King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR 097

April 1987

Defect Correction for Finite Element Discretization

Bengt Lindberg

DHAHRAN 31261 ● SAUDI ARABIA ● www.kfupm.edu.sa/math/ ● E-mail: mathdept@kfupm.edu.sa

Defect correction for finite element discretization

Bengt Lindberg

Abstract: Defect correction for a finite element disretization of a linear one-dimensional two-point boundary value problem is studied. For a piecewise linear approximation the defect is chosen to be the Petrov-Galerkin approximation using piecewise quadratic elements and piecewise linear test functions. Using matrix-analysis it is proved that iterated defect correction converges to the Petrov-Galerkin solution without the need for error expansions for the basic discretization.

1. Introduction: To examine the possibility of defect correction in connection with finite element discretizations we study the following class of one-dimensional linear boundary value problems,

(1) $\quad Lu = f \qquad u(0) = 0; \quad u(1) = 0$

with

$$L = -\frac{d}{dx}\left(p(x)\frac{d}{dx}\right) + q(x)$$

$$p(x) > 0 \qquad q(x) \geq 0.$$

2. A finite element discretization:

Introduce the notation

(2) $$(u, v) = \int_0^1 uv \, dx$$

(3) $$a(u, v) = \int_0^1 p \frac{du}{dx} \frac{dv}{dx} + quv \, dx$$

(4) H_0^1 = the space of functions defined on (0, 1) with

$v(0) = v(1) = 0$ and with finite norm

$$\|v\|_1 = \sqrt{\int_0^1 v^2 + (v')^2 \, dx}$$

Define

(5) $$u^h = \sum_{v=1}^{N} c_v \phi_v$$

where $\{\phi_v\}_{v=0}^{N+1}$ are the basis functions for piecewise linear functions with $u(0) = u(1) = 0$ and with knots at x_i, $i = 0, 1, \ldots, N + 1$ such that $0 = x_0 < x_1 < \ldots x_N < x_{N+1} = 1$.

The basis functions are the well-known roof-functions

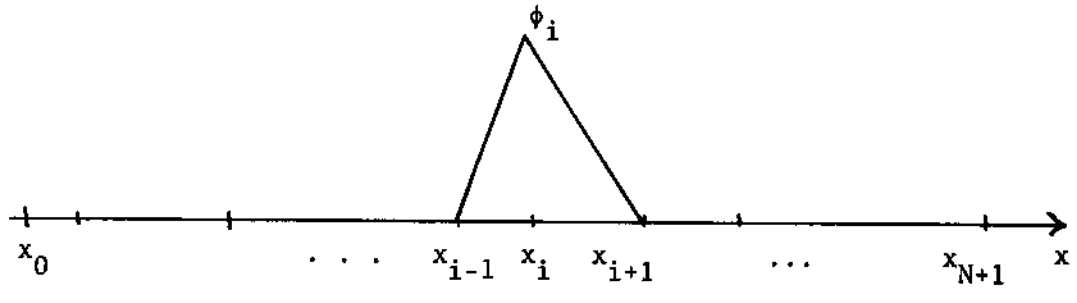


Figure 1: Roof-functions

A finite element solution of our problem (1) is given by (5) where the coefficients c_v are the solution of the system of linear equations.

(6) $Kc = G$

where $K = \{a(\phi_i, \phi_j)\}$ N×N matrix

$G = \{(\phi_i, f)\}$ N-vector

$c = \{c_i\}$ N-vector

The elements of the matrix K for a model problem are given in section 5.1.

Note that c_k is an approximation to $u(x_k)$, see figure 2.

(7) For future reference denote by S^h the space of piecewise linear functions with $v(0) = v(1) = 0$ and with norm $\|v\|_1 = \sqrt{\int v^2 + (v')^2 \, dx}$. Note that $S^h \subset H_0^1$.

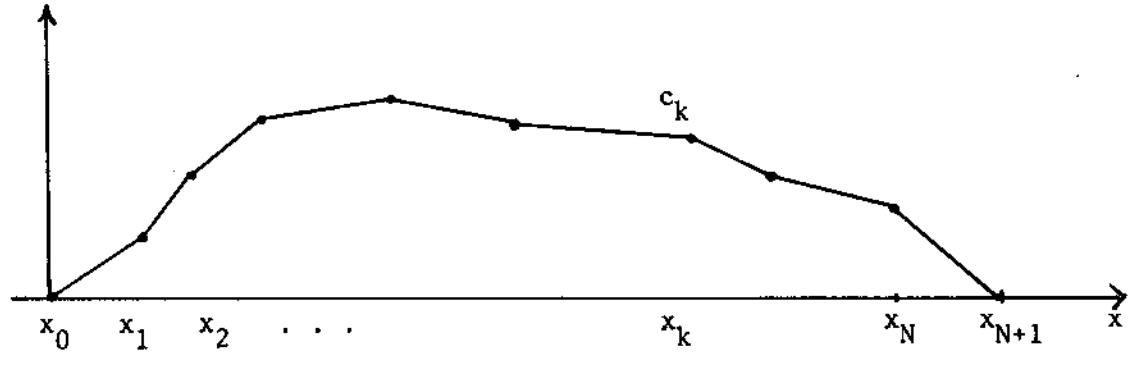


Figure-2: $u^h(x)$

Furthermore we know that u^h is the best approximation to u in S^h. Thus we can not hope to get any better approximation to u from S^h. However, if we look in other subspaces of H_0^1 we can hope to do better. For example,

for $v^h \in Q^h$ where

(8) Q^h = the space of piecewise quadratic functions with $v(0) = v(1) = 0$ and with knots at x_i, $i = 0, 1, \ldots, N + 1$ and norm $\|v\|_1$

a better approximation may be obtained.

3. The defect

We introduce the general formalism of Lindberg (1980) and describe our finite-element discretization in that formalism in order to arrive at a useful definition of the defect for the numerical method.

Consider linear functional equations

(9) $$F(y) = 0$$

where $F: E \to E^0$ is a linear operator from a Banach space E into another Banach space E^0. The functional equation above is approximated by

(10) $$\Phi_h(\eta) = 0$$

where $\Phi_h: E_h \to E_h^0$ is a linear operator from one Banach space E_h into another Banach space E_h^0. Define bounded linear transformations Δ_h and Δ_h^0 mapping E, E^0 into E_h, E_h^0 respectively.

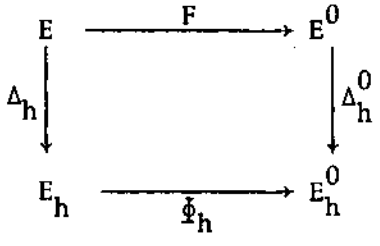


Figure-3: Mappings

For our problem the operators Δ_h and Δ_h^0 are implicitly defined according to

(11) $\Delta_h z = [z(x_1), z(x_2), \ldots, z(x_N)]$,

(12) $\Delta_h^0 g = [(\phi_1, g), (\phi_2, g), \ldots, (\phi_N, g)]$.

Note the operator Δ_h^0 which is quite different from the corresponding operators for finite difference discretizations.

The spaces E_h and E_h^0 are R^N with suitable norms. For our present needs the Euclidean norm will be suitable.

The operator $\Phi_h(\eta) = K\eta - G$. For the theory of Lindberg (1980) to be applicable we need that

$$\Phi_h(\Delta_h z) = \Delta_h^0 \{F(z) + \sum_{\nu=p}^{M} h^\nu f_\nu(z)\} + O(h^{M+1})$$

i.e. Φ_h should be consistent with F with order of consistency p and the error must have a certain regularity. The parameter h is the gridspacing of the discretization which is implicitly assumed to be constant. For our present application these explicit and implicit assumptions are hard to prove or not satisfied. In section 4 we will consider other types of conditions.

To estimate the error of the solution η of the discretization we need a more accurate approximation $\Phi_h^E : E_h \to E_h^0$ of F, i.e., Φ_h^E should be such that

$$\Phi_h^E(\Delta_h z) = \Delta_h^0\{F(z)\} + O(h^q), \quad q > p.$$

This means that Φ_h^E is consistent with F with order of consistency q. For the method of section 2 a natural definition of the defect operator seems to be

$$(13) \qquad [\Phi_h^E(\Delta_h z)]_i = a(\phi_i, \Pi(\Delta_h z)) - (\phi_i, f) \qquad i = 1, 2, \ldots, N$$

where $\Pi: E_h \to D \subset E$ is a linear mapping from E_h into a subspace of E. One example of Π is given by

$$(14) \qquad \Pi(\Delta_h z) = \sum_{\nu=1}^{N} z(x_\nu)\, \psi_\nu$$

where $\{\psi_\nu\}_{\nu=1}^N$ is the standard basis of the space of piecewise quadratic functions with nodes at the same points as $\{\phi_\nu\}_{\nu=1}^N$. To be specific let $N = 2M$ and $x_{2i-1} = \frac{1}{2}(x_{2i-2} + x_{2i})$, $i = 1, 2, \ldots, M$.

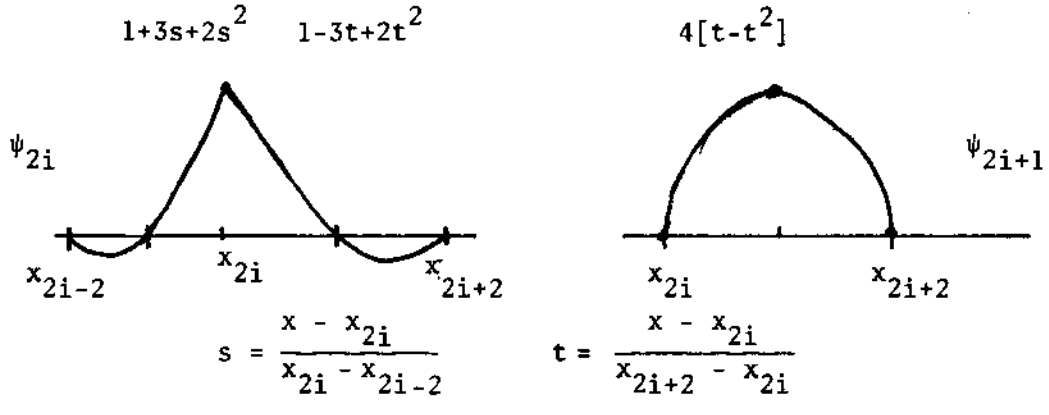


Figure 4: Basis functions for piecewise quadratics

In the rest of this report we will only consider Π according to (14). Note that

(15) $\phi_h^E(\eta) = M\eta - G$

where

(16) $M = \{a(\phi_i, \psi_j)\}$

and G is as in (6).

The defect correction according to Lindberg (1980) then gives that $\|\eta^E - \eta\|_{E_h}$ where

$$\Phi_h(\eta) = 0$$

$$\Phi_h(\eta^E) + \phi_h^E(\eta) = 0$$

is an estimate of the erro in η. In the matrix notation we have

$$K\eta = G$$

$$K\eta^E - G + M\eta - G = 0$$

or after rearranging and subtracting $K\eta - G$ from the right hand side

$$K\eta^E = (K - M)\eta + G.$$

4. A defect correction for linear problems

Given a linear problem

(17) $Lu = f$

and two approximations to it

(18) $$L_0 u_h = f_h$$

(19) $$L_1 w_h = f_h$$

with (19) more accurate than (18).

Then w_h can be computed iteratively according to

$$L_0 w_h^0 = f_h$$

(20) $$L_0 w_h^{(j+1)} = (L_0 - L_1) w_h^{(j)} + f_h \qquad j = 0, 1, \ldots, .$$

If this iteration converges it converges to the solution w_h of (19).

The equations (18) and (19) correspond in our case to

(18') $$K\underline{c} = \underline{G}$$

(19') $$M\underline{d} = \underline{G}$$

where $$u_h = \sum_{v=1}^{N} c_v \phi_v$$

$$w_h = \sum_{v=1}^{N} d_v \psi_v$$

gives the relation between u_h, w_h and $\underline{c}$, $\underline{d}$.

Thus, the iterations are

(20') $$K\underline{d}^{j+1} = (K - M)\,\underline{d}^j + \underline{G} \qquad j = 0, 1, \ldots$$

with $\underline{d}^0$ defined from

$$K\underline{d}^0 = \underline{G}\,.$$

Note that the first iteration $j = 0$, is exactly the defect correction from the previous section. To study the convergence of (20') we rewrite it as

$$(21) \qquad K\underline{d}^{j+1} = (K - M)\, K^{-1}\, K\underline{d}^{j} + \underline{G}.$$

$(K - M)K^{-1}$ is the iteration matrix.

If $\|(K - M)K^{-1}\| = \sigma < 1$ the iterations will converge at a rate determined by σ. If (21) converges it converges to the solution of

$$K\underline{d} = (K - M)\underline{d} + \underline{G}$$

i.e. to the solution of

$$M\underline{d} = \underline{G}.$$

In the rest of this report, we will find estimate of σ for some special cases.

5. A model problem

5.1 Introduction

Consider

$$-u'' + u = f \qquad u(0) = u(1) = 0.$$

Let $I = \{x_0, x_1, x_2, \ldots, x_N, x_{N+1}\}$

with

$$0 = x_0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = 1$$

be a partition of $[0, 1]$. Let

$$h_i = x_i - x_{i-1} \qquad i = 1, 2, \ldots, N + 1,$$

and introduce the notation

$$I_0 = I \quad \text{with} \quad h_i = h = \frac{1}{N+1} \qquad i = 1, 2, \ldots, N + 1$$

$$I_1 = I \quad \text{with} \quad h_{2k} = h_{2k-1} \qquad k = 1, 2, \ldots, \frac{N+1}{2}$$

$$N+1 \quad \text{even}$$

I_0 is an equidistant partition and I_1 is illustrated in figure 5:

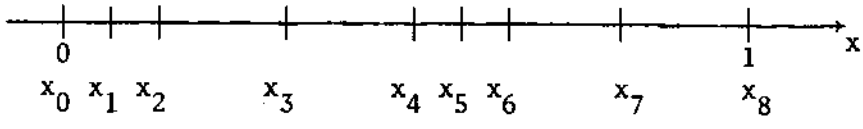


Figure 5: I_1 with $N + 1 = 8$

For our model problem

(22) $\quad K = a(\phi_i, \phi_j) = K_1 + K_2$

where $\quad K_1 = (\phi_i', \phi_j') \qquad K_2 = (\phi_i, \phi_j)$

Lengthy trivial calculations give

$$
(23)\qquad K_1 = \begin{bmatrix}
\frac{1}{h_1}+\frac{1}{h_2} & -\frac{1}{h_2} & & & & & \\
-\frac{1}{h_2} & \frac{1}{h_2}+\frac{1}{h_3} & -\frac{1}{h_3} & & & & \\
 & -\frac{1}{h_3} & \frac{1}{h_3}+\frac{1}{h_4} & -\frac{1}{h_4} & & & \\
 & & \cdot & & & & \\
 & & & \cdot & & & \\
 & & & & \cdot & & \\
 & & & & -\frac{1}{h_{N-1}} & \frac{1}{h_{N-1}}+\frac{1}{h_N}
\end{bmatrix}
$$

$$
(23')\qquad K_2 = \frac{1}{6}\begin{bmatrix}
2(h_1+h_2) & h_2 & & & & & \\
h_2 & 2(h_2+h_3) & h_3 & & & & \\
 & h_3 & 2(h_2+h_3) & h_4 & & & \\
 & & \cdot & & & & \\
 & & & \cdot & & & \\
 & & & & \cdot & & \\
 & & & & h_{N-1} & 2(h_{N-1}+h_N)
\end{bmatrix}
$$

For the partition I_1 we have $h_1 = h_2$, $h_3 = h_4 \ldots$, $h_{N+1} = h_N$ and

(24)

$$K_1 = \begin{bmatrix} \frac{2}{h_2} & -\frac{1}{h_2} & & & & & \\ -\frac{1}{h_2} & \frac{1}{h_2}+\frac{1}{h_4} & -\frac{1}{h_4} & & & & \\ & -\frac{1}{h_4} & \frac{2}{h_4} & -\frac{1}{h_4} & & & \\ & & -\frac{1}{h_4} & \frac{1}{h_4}+\frac{1}{h_6} & -\frac{1}{h_6} & & \\ & & & & & & \\ & & & & & -\frac{1}{h_{N+1}} & \frac{2}{h_{N+1}} \end{bmatrix}$$

(24')

$$K_2 = \frac{1}{6}\begin{bmatrix} 4h_2 & h_2 & & & \\ h_2 & 2(h_2+h_4) & h_4 & h_4 & \\ & h_4 & 4h_4 & h_6 & \\ & & h_6 & 2(h_6+h_8) & h_8 \end{bmatrix}$$

For the partition I_0 all h_i are equal, so

(25) $$K = \mathrm{trid}\left(\frac{h}{6} - \frac{1}{h},\ \frac{4h}{6} + \frac{2}{h},\ \frac{h}{6} - \frac{1}{h}\right)$$

where trid stands for tridiagonal.

In the appendix we calculate the elements of the matrix

$$M = a(\phi_i, \psi_j) = M_1 + M_2$$

where

(26) $$M_1 = (\phi_i', \psi_j'),\quad M_2 = (\phi_i, \psi_j)\ .$$

For the partition I_1 we get with $P = \frac{N+1}{2}$

$K - M = K_1 - M_1 + K_2 - M_2 =$

(27)
$$= \frac{1}{24}\begin{bmatrix} 16h_2 & 4h_2 & & & & \\ 4h_2 & 8(h_2+h_4) & 4h_4 & & & \\ & 4h_4 & 16h_4 & 4h_4 & & \\ & & 4h_4 & 8(h_4+h_6) & 4h_6 & \\ & & & 4h_6 & 16h_6 & 4h_6 \\ & & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot \end{bmatrix}$$

$$
-\frac{1}{24}\begin{bmatrix}
20h_2 & 2h_2 & & & & \\
6h_2 & 7(h_2+h_4) & 6h_4 & -h_4 & & \\
 & 2h_4 & 20h_4 & 2h_4 & & \\
 & -h_4 & 6h_4 & 7(h_6+h_6) & 6h_6 & -h_6
\end{bmatrix}
$$

$$
=\frac{1}{24}\begin{bmatrix}
-4h_2 & 2h_2 & & & & & \\
-2h_2 & (h_2+h_4) & -2h_4 & h_4 & & & \\
 & 2h_4 & -4h_4 & 2h_4 & & & \\
 & h_4 & -2h_4 & h_4+h_6 & -2h_6 & h_6 & \\
 & & & 2h_6 & -4h_6 & 2h_6 & \\
 & & & & & 2h_{2p} & -4h_{2p}
\end{bmatrix}
$$

Hence for any vector $\underline{y}$ we have

$$(28) \qquad (K - M)\,\underline{y} = \frac{1}{24} \begin{bmatrix} 2h_2\Delta^2 y_0 \\ h_2\,\Delta^2 h_0 + h_4\,\Delta^2 y_2 \\ 2h_4\Delta^2 y_2 \\ h_4\Delta^2 y_2 + h_6\Delta^2 y_4 \\ \vdots \\ 2h_{2p}\Delta^2_{2p-2} \end{bmatrix}$$

where we have defined $y_0 = 0$ and $y_{2p} = 0$.

5.2 Crude estimates of the norm of the iteration matrix

Note that for any square matrix A we have

$$\|A\|_2 \leq \sqrt{\|A\|_1 \cdot \|A\|_\infty}$$

Directly from (27) we get for the partition I_1

$$\|K - M\|_1 = \frac{8}{24} h_{max} = \|K - M\|_\infty$$

Hence

$$(29) \qquad \|K - M\|_2 \leq \frac{1}{3} h_{max}$$

For symmetric positive definite matrices we know

$$\|A^{-1}\| \leq \frac{1}{\lambda_{min}}$$

where λ_{min} is the smallest eigenvalue of A.

Gerschgorin's circle theorem gives

$$\lambda_{min} \geq \min_i \left(a_{ii} - \sum_{j \neq i} |a_{ij}|\right)$$

Hence for the partition I_1 we get for the matrix K

$$\lambda_{min} \geq \min \left\{\frac{1}{h_{2i}} + \frac{1}{h_{2i+2}} + \frac{2}{6}(h_{2i} + h_{2i+2}) \right.$$

$$\left. - \left(\frac{1}{h_{2i}} + \frac{1}{h_{2i+2}} - \frac{1}{6}(h_{2i} + h_{2i+2})\right)\right\}$$

$$= \min_i \frac{3}{6}(h_{2i} + h_{2i+2}) \geq h_{min}.$$

Hence

$$\|K^{-1}\| \leq \frac{1}{h_{min}} \tag{30}$$

and

$$\|(K - M)K^{-1}\| \leq \frac{1}{3}\frac{h_{max}}{h_{min}}. \tag{31}$$

For the equidistant partition I_0 the eigenvalues of K are given by

$$\lambda_s = 4h + \frac{2}{h} + 2\left(\frac{1}{h} - h\right)\cos\frac{s\pi}{N+1}, \quad s = 1, 2, \ldots, N$$

(assuming that $h < 1$).

Hence

$\lambda_{min} \geq 2h$ giving

$$\|(K - M)K^{-1}\|_2 \leq \frac{1}{6}.$$

However, numerical experiments with the problem indicate that $\|(K - M)K^{-1}\|_2$ is much smaller than these bounds. In fact they indicate that the norm is proportional to h^2 for I_0. In the next subsection we will derive a sharper estimate for $\|(K - M)K^{-1}\|_2$.

5.3 Estimate of $\|(K - M)K^{-1}\|_2$, the equidistant case

For the matrix

$$K = \text{trid}(b, a, b) \qquad N \times N$$

the eigenvalues are

$$\lambda_s = a + 2\,|b|\cos\frac{s\pi}{N+1}, \quad s = 1, 2, \ldots, N \tag{32}$$

and the eigenvectors are

$$\underset{\sim}{x}_s = (\sin\frac{s\pi}{N+1}, \sin\frac{2s\pi}{N+1}, \ldots, \sin\frac{Ns\pi}{N+1})^T$$

$$s = 1, 2, \ldots, N.$$

Note that the eigenvectors don't depend on the values of a and b.

For our model problem we have

$$a = 4h + \frac{2}{h} \qquad b = h - \frac{1}{h}$$

so

$$\lambda_s = 4h + \frac{2}{h} + 2(\frac{1}{h} - h)\cos\frac{s\pi}{N+1}, \quad s = 1, 2, \ldots, N \tag{33}$$

(assuming that $h < 1$).

Denote by $\underline{t}_s$ the normalized eigenvectors corresponding to $\underline{x}_s$, i.e., $\underline{t}_s^T \underline{t}_s = 1$. Note that $\underline{t}_i^T \underline{t}_j = 0$ if $i \neq j$, because a symmetric matrix has orthogonal eigenvectors.

Introduce the orthogonal matrix

$$Y = [\underline{t}_1, \underline{t}_2, \ldots, \underline{t}_N] \tag{34}$$

and the diagonal matrix

$$\Lambda = \operatorname{diag}(\lambda_s)$$

Then

$$K = Y \Lambda Y^T \tag{35}$$

and

$$K^{-1} = Y\Lambda^{-1}Y^T = \sum_{s=1}^{N} \lambda_s^{-1} \underline{t}_s \underline{t}_s^T \tag{36}$$

Hence

$$(K - M)K^{-1} = \sum_{s=1}^{N} (K - M)\underline{t}_s \lambda_s^{-1} \underline{t}_s^T \tag{37}$$

For the equidistant grid I_0 we have according to (28)

$$(K - M)y = \frac{h}{24} \begin{bmatrix} 2\Delta^2 y_0 \\ \Delta^2 y_0 + \Delta^2 y_2 \\ 2\Delta^2 y_2 \\ \Delta^2 y_2 + \Delta^2 y_4 \\ \vdots \\ 2\Delta^2 y_{N-1} \end{bmatrix} \tag{38}$$

If y is an eigenvector of K we know that

$$by_{2i} + ay_{2i+1} + by_{2i+2} = \lambda y_{2i+1}$$

where λ is the corresponding eigenvalue. Hence

$$(h - \frac{1}{h})y_{2i} + (4h + \frac{2}{h})y_{2i+1} + (h - \frac{1}{h})y_{2i+2} = \lambda y_{2i+1}$$

i.e.

$$h(y_{2i} + 4y_{2i+1} + y_{2i+2}) - \frac{1}{h}\Delta^2 y_{2i} = \lambda y_{2i+1}$$

or

$$\Delta^2 y_{2i} = h^2(y_{2i} + 4y_{2i+1} + y_{2i+2}) - h\lambda y_{2i+1}$$

However, all matrices $\text{trid}(\alpha, \beta, \alpha)$ have the same eigenvectors, hence $\underline{y}$ is also an eigenvector of $\text{trid}(1, 4, 1)$.

Thus

$$y_{2i} + 4y_{2i+1} + y_{2i+2} = \mu y_{2i+1}$$

where

$$\mu_s = 4 + 2\cos\frac{s\pi}{N+1} \quad s = 1, 2, \ldots, N.$$

From these formulas we get

$$2\Delta^2 y_{2i} = 2h^2(4 + 2\cos\frac{s\pi}{N+1})y_{2i+1}$$

$$- 2h(4h + \frac{2}{h} + 2(\frac{1}{h} - h)\cos\frac{3\pi}{N+1})y_{2i+1}$$

$$= (-4 - 4\cos\frac{s\pi}{N+1} + 8h^2\cos\frac{s\pi}{N+1})y_{2i+1}$$

or

$$(39) \qquad 2\Delta^2 y_{2i} = 2\kappa_s y_{2i+1}$$

with

$$(40) \qquad \kappa_s = -2 - 2\cos\frac{s\pi}{N+1} + 4h^2\cos\frac{s\pi}{N+1}$$

Furthermore

$$(41) \qquad \Delta^2 y_{2i-2} + \Delta^2 y_{2i} = \kappa_s(y_{2i-1} + y_{2i+1})$$

$$= \kappa_s(0 + 2\cos\frac{s\pi}{N+1})y_{2i}$$

$$= 2\kappa_s \cdot \cos\frac{s\pi}{N+1}\, y_{2i}$$

Thus for $\underline{y} = \underline{t}_s$ we get

$$(42) \qquad (K - M)\underline{y} = \frac{h}{24}\, 2\kappa_s \underline{\tilde{y}}$$

where

$$(43) \qquad \underline{\tilde{y}} = (y_1, \cos\frac{s\pi}{N+1}\, y_2, y_3, \cos\frac{s\pi}{N+1}\, y_4, \ \ldots\ )^T$$

Furthermore from (33), (40) and (42)

$$(44) \qquad (K - M)\lambda_s^{-1}\underline{y}\underline{y}^T = \frac{h}{24}\,\frac{2\kappa_s}{\lambda_s}\,\underline{\tilde{y}}\underline{y}^T = -\frac{h^2}{24}\,\beta_s\underline{\tilde{y}}\underline{y}^T$$

where

$$(45) \qquad \beta_s = \frac{4(1 + \cos\frac{s\pi}{N+1} - 2h^2\cos\frac{s\pi}{N+1})}{2 + 4h^2 + 2\cos\frac{s\pi}{N+1} - 2h^2\cos\frac{s\pi}{N+1}}\,.$$

Note that for all $s = 1, 2, \ldots, N$

$$(46) \qquad 0 < \beta_s \leq 2$$

Thus

$$(47) \qquad (K - M)K^{-1} = -\frac{h^2}{24} \sum_{s=1}^{N} \beta_s \tilde{t}_s t_s^T$$

where $\tilde{t}_s$ is defined as in (43).

Now consider the matrix

$$(48) \qquad C = \sum_{s=1}^{N} \beta_s \tilde{t}_s t_s^T$$

The set of vectors $\{t_s\}_{s=1}^{N}$ constitutes an orthonormal basis for R^N.

The unit ball

$$S = \{x \mid \|x\|_2 = 1\}$$

can be represented as

$$(49) \qquad x = \sum_{k=1}^{N} \alpha_k t_k$$

with $\alpha_k = t_s^T x$ and $\sum_{k=1}^{N} \alpha_k^2 = 1$.

Thus for $x \in S$ we have

$$(50) \qquad Cx = \sum_{s=1}^{N} \beta_s \tilde{t}_s t_s^T x$$

$$= \sum_{s=1}^{N} \beta_s \tilde{t}_s \alpha_s = \tilde{Y} v$$

where

$$\tilde{Y} = [\tilde{t}_1 \ \tilde{t}_2 \ \cdots \ \tilde{t}_N]$$

$$v^T = [\beta_1\alpha_1, \ \beta_2\alpha_2, \ \ldots, \ \beta_N\alpha_N]$$

Note that

$$\|\underline{v}\|_2^2 = \sum_{i=1}^{N} \beta_i^2 \alpha_i^2 \leq \beta_{max}^2 \sum_{i=1}^{N} \alpha_i^2 = \beta_{max}^2$$

so $\|\underline{v}\|_2 \leq \beta_{max} \leq 2$ and

$$\|C\underline{x}\|_2 \leq 2\|\tilde{Y}\|_2$$

Thus

$$(51) \qquad \|C\|_2 = \max_{\|\underline{x}\|_2 = 1} \|C\underline{x}\|_2 \leq 2\|\tilde{Y}\|_2$$

But

$$\tilde{Y} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1N} \\ \cos\frac{\pi}{N+1}\, t_{21} & \cos\frac{2\pi}{N+1}\, t_{22} & & \cos\frac{N\pi}{N+1}\, t_{2N} \\ t_{31} & t_{32} & & t_{3N} \\ \vdots & \vdots & & \vdots \\ \cos\frac{\pi}{N+1}\, t_{N-1,1} & \cos\frac{2\pi}{N+1}\, t_{N-1,2} & & \cos\frac{N\pi}{N+1}\, t_{N-1\,N} \\ t_{N1} & t_{N2} & & t_{NN} \end{bmatrix}$$

so for $\underline{z} \in S$ we get

$$\underline{b} = \tilde{Y}\underline{z}$$

with

$$b_i = \begin{cases} \sum_{s=1}^{N} t_{is}\, z_s & i \text{ odd} \\ \sum_{s=1}^{N} t_{is}\, z_s \cos\frac{s\pi}{N+1} & i \text{ even} \end{cases}$$

and

$$b_i^2 = \begin{cases} \left(\sum_{s=1}^{N} t_{is}\, z_s\right)^2 \le \sum_{s=1}^{N} t_{is}^2 \sum_{s=1}^{N} z_s^2 = 1 & i \text{ odd} \\ \left(\sum_{s=1}^{N} t_{is}\, z_s \cos\frac{s\pi}{N+1}\right)^2 \le 1 & i \text{ even} \end{cases}$$

Thus

$$(52) \qquad \|\tilde{Y}\|_2 = \max_{\|z\|_2=1} \|\tilde{Y}z\|_2 = \max_{\|z\|_2=1} \|b\|_2 = \max_{\|z\|_2=1} \sqrt{\sum_{i=1}^{N} b_i^2} \le \sqrt{N}$$

and from (47), (48), (51) and (52)

$$\|(K - M)K^{-1}\|_2 \le \frac{h^2}{12}\sqrt{N}$$

This result is sharper than the result of Section 5.2. However, the numerical experiments do not show any dependence of the dimension N of the matrices. In fact, the experiments indicate that

$$\|(K - M)K^{-1}\|_2 \le \frac{h^2}{12}$$

if $h < 1$. We are not able to prove such a bound at this time.

Appendix

The matrix

$$M = \{a(\phi_i, \psi_j)\} = M_1 + M_2$$

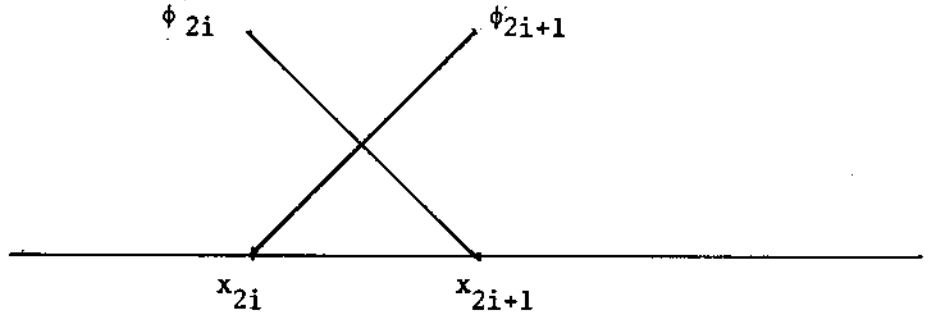
with

$$M_1 = (\phi_i', \psi_j') \qquad M_2 = (\phi_i, \psi_j)$$

are constructed by elementary integration.

1. The element matrices

We have to consider intervals $[x_{2i}, x_{2+1}]$ and $[x_{2i+1}, x_{2i+2}]$ separately. First consider an interval $[x_{2i}, x_{2i+1}]$. On this interval the basis functions graphed in figure 6. are the only ones that are non-zero.



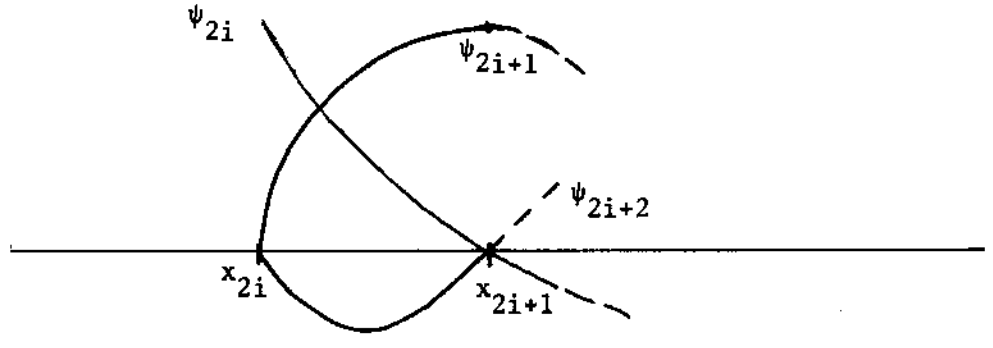


Figure 6. Non-zero basis functions on $[x_{2i}, x_{2i+1}]$.

Introducing a local labelling of the basis functions according to figure 7 and a change of coordinate system so $(x_{2i}, 0)$ becomes the origin and $t = \frac{x - x_{2i}}{h}$ a local coordinate, $h = x_{2i+1} - x_{2i}$, we get

$$\int uv\, dx = h \int uv\, dt \qquad \int u'\, v'\, dx = \frac{1}{h} \int \dot{u}\, \dot{v}\, dt$$

where

$$' = \frac{d}{dx} \quad \text{and} \quad \cdot = \frac{d}{dt} .$$

and

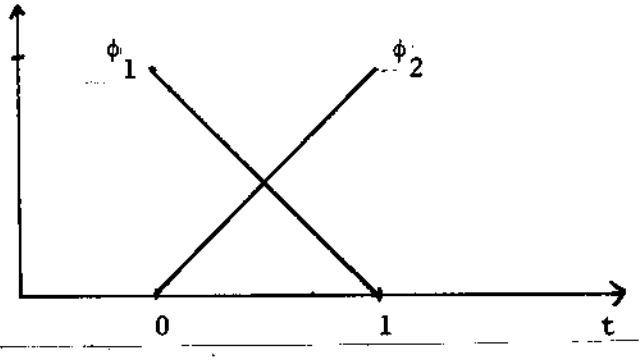


Figure 7a: Local labelling

$$\phi_1(t) = 1 - t \qquad \dot{\phi}_1(t) = -1$$

$$\phi_2(t) = t \qquad \dot{\phi}_2(t) = 1$$

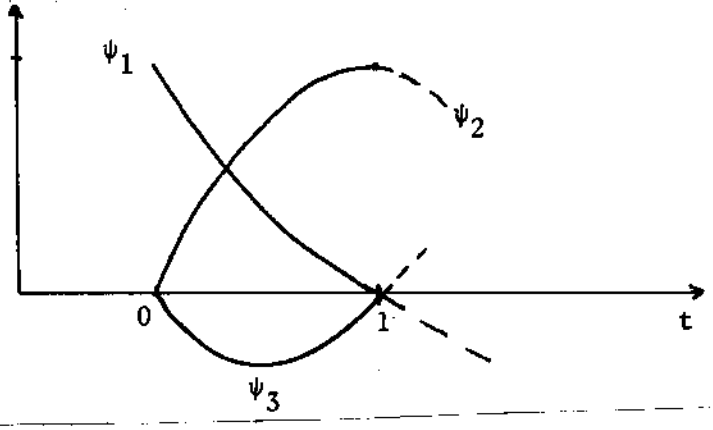


Figure 7b: Local labelling

$$\psi_1 = 1 - \frac{3}{2}t + \frac{1}{2}t^2 \qquad \dot{\psi}_1 = -\frac{3}{2} + t$$

$$\psi_2 = 2t - t^2 \qquad \dot{\psi}_2 = 2 - 2t$$

$$\psi_3 = \frac{1}{2}t^2 - \frac{1}{2}t \qquad \dot{\psi}_3 = t - \frac{1}{2}$$

Straightforward calculations give the element-matrices

$$E_1 = \{(\dot{\phi}_i, \dot{\psi}_j)\} \qquad 2 \times 3$$

$$E_2 = \{(\phi_i, \dot{\psi}_j)\} \qquad 2 \times 3$$

where

$$E_1 = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{bmatrix}; \quad E_2 = \frac{1}{24}\begin{bmatrix} 7 & 6 & -1 \\ 3 & 10 & -1 \end{bmatrix}$$

Now consider an interval $[x_{2i+1}, x_{2i+2}]$. On this interval the basis functions graphed in figure 8 are the only ones that are non-zero.

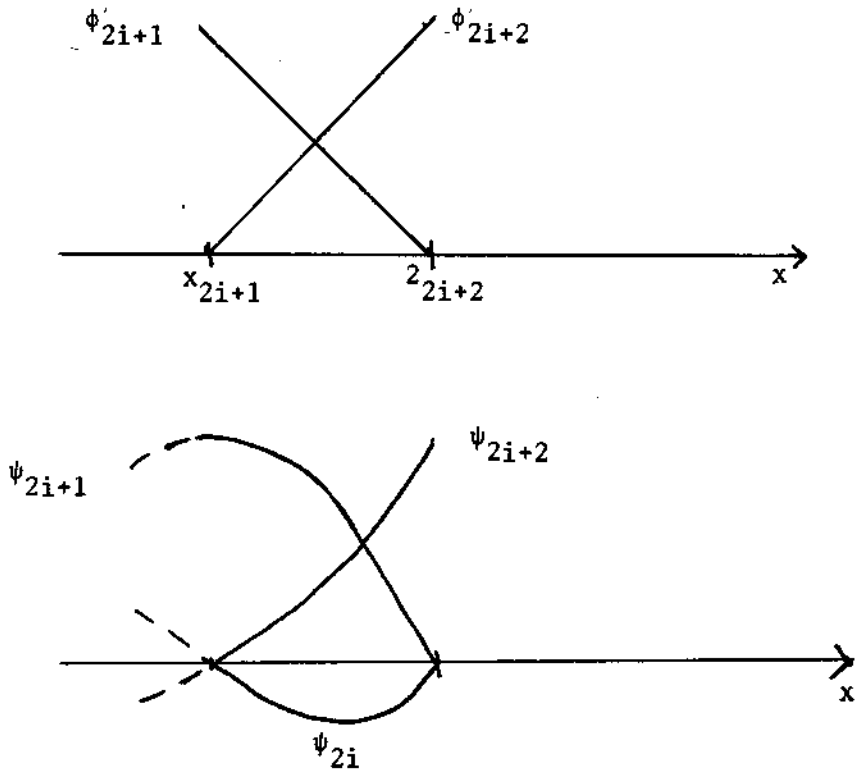


Figure 8: Non-zero basis functions on $[x_{2i+1}, 2_{2i+2}]$

With the local labelling of figure 9 and $t = \frac{x - x_{2i+1}}{h}$, $h = x_{2i+2} - x_{2i+1}$, we get

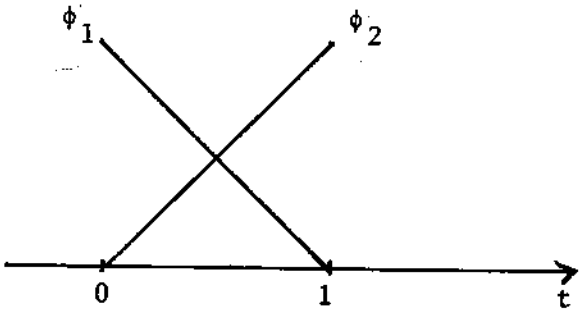


Fig. 9a: Local labelling

$\phi_1 = 1 - t$ $\qquad$ $\dot{\phi}_1 = -1$

$\phi_2 = t$ $\qquad$ $\dot{\phi}_2 = 1$

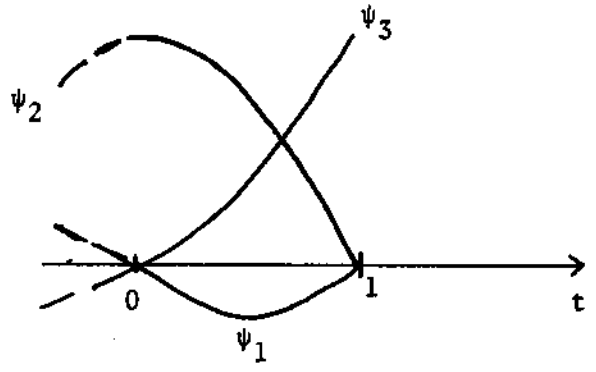


Figure 9b: Local labelling

$$\psi_1 = \frac{1}{2}t^2 - \frac{1}{2}t \qquad \dot{\psi}_1 = t - \frac{1}{2}$$

$$\psi_2 = 1 - t^2 \qquad \dot{\psi}_2 = -2t$$

$$\psi_3 = \frac{1}{2}t^2 + \frac{1}{2}t \qquad \dot{\psi}_3 = t + \frac{1}{2}$$

Straight-forward calculations give the element-matrices

$$H_1 = \{(\dot{\phi}_i, \dot{\psi}_j)\} \qquad 2 \times 3$$

$$H_2 = \{(\phi_i, \psi_j)\} \qquad 2 \times 3$$

where

$$H_1 = \begin{bmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix} \quad ; \quad H_2 = \frac{1}{24}\begin{bmatrix} -1 & 10 & 3 \\ -1 & 6 & 7 \end{bmatrix}$$

2. Assembly of M_1 and M_2

Consider first a row with odd index number, i.e., a row corresponding to ϕ_{2i+1}. Then $a(\phi_{2i+1}, \sum u_v \psi_v)$ will get contributions from the two intervals $[x_{2i}, x_{2i+1}]$ and $[x_{2i+1}, x_{2i+2}]$. The contributions will be from the three basis functions denoted by their local labels in figure 10

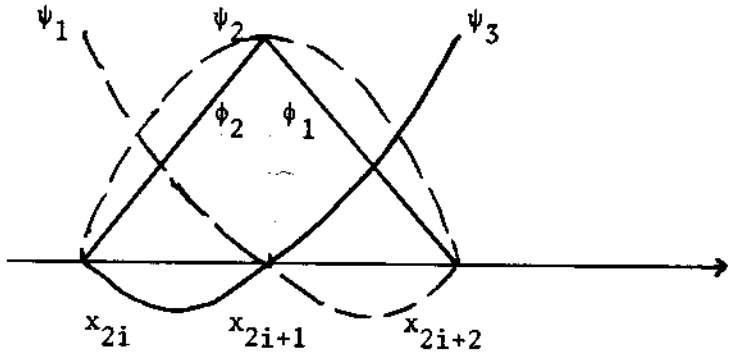


Fig. 10: Local labelling for $[x_{1i}, x_{2i+2}]$

From the figure we get

$$\psi_1 \equiv \psi_{2i}, \quad \psi_2 \equiv \psi_{2i+1}, \quad \psi_3 = \psi_{2i+2}$$

From the interval $[x_{2i}, x_{i+1}]$ the second rows of E_1 and E_2 correspond to the matrix elements $(2i+1, 2i)$, $(2i+1, 2i+1)$ and $(2i+1, 2i+2)$.

From the interval $[x_{2i+1}, x_{2i+2}]$ the first rows of H_1 and H_2 correspond to the matrix elements

$$(2i+1, 2i), \quad (2i+1, 2i+1), \quad (2i+1, 2i+2).$$

Putting these together, remembering the transformation from global to local coordinates we get for row $2i+1$ of M_1

$$\frac{1}{h_{2i+2}}[-1 \quad 1 \quad 0] + \frac{1}{h_{2i+2}}[0 \quad 1 \quad -1]$$

Column: $\quad 2i \quad 2i+1 \quad 2i+2 \qquad 2i \quad 2i+1 \quad 2i+2$

$$= \frac{1}{h_{2i+2}}[\quad -1 \quad 2 \quad -1]$$

and for row $2i+1$ of M_2

$$\frac{h_{2i+2}}{24}[3 \quad 10 \quad -1] + \frac{h_{2i+2}}{24}[-1 \quad 10 \quad 3]$$

$$= \frac{h_{2i+2}}{24}[2 \quad 20 \quad 2]$$

Column: $\quad 2i \quad 2i+1 \quad 2i+2$

Now consider a row with even index number, i.e., a row corresponding to ϕ_{2i}. Then $a(\phi_{2i}, \sum u_v \psi_v)$ will get contributions from the two intervals $[x_{2i-1}, x_{2i}]$ and $[x_{2i}, x_{2i+1}]$. The local labelling for these two intervals is not the same as can be seen from figure 11a. The correspondence to the global labelling can be seen from figure 11b.

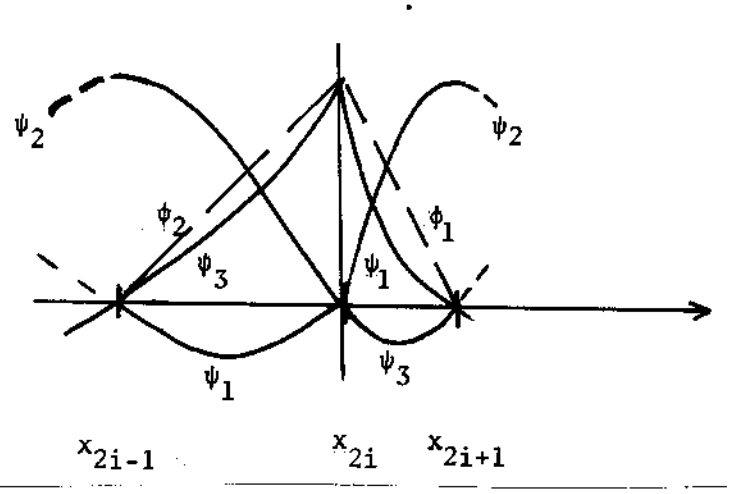


Figure 11-a: Local labellings for $[x_{2i-1}, x_{2i+1}]$

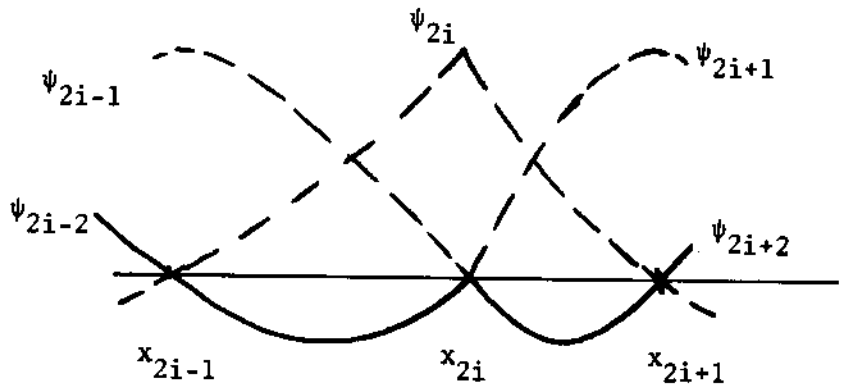


Figure 11-b: Global labelling

From the interval $[x_{2i-1}, x_{2i}]$ the second rows of H_1 and H_2 correspond to the matrix elements

$$(2i, 2i-2), \quad (2i, 2i-1) \quad \text{and} \quad (2i, 2i).$$

From the interval $[x_{2i}, x_{2i+1}]$ the first rows of E_1 and E_2 correspond to the matrix elements

$$(2i, 2i), \quad (2i, 2i+1) \quad \text{and} \quad (2i, 2i+2)$$

Putting these together we get for row $2i$ of M_1

$$\frac{1}{h_{2i}} [0 \quad -1 \quad 1] + \frac{1}{h_{2i+2}} [1 \quad -1 \quad 0]$$

Column $\quad 2i-2 \quad 2i-1 \quad 2i \qquad\qquad 2i \quad 2i+1 \quad 2i+2$

$$= [- \frac{1}{h_{2i}} \quad \frac{1}{h_{2i}} + \frac{1}{h_{2i+2}} \quad - \frac{1}{h_{2i+2}}]$$

and for row $2i$ of M_2

$$\frac{h_{2i}}{24}\,[-1 \quad 6 \quad 7] \;+\; \frac{h_{2i+2}}{24}\,[7 \quad 6 \quad -1]$$

Column $\quad 2i-2 \quad 2i-1 \quad 2i \qquad\qquad 2i \quad 2i+1 \quad 2i+2$

$$= \frac{1}{24}\,[-h_{2i} \quad 6h_{2i} \quad 7(h_{2i}+h_{2i+2}) \quad 6h_{2i+2} \quad -h_{2i+2}]$$

Column $\quad 2i-2 \quad 2i-1 \qquad\qquad 2i \qquad\qquad 2i+1 \quad 2i+2$

Assembling these we get

$$M_1 = \begin{bmatrix} \frac{2}{h_2} & -\frac{1}{h_2} & & & & \\ -\frac{1}{h_2} & \frac{1}{h_2}+\frac{1}{h_4} & -\frac{1}{h_4} & & & \\ & -\frac{1}{h_4} & \frac{2}{h_4} & -\frac{1}{h_4} & & \\ & \ddots & \ddots & \ddots & & \\ & & -\frac{1}{h_{N+1}} & \frac{2}{h_{N+1}} & & \end{bmatrix}$$

and

$$M_2 = \frac{1}{24}\begin{bmatrix} 20h_2 & 2h_2 & & & & \\ 6h_2 & 7(h_2+h_4) & 6h_4 & -h_4 & & \\ & 2h_4 & 20h_4 & 2h_4 & & \\ & -h_4 & 6h_4 & 7(h_4+h_6) & 6h_6 & -h_6 \\ & & \ddots & \ddots & \ddots & \ddots \\ & & -h_{N-1} & 6h_{N-1} & 7(h_{N1}+h_{N+1}) & 6h_{N+1} \\ & & & & 2h_{N+1} & 20h_{N+1} \end{bmatrix}$$

References

Lindberg, B. (1980). Error estimation and iterative improvement for discretization algorithms, B1T20, 486-500.

Strang, G, Fix, G. (1973). An analysis of the finite element method. Prentice-Hall, Inc., Englewood Cliffs, N.J.