{m}_{ij} = \frac{\hat{a}_{ij}^{(j)}}{\hat{a}_{jj}^{(j)}} + \eta_{ij}$$

where η_{ij} is the rounding error in the division. The equations are therefore: ——

For r even

$$e_{r+1,r} = a_{r+1,r}^{(1)} - \hat{m}_{r+1,r-2}\,\hat{a}_{r-2,r}^{(r-2)} - \hat{m}_{r+1,r-1}\,\hat{a}_{r-1,r}^{(r-1)} - \hat{m}_{r+1,r}\,\hat{a}_{r,r}^{(r)}$$

$$e_{r+2,r} = a_{r+2,r}^{(1)} - \hat{m}_{r+2,r}\,\hat{a}_{r,r}^{(r)}$$

$$e_{r+3,r} = a_{r+3,r}^{(1)} - \hat{m}_{r+3,r}\,\hat{a}_{r,r}^{(r)}$$

For r odd

$$e_{r+1,r} = a_{r+1,r}^{(1)} - \hat{m}_{r+1,r-1}\,\hat{a}_{r-1,r}^{(r-1)} - \hat{m}_{r+1,r}\,\hat{a}_{r,r}^{(r)}$$

$$e_{r+2,r} = a_{r+2,r}^{(1)} - \hat{m}_{r+2,r-1}\,\hat{a}_{r-1,r}^{(r-1)} - \hat{m}_{r+2,r}\,\hat{a}_{r,r}^{(r)}$$

Therefore we see that for r even

$$e_{r,r} = \varepsilon_{r,r}^{(r)} + \varepsilon_{r,r}^{(r-1)}$$

$$e_{r,r+1} = \varepsilon_{r,r+1}^{(r)} + \varepsilon_{r,r+1}^{(r-1)}$$

$$e_{r,r+2} = 0$$

$$e_{r,r+3} = 0$$

$$e_{r+1,r} = \varepsilon^{(r-1)}_{r+1,r} + \varepsilon^{(r)}_{r+1,r} + \varepsilon^{(r+1)}_{r+1,r}$$

$$e_{r+2,r} = \varepsilon^{(r+1)}_{r+2,r}$$

$$e_{r+3,r} = \varepsilon^{(r+1)}_{r+3,r}$$

For r odd

$$e_{r,r} = \varepsilon^{(r)}_{r,r} + \varepsilon^{(r-1)}_{r,r} + \varepsilon^{(r-2)}_{r,r}$$

$$e_{r,r+1} = \varepsilon^{(r)}_{r,r+1}$$

$$e_{r,r+2} = \varepsilon^{(r)}_{r,r+2}$$

$$e_{r+1,r} = \varepsilon^{(r)}_{r+1,r} + \varepsilon^{(r+1)}_{r+1,r}$$

$$e_{r+2,r} = \varepsilon^{(r)}_{r+2,r} + \varepsilon^{(r+1)}_{r+2,r}$$

Where $\varepsilon^{(k)}_{ij} = \hat{a}^{(k)}_{ij} - \hat{a}^{(k-1)}_{ij} + \hat{m}^{(k-1)}_{i,k-1}\ \hat{a}^{(k-1)}_{k-1,j}$

$$k = 2 \ldots\ldots i$$

$$\varepsilon^{j+1}_{ij} = a^{(j)}_{jj}\,\eta_{ij}$$

Now in floating point the computed $\hat{a}^{(k)}_{ij}$ is defined by: ——

$$\hat{a}^{(k)}_{ij} = f\ell\left(\hat{a}^{(k-1)}_{ij} - f\ell\left(\hat{m}_{i,k-1}\ \hat{a}^{(k-1)}_{k-1,j}\right)\right)$$

$$= \frac{\left[\hat{a}^{(k-1)}_{ij} - \hat{m}_{i,k-1}\ \hat{a}^{(k-1)}_{k-1,j}\ (1+\varepsilon_1)\right]}{1+\varepsilon_2} \qquad |\varepsilon_1|, |\varepsilon_2| \leq \nu$$

Now $$\varepsilon_{ij}^{(k)} = \hat{a}_{ij}^{(k)} - \left(\hat{a}_{ij}^{(k-1)} - \hat{m}_{i,k-1}\ \hat{a}_{k-1,j}^{(k-1)}\right)$$

$$\hat{a}_{ij}^{(k-1)} = \hat{a}_{ij}^{(k)}\ (1+\varepsilon_2) + \hat{m}_{i,k-1}\ \hat{a}_{k-1,j}^{(k-1)}\ (1+\varepsilon_1)$$

$$\varepsilon_{ij}^{(k)} = \hat{a}_{ij}^{(k)} - \left[\hat{a}_{ij}^{(k)}\ (1+\varepsilon_2) + \hat{m}_{i,k-1}\ \hat{a}_{k-1,j}^{(k-1)}\ (1+\varepsilon_1) - \hat{m}_{i,k-1}\ \hat{a}_{k-1,j}^{(k-1)}\right]$$

$$= -\hat{a}_{ij}^{(k)}\varepsilon_2 + \hat{m}_{i,k-1}\ \hat{a}_{k-1,j}^{(k-1)}\ \varepsilon_1$$

$$|\varepsilon_{ij}^{(k)}| \leq \left|\hat{a}_{ij}^{(k)} + \hat{m}_{i,k-1}\ a_{k-1,j}^{(k-1)}\right|\ \nu$$

Now we show later that $\left|\hat{a}_{ij}^{(k)}\right| \leq 1$ for all k

and now we consider $\hat{m}_{i,k-1}\ \hat{a}_{k-1,j}^{(k-1)}$.

$$m_{i,k-1}\ a_{k-1,j}^{(k-1)} = \frac{\hat{a}_{i,k-1}^{(k-1}\ \hat{a}_{k-1,j}^{(k-1)}}{\hat{a}_{k-1,k-1}^{(k-1)}}\quad (1+\varepsilon) \qquad |\varepsilon| \leq \nu$$

if $i \leq j$ then k has values $k = 2 \ldots . i$

if $i < j$ then k has values $k = 2 \ldots . j$

and for either case $\hat{a}_{i,k-1}^{(k-1)}$, $\hat{a}_{k-1,j}^{(k-1)}$ and $\hat{a}_{k-1,k-1}^{(k-1)}$ are all elements of $\hat{B}^{(k-1)}$. Since $\hat{B}^{(k-1)}$ is positive definite, then

$$\left(\hat{a}_{i,k-1}^{(k-1)}\right)^2 \leq \hat{a}_{ii}^{(k-1)}\ \hat{a}_{k-1,k-1}^{(k-1}$$

and $$\left(\hat{a}^{(k-1)}_{k-1,j}\right)^2 \leq \hat{a}^{(k-1)}_{jj}\ \hat{a}^{(k-1)}_{k-1,k-1}$$

$$\hat{m}_{i,k-1}\ \hat{a}^{(k-1)}_{k-1,j} \leq \left(\sqrt{a^{(k-1)}_{ii}\ a^{(k-1)}_{jj}}\right)\ \left(1+\nu\right) \leq 1+\nu$$

$$|\varepsilon^{(k)}_{ij}| \leq \nu + \nu\ (1+\nu)$$

$$\leq 2\nu$$

This result applies to all $\varepsilon^{(k)}_{ij}$ except $\varepsilon^{(j+1)}_{ij}$ $(i>j)$
For $\varepsilon^{(j+1)}_{ij}$ we have

$$\hat{m}_{ij} = f\ell\left(\frac{\hat{a}^{(j)}_{ij}}{\hat{a}^{(j)}_{jj}}\right) = \frac{\hat{a}^{(j)}_{ij}}{\hat{a}^{(j)}_{jj}}\ (1+\varepsilon)$$

$$\eta_{ij} = \frac{\hat{a}^{(j)}_{ij}}{\hat{a}^{(j)}_{jj}}\ \varepsilon$$

$$\left|\varepsilon^{(j+1)}_{ij}\right| = \left|\frac{\hat{a}^{(j)}_{jj}\ \hat{a}^{(j)}_{ij}}{\hat{a}^{(j)}_{jj}}\ \varepsilon\right| = \left|\hat{a}^{(j)}_{ij}\,\varepsilon\right| \leq \nu < 2\nu$$

and therefore we need not give this element special treatment.

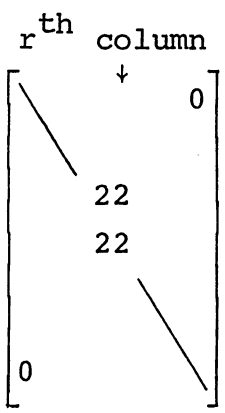
Combining all these results we have that: ——

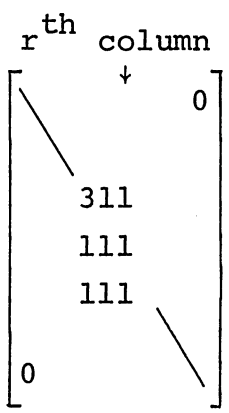
$$|E| \leq \nu \begin{bmatrix} 0 & 0 & 0 & & & & \\ 1 & 1 & 1 & & & & \\ 1 & 2 & 2 & 1 & 1 & & \\ & 1 & 2 & 2 & 2 & & \\ & 1 & 2 & 3 & 3 & 1 & 1 \\ & & & 1 & 2 & 2 & 2 \\ & & & 1 & 2 & 3 & 3 & 1 & 1 \\ & & & & & \text{etc.} & & & \end{bmatrix}$$

and therefore that $||E||_\infty \leq 22\nu$

$$||E||_1 \leq 20\nu \ .$$

To show that $\hat{B}^{(r)}$ is positive definite we write $\hat{B}^{(r)} = B^{(r)} + J^{(r)}$ where $J^{(r)}$ is the matrix of errors introduced up to the r^{th} stage of the elimination, i.e., $J^{(r)}$ has as its elements $e_{ij}^{(k)}$, $k = 2 \ldots . r$. Thus it can be seen that $J^{(r)}$ is symmetric and minimum eigenvalue $\hat{B}^{(r)} \geq$ minimum eigenvalue $B^{(r)}$ + minimum eigenvalue $J^{(r)}$.

Now $|J^{(r)}| \leq \nu$ 

or $\leq \nu$ 

depending on whether r is even or odd. Thus, for any r, the minimum eigenvalue of $J^{(r)} \geq -5\nu$

Therefore, provided minimum eigenvalue $B^{(r)} > 5\nu$ for any r, then minimum eigenvalue of $\hat{B}^{(r)} > 0$ and thus $\hat{B}^{(r)}$ is positive definite.

Now $$\hat{a}_{ii}^{(2)} = f\ell\left[a_{ii}^{(1)} - f\ell\left(m_{il}\ a_{lj}^{(1)}\right)\right]$$

$$= \left[a_{ii}^{(1)} - m_{il}\ a_{lj}^{(1)}\ (1+\varepsilon_1)\right]\ \left(1+\varepsilon_2\right)$$

$$\leq a_{ii}^{(1)}\ \left(1+\varepsilon_2\right)$$

and in general $\hat{a}_{ii}^{(k)} \leq \hat{a}_{ii}^{(k-1)}\left(1+\varepsilon_2\right) \qquad |\varepsilon_2| \leq \nu$

Now the largest element in a positive definite matrix lies on the diagonal and since we operate on each row at most 3 times, if we scale originally so that $a_{ii}^{(1)} \leq (1+3\nu)$ then we can be sure that $\hat{a}_{ij}^{(k)} \leq 1$ for any k.

We shall now turn our attention to the back solution and consider the solution of a lower triangular set where the matrix of coefficients does not necessarily have unit elements on the diagonal, but where if we are solving Ly = d

$L = \hat{m}_{ij}$ then $\hat{m}_{ij} = 0$ $\qquad$ i even $\quad i-j \geq 3$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ i odd $\quad i-j \geq 4$

for r even

$$\hat{y}_r = f\ell\left(\frac{d_r - \hat{m}_{r,r-2}\,\hat{y}_{r-2} - \hat{m}_{r,r-1}\,\hat{y}_{r-1}}{\hat{m}_{rr}}\right)$$

$$\hat{y}_r = \frac{d_r - \hat{m}_{r,r-2}\,\hat{y}_{r-2}\,(1+E_{r,r-2}) - \hat{m}_{r,r-1}\,\hat{y}_{r-1}\,(1+E_{r,r-1})}{\hat{m}_{rr}\,(1+\eta_r)\,(1+\varepsilon_r)}$$

for r odd

$$\hat{y}_r = \frac{d_r - \hat{m}_{r,r-3}\hat{y}_{r-3}(1+E_{r,r-3}) - \hat{m}_{r,r-2}\hat{y}_{r-2}(1+E_{r,r-2}) - \hat{m}_{r,r-1}\hat{y}_{r-1}(1+E_{r,r-1})}{\hat{m}_{rr}\,(1+\eta_r)\,(1+\varepsilon_r)}$$

where $\left|\eta_r\right| \leq \nu \qquad \left|\varepsilon_r\right| \leq \nu$

$$E_{ri} \leq (r+1-i)\ 1{\cdot}06\nu$$

It is more convenient to express the above in the following form: ——

for r even

$$d_r = \hat{y}_r\hat{m}_{rr}(1+E_{rr}) - \hat{m}_{r,r-2}\hat{y}_{r-2}(1+E_{r,r-2}) - \hat{m}_{r,r-1}\hat{y}_{r-1}(1+E_{r,r-1})$$

and for r odd

$$d_r = \hat{y}_r m_{rr}(1+E_{rr}) + \hat{m}_{r,r-3}\hat{y}_{r-3}(1+E_{r,r-3}) + \hat{m}_{r,r-2}\hat{y}_{r-2}(1+E_{r,r-2}) + \hat{m}_{r,r-1}\hat{y}_{r-1}(1+E_{r,r-1})$$

where $1+E_{rr} = \left(|+\eta_r\right)\left(1+\epsilon_r\right)$ $\qquad |E_{rr}| \leq 2\nu$

and in particular

$$|E_{r,r-2}| \leq \left(r+1-(r-2)\right)1\cdot06\nu = 3\times1\cdot06\nu$$

$$|E_{r,r-1}| \leq \left(r+1-(r-1)\right)1\cdot06\nu = 2\times1\cdot06\nu$$

$$|E_{r,r-3}| \leq \left(r+1-(r-3)\right)1\cdot06\nu = 4\times1\cdot06\nu$$

Thus the computed vector is the exact solution of

$$(\hat{L} + d\hat{L})\hat{y} = d$$

where $d\hat{L}$ is bounded as shown below

$$|d\hat{L}| \leq 1\cdot06\nu \begin{bmatrix} |\hat{m}_{11}| & & & & \\ 2|\hat{m}_{21}| & 2|\hat{m}_{22}| & & & \\ 3|\hat{m}_{31}| & 2|\hat{m}_{32}| & 2|\hat{m}_{33}| & & \\ & 3|\hat{m}_{42}| & 2|\hat{m}_{43}| & 2|\hat{m}_{44}| & \\ & 4|\hat{m}_{52}| & 3|\hat{m}_{52}| & 2|\hat{m}_{54}| & 2|\hat{m}_{55}| \\ & & & \text{etc.} & \end{bmatrix}$$

If the diagonal elements of L are unity then the diagonal elements of $|d\hat{L}|$ are unity.

An analysis of $d\hat{U}$ for $(\hat{U}+d\hat{U})\hat{w} = y$ can be computed in an exactly analogous fashion and

$$|d\hat{U}| \leq 1\cdot 06\nu \begin{bmatrix} 2|a_{11}^{(1)}| & 2|a_{12}^{(1)}| & 3|a_{13}^{(1)}| & & & \\ & 2|a_{22}^{(2)}| & 2|a_{23}^{(2)}| & 3|a_{24}^{(2)}| & 4|a_{25}^{(2)}| & \cdots \\ & & & \text{etc.} & & \end{bmatrix}$$

We now need a bound on $||\hat{L}d\hat{U}||_\infty$

$$\hat{L}d\hat{U} = 1\cdot 06H\nu$$

where for i even $\quad h_{ij} = 0 \quad j<i \quad i-j \geq 3$

$\qquad\qquad\qquad\qquad\qquad\qquad i \leq j \quad j-i \geq 4$

$$|h_{i,i-2}| \leq 2|\hat{m}_{i,i-2}\ \hat{a}_{i-2,i-2}^{(i-2)}|$$

$$|h_{i,i-1}| \leq 2|\hat{m}_{i,i-2}\ \hat{a}_{i,i-1}| + 2|\hat{m}_{i,i-1}\ \hat{a}_{i-1,i-1}^{(i-1)}|$$

$$|h_{ii}| \leq 3|\hat{m}_{i,i-2}\ \hat{a}_{i-2,i}^{(i-2)}| + 2|\hat{m}_{i,i-1}\ a_{i-1,i}^{(i-1)}| + 2|\hat{m}_{ii}\ \hat{a}_{ii}^{(i)}|$$

$$|h_{i,i+1}| \leq 4|\hat{m}_{i,i-2}\ \hat{a}_{i-2,i+1}^{(i-2)}| + 3|\hat{m}_{i,i-1}\ \hat{a}_{i-1,i+1}^{(i-1)}| + 2|\hat{m}_{ii}\ \hat{a}_{i,i+1}^{(i)}|$$

$$|h_{i,i+2}| \leq 3|\hat{m}_{ii}\ \hat{a}_{i,i+2}^{(i)}|$$

$$|h_{i,i+3}| \leq 4|\hat{m}_{ii}\ \hat{a}_{i,i+3}^{(i)}|$$

for i odd $\quad h_{ij} = 0 \quad j<i \quad i-j \geq 4$

$\qquad\qquad\qquad\qquad\qquad\qquad i \leq j \quad j-i \geq 3$

$$\left|h_{i,i-3}\right| \leq 2\left|\hat{m}_{i,i-3}\ \hat{a}^{(i-3)}_{i-3,i-3}\right|$$

$$\left|h_{i,i-2}\right| \leq 2\left|\hat{m}_{i,i-3}\ \hat{a}^{(i-3)}_{i-3,i-2}\right| + 2\left|\hat{m}_{i,i-2}\ \hat{a}^{(i-1)}_{i-2,i-2}\right|$$

$$\left|h_{i,i-1}\right| \leq 3\left|\hat{m}_{i,i-3}\ \hat{a}^{(i-3)}_{i-3,i-1}\right| + 2\left|\hat{m}_{i,i-2}\ \hat{a}^{(i-2)}_{i-2,i-1}\right| + 2\left|\hat{m}_{i,i-1}\ \hat{a}^{(i-1)}_{i-1,i-1}\right|$$

$$\left|h_{ii}\right| \leq 4\left|\hat{m}_{i,i-3}\ \hat{a}^{(i-3)}_{i-3,i}\right| + 3\left|\hat{m}_{i,i-2}\ \hat{a}^{(i-2)}_{i-2,i}\right| + 2\left|\hat{m}_{i,i-1}\ \hat{a}^{(i-1)}_{i-1,i}\right| + 2\left|\hat{m}_{ii}\ \hat{a}^{(i)}_{ii}\right|$$

$$\left|h_{i,i+1}\right| \leq 3\left|\hat{m}_{i,i-1}\ \hat{a}^{(i-1)}_{i-1,i+1}\right| + 2\left|\hat{m}_{ii}\ \hat{a}^{(i)}_{i,i+1}\right|$$

$$\left|h_{i,k+2}\right| \quad 4\left|\hat{m}_{i,i-1}\ \hat{a}^{(i-1)}_{i-1,i+2}\right| + 3\left|\hat{m}_{ii}\ \hat{a}^{(i)}_{i,i+2}\right|$$

By the arguments used in calculating $||E||_\infty$ we see that

$\left|\hat{m}_{k\ell}\ a^{(\ell)}_{\ell m}\right| \leq 1$

and therefore $||\hat{L}d\hat{U}||_\infty \leq ||\ |\hat{L}d\hat{U}|\ ||_\infty \leq 1{\cdot}06\times 35\nu$

$||\hat{L}d\hat{U}||_1 \leq 1{\cdot}06\times 39\nu$

In exactly the same way we can calculate bounds on $||d\hat{L}\hat{U}||_\infty$ and $||d\hat{L}d\hat{U}||_\infty$ and we obtain

$$||d\hat{L}\hat{U}||_\infty \leq 1{\cdot}06\times 36\nu \qquad ||d\hat{L}\hat{U}||_1 \leq 1{\cdot}06\times 32\nu$$

$$||d\hat{L}d\hat{U}||_\infty \leq (1{\cdot}06\nu)^2 101 \qquad ||d\hat{L}d\hat{U}||_1 \leq (1{\cdot}06\nu)^2 90$$

Therefore $||K||_\infty \leq 2\times 11\nu + 1{\cdot}06\times 35\nu + 1{\cdot}06\times 36\nu + (1{\cdot}06\nu)^2 101$

$$||K||_1 \leq 2\times 10\nu + 1{\cdot}06\times 39\nu + 1{\cdot}06\times 32\nu + (1\ 06\nu)^2 90$$

ignoring terms of order ν^2

$$||K||_2 \leq \sqrt{||K||_\infty ||K||_1} \leq 97\nu$$

We can now obtain a bound P on $||R||_A$. If we write $||A^{-\frac{1}{2}}||_2 \quad ||A^{\frac{1}{2}}||_2 \leq c$ and $||Q||_2 \leq \frac{c'}{h}$ where c and c' are both constant. Then using $||K||_2 \leq 97\nu$ and substituting into equation (9)

$$||R||_A \leq \frac{cc'97\nu}{h}\left(1 - \frac{c'97\nu}{h}\right)^{-1}$$

$$\sim \frac{cc'97\nu}{h}\left(1 + \frac{c'97\nu}{h} \ldots .\right)$$

$$\sim \frac{cc'97\nu}{h} + O\left(\frac{\nu}{h}\right)^2$$

$$\therefore \quad P = \frac{cc'97\nu}{h}$$

We are now in a position to substitute into equation (7). We shall obtain results showing the asymptotic behaviour of the error as $k,h \to 0$ provided $\frac{\nu}{kh}$ is kept constant.

Consider first the case when $k = c_1 h$

then $\quad || \; |M| \; ||_A \leq pc_1 + p'h \qquad\qquad p,p'$ constant.

$$\phi_1 \leq \left(\frac{1 + cc'97\nu}{h}\right)\left(\frac{1 + 6\cdot 36\nu cc'}{h}\left(pc_1+p'h\right)\right)$$

$$\leq \frac{1 + cc'97\nu}{h} + \frac{6\cdot 36cc'pc_1\nu}{h} + O\left(\frac{\nu}{h}\right)^2$$

$$\sim 1 + \frac{cc'\nu}{h}\left(97 + 6\ 36pc_1\right)$$

$$\phi_2 \leq \frac{cc'97\nu}{h} + \left(\frac{1 + cc'97\nu}{h}\right)\frac{6\cdot 36cc'\nu}{h}\left(pc_1+p'h\right)$$

$$\leq \frac{cc'97\nu}{h} + \frac{6\cdot 36cc'pc_1\nu}{h} + O\left(\frac{\nu}{h}\right)^2$$

$$\sim \frac{cc'\nu}{h}\left(97 + 6\cdot 36pc_1\right)$$

Therefore

$$\phi_1 \sim 1 + \frac{k'\nu}{h}$$

$$\phi_2 \sim \frac{k'\nu}{h} \qquad \text{where } k' = cc'\left(97 + 6\cdot 36pc_1\right)$$

$$||w_{n+1}||_A \leq \frac{\beta k'\nu}{h} \frac{\left(\left(1 + \frac{k'\nu}{h}\right)^n - 1\right)}{k'\frac{\nu}{h}}$$

$$\leq \beta \left(\left(1 + \frac{k'\nu}{h}\right)^n - 1\right)$$

$$\leq \frac{\beta n k'\nu}{h} e^{\frac{nk'\nu}{h}}$$

But $n = \frac{T}{k}$ and if $\frac{\nu}{kh} = K$ a constant

then $||w_{n+1}||_A = ||\hat{U}_{n+1} - U_{n+1}||_{L^2} \leq \beta T k' K e^{TKk'} \leq K''$

where K" is a constant

Now consider the case where $k = c_2 h^2$, then

$||Q||_2 \leq \frac{c''}{h}$ $\quad$ $||\,|M|\,||_2 \leq p_2 h$ $\quad$ c",p_2 constant

$$\phi_1 = \left(1 + \frac{cc''97\nu}{h}\right)\left(1 + \frac{6\cdot 36 cc'' p_2 h\nu}{h}\right)$$

$$= 1 + \frac{cc''97\nu}{h} + O(\nu)$$

$$\phi_2 = \frac{cc''97\nu}{h} + \left(1 + \frac{cc''97\nu}{h}\right)\frac{6\cdot 36 cc'' p_2 h\nu}{h}$$

$$= \frac{cc''97\nu}{h} + O(\nu)$$

Therefore $\phi_1 = 1 + \frac{k'\nu}{h}$

$\phi_2 = \frac{k'\nu}{h}$ $\quad$ where k' = cc"97

and $$\left|\left|w_{n+1}\right|\right|_A \leq \beta\left(\left(1 + \frac{k'\nu}{h}\right)^n - 1\right)$$

$$\leq \frac{\beta n k'\nu}{h} e^{\frac{nk'\nu}{h}}$$

But $n = \frac{T}{k}$ and if we keep $\frac{\nu}{kh}$ constant

then $\left|\left|w_{n+1}\right|\right|_A \leq K''$ where K" is a constant.

We have just proved the following two theorems.

Theorem 1:

Let $\nu = N^{1-\tau}$ where τ is the number of floating base -N digits in the mantissa. For the boundary value problem satisfying (1) whose solution is approximated by the function $U(x,t) = \sum_{\ell} \alpha_{\ell}(t) v_{\ell}(x)$ defined by (2) and using the difference scheme set out in (4) then, if the approximate solution is required to lie in a Hermite space of piecewise polynomials of degree 3, based on a rectangular grid of mesh size h, and if $\Delta t = c_1 h$, where c_1 is a constant, and if as $\Delta t, h \to 0$, $\frac{\nu}{\Delta t h}$ is kept constant, then the computed solution $\hat{U}$ satisfies

$$\left|\left|\hat{U}-U\right|\right|_{L^2} \leq K' \qquad \text{where K' is a constant.}$$

Theorem 2:

For the boundary value problem described in Theorem 1, but where $\Delta t = c_2 h^2$, where c_2 is a constant and if as $\Delta t, h \to 0$, $\frac{\nu}{\Delta t h}$ is kept constant, then the computed solution $\hat{U}$ satisfies

$$\left|\left|\hat{U}-U\right|\right|_{L^2} \leq K'' \qquad \text{where } K'' \text{ is a constant.}$$

If we compare these results with those obtained in [2] for the finite difference method we see that exactly the same asymptotic behaviour is exhibited in both cases.

References

1. Douglas, J., Jr., and Dupont, T., Galerkin Methods for Parabolic Equations. To appear.

2. Rachford, H. H., Jr., Rounding Errors in Parabolic Problems 1, The one space variable case. SIAM J. of Num. Analysis, 5, no.1, 156-171(1968).

3. Wilkinson, J. H., Rounding Errors in Algebraic Processes, Prentice Hall (1963).

4. Wilkinson, J. H., Error Analysis of Direct Methods of Matrix Inversion. J. Assoc. Comp. Mach. 8, 281-330(1961).