ON THE STRUCTURE OF THE MOVING FINITE ELEMENT EQUATIONS

A.J. Wathen and M.J. Baines

NUMERICAL ANALYSIS REPORT NO. 5/83

ABSTRACT

The Moving Finite Element Method for evolutionary partial differential equations leads to a coupled non-linear system of ordinary differential equations in time, with a coefficient matrix A, say, for the time derivatives. We show that for linear elements in any number of dimensions, A can be written in the form M^TCM, where the matrix C depends solely on the mesh geometry and the matrix M on the gradient of the solution. As a simple consequence we show that A is singular only in the cases (i) element degeneracy ($|C| = 0$) and (ii) collinearity of nodes (M not of full rank). We give constructions for the inversion of A in all cases.

In one dimension, if A is non-singular, it has a simple explicit inverse. If A is singular we replace it by a reduced matrix A*. We show that in every case the spectral radius of the Jacobi iteration matrix is $\frac{1}{2}$ and that A or A* can be efficiently inverted by conjugate gradient methods.

Finally, we discuss the applicability of these arguments to systems of equations in any number of dimensions.

are degenerate (in the case $|C| = 0$). We describe in Section 5 procedures in the one-dimensional case for dealing with these situations. If the nodes are collinear the system is replaced by a reduced system with a matrix of similar type The inversion of the reduced matrix may still be readily obtained by iteration, since the spectral radius of the Jacobi iteration matrix for the reduced system remains equal to $\frac{1}{2}$, a result also proved in the Appendix. For a degenerate element the system decouples into two separate systems which however need an "internal boundary condition" in order to be solvable. In hyperbolic problems this can be provided by associating element degeneracy with the formation of shocks and adding a form of jump condition at the interface. As nodes run into a shock it is always possible to reduce the number of shocked nodes to two, each with the same nodal position but with different amplitudes.

In Section 6 the questions of degeneracy of the MFE system and the inversion of the MFE matrix are discussed for higher dimensional problems, with a detailed description given in the two-dimensional case. Although the situation is more complicated here the principles are broadly the same. In higher dimensions, however, it is no longer easy to provide an explicit inverse of the matrix A since M is rectangular in general and in these cases iterative methods, which remain available, are required. Furthermore, there is in the hyperbolic case now no reduction in the number of elements involved in the vicinity of a shock as nodes run into the shock.

For systems of equations, provided that each component is given its own moving finite element mesh, the structure of the MFE equations is unaltered and the results and procedures of the single component system are applicable. A discussion (for the one-dimensional case only) is given in Section 7.

Finally, the content of the paper is reviewed in Section 8, together with a discussion of the effectiveness of the procedures on a number of one-dimensional problems, and a summary of the procedures is given in Section 9.

1. INTRODUCTION

The Moving Finite Element (MFE) method, introduced by Miller & Miller (1981) and Miller (1981), has been used with considerable success for the solution of time dependent partial differential equations in one space dimension. For parabolic problems see, for example, Gelinas Doss & Miller (1981) and Herbst (1982), and for hyperbolic problems, Wathen (1982). Despite this success there is a limited amount of analysis of the method to date, due no doubt to the inherent non-linearity of the method which stems from treating the mesh locations as unknown as well as the nodal parameters (see however Dupont (1982)).

In the present paper we study the coupled system of ordinary differential equations

$$A(\underline{y})\dot{\underline{y}} = \underline{g}(\underline{y}) \qquad (1.1)$$

which arises from the MFE approach and study questions concerning the structure, degeneracy and solution of the system. The matrix $A(\underline{y})$ of (1.1) is positive definite except in particular circumstances, and the strategy in the papers mentione above is to avoid such situations by special devices. The approach in the present paper is rather to analyse the singularities of A in these situations and to provide special inversion procedures in such cases.

After a section recalling the features of the Moving Finite Element method which we need, we discuss in Section 3 the structure of A in the case of linear elements in any number of dimensions, showing in particular that A may be decomposed into the form M^TCM, where the matrices M and C are block diagonal (although with different orderings). From this decomposition we see clearly the sources of singularity of A. In the one-dimensional case with A non-singular, we can write down the explicit inverse of A and this is shown in Section 4. Moreover we can show that the spectrum of the Jacobi matrix is particularly simple, with a spectral radius of $\frac{1}{2}$. This result, which is proved in the Appendix, is useful when considering various iterative procedures for inverting A.

If A is singular there are two sources of singularity: either the nodes are collinear (in which case M is not of full column rank) or the elements

2. THE MOVING FINITE ELEMENT METHOD

We begin by describing the essentials of the MFE method in one space dimension, using mainly the notation of Gelinas, Doss & Miller (1981). Given the scalar evolutionary problem

$$u_t - L(u) = 0, \tag{2.1}$$

where L is a differential operator in space[+], together with boundary conditions, we seek a semi-discrete approximate solution of the form

$$v(x,t) = \sum_{j=0}^{N+1} a_j(t)\,\alpha_j\{x,\underline{s}(t)\} \quad , \tag{2.2}$$

where $\underline{a} = (a_0,a_1,\ldots,a_{N+1})$ is a vector of finite element nodal amplitudes, $\underline{s} = (s_0,s_1,\ldots,s_{N+1})$ is a vector of (time-dependent) nodal positions and α_j is any standard finite element basis function on the grid defined by $\underline{s}$. If $\underline{s}$ were a fixed vector, independent of time, (2.2) would be the usual finite element approximation but in the MFE method $\underline{s}$ is allowed to vary with time and is determined along with $\underline{a}$. A Dirichlet-type boundary condition at the left hand end point s_0 is modelled by fixing s_0 and a_0. Similarly, a Neumann-type condition is modelled by fixing s_0 but allowing a_0 to vary. For definiteness we shall assume Dirichlet-type conditions at both ends s_0 and s_{N+1} throughout this section. The function $v(x,t)$ in (2.2) thus has $2N$ degrees of freedom.

Partial differentiation of (2.2) with respect to time yields

$$v_t(x,t) = \sum_{j=1}^{N} \left[\dot{a}_j(t)\,\frac{\partial v}{\partial a_j} + \dot{s}_j(t)\,\frac{\partial v}{\partial s_j} \right]$$

$$= \sum_{j=1}^{N} \left[\dot{a}_j\alpha_j(x,\underline{s}(t)) + \dot{s}_j\beta_j(x,\underline{a}(t),\underline{s}(t)) \right], \tag{2.3}$$

where

$$\beta_j(x,\underline{a}(t),\underline{s}(t)) = \frac{\partial v}{\partial s_j} \tag{2.4}$$

and the dot denotes time differentiation.

The function β_j may be regarded as a second type of basis function, auxiliary to α_j. In fact both have the same support, and indeed for linear elements it can be shown by direct differentiation or by arguments similar to those of Lynch (1982), that

$$\beta_j = - v_x\alpha_j \tag{2.5}$$

In the case of

+ The class of operators to which L must belong is limited only by the operations subsequently performed on $L(u)$.

linear elements, there is an advantage in regarding each as combinations of base functions having support on a single element (see below).

Equations determining the 2N unknown parameters a_j, s_j $(j=1,2,\ldots,N)$ are obtained by minimising the L_2 norm of the residual, namely

$$\| v_t - L(v) \|_{L_2} , \tag{2.6}$$

with respect to the time derivatives of the parameters $\dot{a}_j, \dot{s}_j$ $(j=1,2,\ldots,N)$. This gives the set of 2N equations

$$\begin{aligned} \langle v_t - L(v),\ \alpha_j \rangle &= 0 \\ \langle v_t - L(v),\ \beta_j \rangle &= 0 \end{aligned} \tag{2.7}$$

$(j=1,2,\ldots,N)$, where $\langle \cdot\ , \cdot \rangle$ is the L_2 inner product. Substituting for v_t from (2.3) we obtain the set of MFE equations

$$A(\underline{y})\dot{\underline{y}} = \underline{g}(\underline{y}) , \tag{2.8}$$

where

$$\underline{y} = \{a_1, s_1; a_2, s_2; \ldots; a_N, s_N\}^T , \tag{2.9}$$

$A(\underline{y})$ is the MFE matrix, which is square and symmetric and consisting of inner products of the basis functions α and β in 2 x 2 blocks,

$$\begin{pmatrix} \langle \alpha_i, \alpha_j \rangle & \langle \alpha_i, \beta_j \rangle \\ \langle \beta_i, \alpha_j \rangle & \langle \beta_i, \beta_j \rangle \end{pmatrix} \tag{2.10}$$

and $\underline{g}(\underline{y})$ is a vector arising from the terms involving $L(v)$ in (2.7).

Inversion (where possible) of the matrix A in the non-linear system of differential equations (2.8) yields explicit expressions for $\dot{a}_j$ and $\dot{s}_j$ (i.e. $\dot{\underline{y}}$) in terms of $\underline{y}$ and $\underline{g}(\underline{y})$. We shall here be concerned with two main issues: first, the question of the existence of the inverse of $A(\underline{y})$ and, second, the construction of that inverse.

Note that, from (2.3) and (2.10),

$$\| v_t \|^2_{L_2} = \dot{\underline{y}}^T A \dot{\underline{y}} . \tag{2.11}$$

It follows that the matrix A is positive semi-definite and singular only when $\dot{\underline{y}} \neq 0$ exists such that $v_t = 0$. To see this, let $\lambda_i, \underline{e}_i$ be the eigenvalues and eigenvectors of the symmetric matrix A, for which $\underline{e}_i$ can be taken as a complete orthonormal set. Then, for any $\dot{\underline{y}} \neq 0$ there exist coefficients γ_i, not all zero such that

$$\dot{\underline{y}} = \sum_{i=1}^{2N} \gamma_i \underline{e}_i \quad ; \tag{2.12}$$

then, from (2.11),

$$\| v_t \|_{L_2}^2 = \dot{\underline{y}}^T A \dot{\underline{y}} = \sum_{i=1}^{2N} \gamma_i^2 \lambda_i \tag{2.13}$$

vanishes only if $\lambda_i = 0$ for at least one i. In that case A must be singular. In Section 5 we shall characterise the situations when A is singular and give procedures for determining $\dot{\underline{y}}$ in such cases.

For higher dimensions we take $\underset{\sim}{r}$ to be the position vector of a point and seek an approximate solution of equation (2.1) in the form

$$v(\underset{\sim}{r},t) = \sum_{j=1}^{N+B} a_j(t)\alpha_j \{\underset{\sim}{r}, \underline{\underset{\sim}{s}}(t)\} \quad , \tag{2.14}$$

where

$$\underline{\underset{\sim}{s}} = \{\underset{\sim}{s}_1, \underset{\sim}{s}_2, \ldots, \underset{\sim}{s}_{N+B}\}^T \tag{2.15}$$

contains the nodal position vectors $\underset{\sim}{s}_j$, a_j is a nodal parameter and α_j is a finite element basis function. Here N is the number of internal nodes and B the number of boundary nodes. Any of these nodes may have any of the variables of $\underset{\sim}{s}_j$ and/or a_j constrained to be held fixed, and modelling of boundary conditions is achieved by applying such constraints appropriately. In the following we assume that all of the boundary nodes are held fixed as are the amplitudes at these nodes - this models Dirichlet conditions over a fixed domain. (Note that it may be appropriate, for example with Dirichlet conditions on a square in two-dimensions, to allow boundary nodes to move along the sides of the square, but not into the interior - this is easily achieved). The function $v(\underset{\sim}{r},t)$ thus has $N(1+d)$ degrees of freedom, where d is the dimension of the physical space.

Partial differentiation of (2.14) gives

$$v_t(\underset{\sim}{r},t) = \sum_{j=1}^{N} [\dot{a}_j(t) \frac{\partial v}{\partial a_j} + \dot{\underset{\sim}{s}}_j(t) \cdot \nabla_{\underset{\sim}{s}_j} v]$$

$$= \sum_{j=1}^{N} \left[\dot{a}_j \alpha_j\{\underset{\sim}{r},\underset{\approx}{s}(t)\} + \sum_{m=1}^{d} \dot{s}_{jm}(t)\beta_{jm}\{\underset{\sim}{r},\underline{a}(t),\underset{\approx}{s}(t)\}\right] \qquad (2.16)$$

where $\dot{s}_{jm}(t)$ is the m'th component of $\dot{\underset{\sim}{s}}_j(t)$ and

$$\beta_{jm} = \frac{\partial v}{\partial s_{jm}} \qquad (2.17)$$

is one of d second type basis functions (see (2.4)). As in the one-dimensional case the support of each β_{jm} is the same as that of α_j. The arguments of Lynch (1982) can again be used to show that

$$\beta_{jm} = - \frac{\partial v}{\partial x_m} \alpha_j \qquad (2.18)$$

where we have taken the $\underset{\sim}{r}$ of (2.14) as

$$\underset{\sim}{r} = (x_1 , x_2 , \ldots , x_d) . \qquad (2.19)$$

Minimisation of the L_2 norm (2.6) gives rise to the N(1+d) equations

$$\begin{aligned} \langle v_t - L(v), \alpha_j \rangle &= 0 \\ \langle v_t - L(v), \beta_{jm} \rangle &= 0 \end{aligned} \qquad (2.20)$$

(m = 1,2,...,d),(j = 1,2,...,N). Substituting for v_t from (2.16) then gives the set of MFE equations

$$A(\underline{y})\dot{\underline{y}} = \underline{g}(\underline{y}) \qquad (2.21)$$

where

$$\underline{y} = \{a_1, \underset{\sim}{s}_1; a_2, \underset{\sim}{s}_2; \ldots; a_N, \underset{\sim}{s}_N\}^T \qquad (2.22)$$

and $\underline{g}(\underline{y})$ arises from the terms containing L(v) in (2.20). The matrix A is still square and symmetric consisting of inner products of the α's and β's in blocks. Note that the form (2.11) is preserved.

3. LINEAR ELEMENTS

We now specialise to linear elements in any number of dimensions with nodes at the vertices, e.g. triangles in two dimensions, tetrahedra in three. Thus we choose the function α_j to be a local basis function with the value 1 at node j, zero at its neighbours, being linear in between: elsewhere it is zero. The components of the corresponding $\underset{\sim}{\beta}_j$ are given by (2.17) or (2.18) where v is piecewise linear. Hence each component of $\underset{\sim}{\beta}_j$ is also linear on each element and has the same support as α_j but it is discontinuous, being a different multiple of α_j on each element of its support.

In one dimension α_i is the hat function shown in Fig. 3.2, the approximate solution v is piecewise linear as shown in Fig. 3.1 and the second type basis function β_i is as shown in Fig. 3.3. The actual definitions are given in §4.

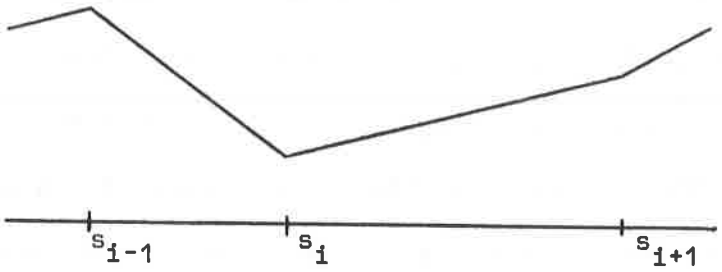

Figure 3.1

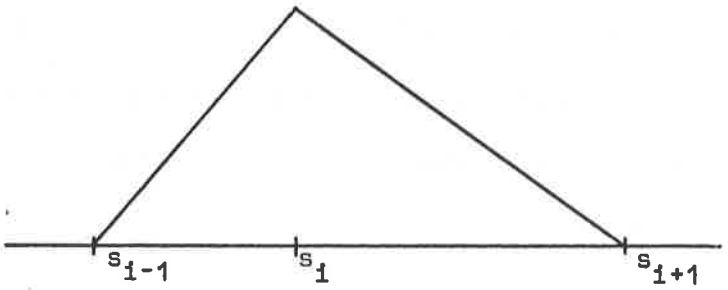

Figure 3.2

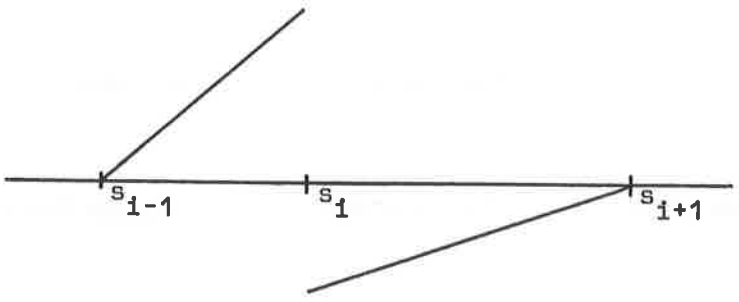

Figure 3.3

Now the matrix A consists of blocks (2.10), which in the present case will vanish except when i=j or i is a neighbour of j. If i is a neighbour of j the integration in (2.10) is over the elements common to the support of α_i and α_j, and the corresponding block is

$$A_{ij} = \begin{bmatrix} \langle \alpha_i, \alpha_j \rangle & \langle \alpha_i, \underset{\sim}{\beta}_j \rangle \\ \langle \underset{\sim}{\beta}_i, \alpha_j \rangle & \langle \underset{\sim}{\beta}_i, \underset{\sim}{\beta}_j \rangle \end{bmatrix} \tag{3.1}$$

which is singular of rank 1 in one dimension and of rank 1 or 2 in two dimensions. If i=j the integration is over a patch of elements surrounding the node i . Also as in the usual finite element assembly, there exists a number δ such that

$$\sum_{j=1}^{N} A_{ij} = \delta A_{ii} \qquad (3.2)$$

(except possibly when i is a boundary node): for example, $\delta = \frac{1}{2}$ in the one-dimensional case .

When the inverse of A_{kk} exists for each k we can define a matrix with (k,ℓ)'th block

$$A_{kk}^{-1} A_{k\ell} \qquad (3.3)$$

and prove some remarkable results concerning its eigenvalues. We shall in fact prove these results (in the Appendix) for the related block Jacobi iteration matrix

$$J_{k\ell} = \begin{cases} -A_{kk}^{-1} A_{k\ell} & k \neq \ell \\ 0 & k = \ell \end{cases} \qquad (3.4)$$

In one dimension, for A non-singular, we show that the eigenvalues of J are precisely $\pm\frac{1}{2}$ and 0+. Moreover, if A is singular, the result $\rho(J^*) = \frac{1}{2}$ can be proved for the spectral radius ρ of the modified matrix A* and the corresponding J* introduced in §5. These results may be used directly to ensure rapid convergence of an iterative method such as SOR iteration with the optimal relaxation parameter

$$w_{opt} = \frac{2}{[1+[1-\{\rho(J)\}^2]^2]^{\frac{1}{2}}} = 8 - 4\sqrt{3} \qquad (3.5)$$

(see Varga, 1962) with a convergence factor better than 0.072. Alternatively, the clustering of the eigenvalues suggests a conjugate gradient method: in particular we may use the generalised conjugate gradient method of Concus, Golub & O'Leary (19).

In higher dimensions Jacobi iteration may not converge since, from (3.2), J may possess an eigenvalue $-\delta$ (depending on the boundary conditions) with $|\delta| \geq 1$. However, since A is symmetric, positive-definiteness is sufficient to guarantee convergence of the SOR iteration. In two dimensions (when δ=1) there is numerical evidence that the eigenvalue spectrum of J is contained in the interval

---

+ There may be 0,1 or 2 zero eigenvalues, depending on the end conditions.

$$[-1, \tfrac{1}{2}] \qquad (3.6)$$

so that the generalised conjugate gradient method applied to the matrix A may remain appropriate.

The form of $\underset{\sim}{\beta}$ suggests another formulation of the MFE system and a slightly different approach. Re-parameterise the function $v_t(\underset{\sim}{r},t)$ of (2.16) in the form

$$v_t(\underset{\sim}{r},t) = \sum_{k=1}^{n} \sum_{\nu=1}^{d+1} \dot{w}_{k\nu}(t)\phi_{k\nu}(\underset{\sim}{r},\underset{\sim}{s}(t)) \qquad (3.7)$$

where each $\phi_{k\nu}$ is a discontinuous basis function with support on a single element k defined as follows : for each element k take $\phi_{k\nu}$ to be 1 at a vertex ν of the element and zero at the other vertices and to be linear in between; elsewhere it is zero. Thus for each element there are the same number of $\phi_{k\nu}$'s as there are vertices and the total number of $\phi_{k\nu}$'s is the product of the number of elements, n, and the number of vertices, $d+1$, where d is the dimension of the physical space. In order to be consistent with the boundary condition that we have assumed in the sequel to (2.14) and (2.15) we set $\dot{w}_{k\nu} = 0$ whenever the ν^{th} node of element k is one of the boundary nodes with position $\underset{\sim}{s}_j$, $j = N + 1 , \ldots , N + B$. We can thus make the comparison with the number of α_j's (or $\underset{\sim}{\beta}_j$'s), which is equal to the number N of internal nodes, as follows.

From (2.16) and (3.7) we may relate the nodal parameters $\dot{\underline{a}},\dot{\underline{s}}$ to the $\dot{w}_{k\nu}$. Writing

$$\underline{\dot{w}} = (\dot{w}_{11},\dot{w}_{12},\ldots,\dot{w}_{1d+1};\dot{w}_{21},\dot{w}_{22},\ldots,\dot{w}_{2d+1};\ldots;\dot{w}_{n2},\ldots,\dot{w}_{nd+1}) \qquad (3.8)$$

(any terms corresponding to boundary nodes being omitted) and using the $\underline{\dot{y}}$ which is the derivative of (2.22) yields

$$\underline{\dot{w}} = M\underline{\dot{y}} \ , \qquad (3.9)$$

where M is the matrix obtained by writing the $\alpha_j,\underset{\sim}{\beta}_j$ in terms of the

$\phi_{k\nu}$. To see this we use the approach of Lynch (1982). Equation (2.16) can be written as

$$v_t = \sum_{j=1}^{N} (\dot{a}_j \alpha_j + \underset{\sim}{\beta}_j \cdot \dot{\underset{\sim}{s}}_j) = \dot{\underline{y}}^T \underline{\alpha} \tag{3.10}$$

where $\underset{\sim}{\beta}_j = -\nabla v \; \alpha_j$,(see (2.18)), (3.11)

and $\underline{\alpha}^T = (\alpha_1, \underset{\sim}{\beta}_1^T, \ldots, \alpha_N, \underset{\sim}{\beta}_N^T)$. (3.12)

Thus $v_t = \sum_{j=1}^{N} (\dot{a}_j - \nabla v \cdot \dot{\underset{\sim}{s}}_j) \; \alpha_j = \dot{\underline{w}}^T \underline{\phi}$ (3.13)

where $\underline{\phi}^T = \{\phi_{k\nu}\}$ is the vector of element base functions.

So we have

$$v_t = \dot{\underline{y}}^T \underline{\alpha} = \dot{\underline{w}}^T \underline{\phi} = \dot{\underline{y}}^T M^T \underline{\phi}$$

from which it follows

$$\underline{\alpha} = M^T \underline{\phi} \quad . \tag{3.14}$$

Since α_j (or $\underset{\sim}{\beta}_j$) is defined on the support of a local patch of elements around the node j, M has a relatively simple structure (see below). Moreover the coefficients of M will only involve the (constant) slope of the solution in the various elements.

The quadratic form (2.11) now becomes

$$\| v_t \|^2_{L_2} = \dot{\underline{w}}^T C \, \dot{\underline{w}}, \tag{3.15}$$

where the matrix $C = \{c_{k\ell}\}$, with entries $c_{k\ell} = \langle \phi_{k\nu}, \phi_{\ell\mu} \rangle$, (3.16)

is an alternative MFE matrix, also symmetric and positive semi-definite, but element-based as compared to the node-based MFE matrix A. Comparing (3.9), (3.15) and (2.11) we see that

$$A = M^T C M \quad . \tag{3.17}$$

From (3.15),(3.16) and the definition of $\phi_{k\nu}$ it follows that the matrix C is square and block diagonal with the elementwise ordering in the vector $\dot{\underline{w}}$ of

(3.8). The entries in the k^{th} block are a simple scalar multiple of the measure of element k. Also, by the same argument as for the matrix A in §2, C is positive semi-definite and singular only when $\underline{\dot{w}} \neq 0$ exists such that $v_t = 0$; i.e. when an element becomes null. The matrix M, on the other hand, is in general rectangular and represents the mapping from the node-based representation to the element-based representation.

4. ONE DIMENSION - POSITIVE DEFINITE A

We now study in detail the case of linear elements in one dimension. A consequence of each node possessing just two neighbours in one dimension is that, numbering the nodes 1 to N sequentially along the axis, the matrix A of (2.8) is 2x2 block tri-diagonal (see (3.1)). Provided that the diagonal blocks are positive definite, A is invertible. In this section we shall assume this property, leaving to Section 5 the case of singular A.

To invert the matrix A in the present case we could use the iterative methods mentioned in Section 3. Since the eigenvalues of the Jacobi iteration matrix are $\pm \frac{1}{2}$ and 0 (the latter depending on boundary conditions), the exact solution may be obtained from the generalised conjugate gradient method after precisely 2 (or, if there is a zero eigenvalue, 3) iterations. However, in this one-dimensional case, with each element associated with just two nodes, the approach using the $\phi_{k\nu}$'s of (3.7) leads to a matrix M in (3.9) which is square, block 2x2 diagonal (except that possibly the extreme diagonal blocks may be 1x1 - see below) and therefore easily invertible. Since the matrix C of (3.16) is in this case also block 2x2 diagonal (although with blocks staggered with respect to the blocks of M - see below) and neither C nor M is singular (since A is non-singular) we have immediately that

$$A^{-1} = M^{-1}C^{-1}M^{-T} \quad . \tag{4.1}$$

At the end-points of the region there are two distinct cases. If there is a fixed end-point s_0 and fixed end value a_0 the matrix C (which is element-based) will contain a 1x1 diagonal entry at one corner, while M is block 2x2 throughout. Alternatively, if either a_0 or s_0 is free, the matrix M will contain a 1 x 1 diagonal entry at a corner, while C remains block 2x2 throughout. The choice will depend on boundary conditions. The boundary behaviour of A will vary accordingly.

For the calculation of the elements of the matrices we have (see e.g. Gelinas Doss & Miller (1981))

$$\alpha_i = \begin{cases} \dfrac{x-s_{i-1}}{s_i-s_{i-1}} & \left(s_{i-1} \le x \le s_i\right) \\ \dfrac{s_{i+1}-x}{s_{i+1}-s_i} & \left(s_i \le x \le s_{i+1}\right) \\ 0 & \text{otherwise} \end{cases} \qquad (4.2)$$

and

$$\beta_i = \begin{cases} -m_{i-\frac{1}{2}}\ \alpha_i & \left(s_{i-1} \le x < s_i\right) \\ -m_{i+\frac{1}{2}}\ \alpha_i & \left(s_i < x \le s_{i-1}\right) \\ 0 & \text{otherwise}, \end{cases} \qquad (4.3)$$

where

$$m_{i+\frac{1}{2}} = \frac{a_{i+1}-a_i}{s_{i+1}-s_i}$$

is the slope of the solution in the element (i,i+1). Note that the support of α_i and β_i is the same, and that

$$\beta_i = -v_x \alpha_i \qquad (4.4)$$

over the whole of that support (loc. cit.).

Introducing the element numbering $k = i-\frac{1}{2}, (k=\frac{1}{2},\ldots,N+\frac{1}{2})$, the basis function ϕ_{k1} is that part of α_{i-1} in $s_{i-1} < x \le s_i$ (see (4.2)) and ϕ_{k2} is that part of α_i in the same interval.

The diagonal blocks of M are therefore

$$\begin{bmatrix} 1 & -m_{i-\frac{1}{2}} \\ 1 & -m_{i+\frac{1}{2}} \end{bmatrix} \tag{4.5}$$

and M non-singular corresponds to $m_{i-\frac{1}{2}} \neq m_{i+\frac{1}{2}}$ for any i.

The diagonal blocks of C are

$$(s_{i+1}-s_i) \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix} \tag{4.6}$$

and $C \neq 0$ corresponds to $s_{i+1} \neq s_i$. If either $m_{i-\frac{1}{2}} = m_{i+\frac{1}{2}}$ or $s_{i+1} = s_i$, for any i, M or C is singular, A is singular and inversion of A using (4.1) is not possible.

5. ONE DIMENSION - SINGULAR A

We consider now the case of linear elements in one dimension when A is singular. We know that there are just two situations in which this can arise. When $m_{i-\frac{1}{2}} = m_{i+\frac{1}{2}}$, for some i, M is singular and we see from (4.2) and (4.3) that β_i is proportional to α_i. Not only are the off-diagonal blocks of A singular but so is the diagonal block A_{ii} (see (3.1) and (3.2)). This type of degeneracy corresponds to two adjacent slopes in the solution being equal, i.e. a node and its two neighbours collinear. We shall refer to such degeneracy as "parallelism" (at a node).

If $s_{i+1} = s_i$, then C is singular as a result of a complete block vanishing and A_{ii} is again singular. This kind of degeneracy, which corresponds to node overtaking (as in shock formation), we shall term "shock" type degeneracy.

Consider parallelism first. Let $m_{i-\frac{1}{2}} = m_{i+\frac{1}{2}} = m_i$, say, for some particular i. From the fact (4.3) that $\beta_i(x) = -m_i \alpha_i(x)$ it follows that the two MFE equations (2.7) are linearly dependent for this particular i. Hence $A(\underline{y})$ is singular with a null space spanned by the single vector

$$\underline{u}_i = \{0,0;0,0;\ldots;\overbrace{m_i,1}^{i'\text{th block}};\ldots;0,0\}^T. \tag{5.1}$$

We may still obtain $\dot{\underline{y}}$ from (2.8) as follows: delete the equation which arises from the inner product (2.7) involving β_i and replace it by the equation

$$\dot{s}_i = 0 \quad . \tag{5.2}$$

Without loss of generality we may now delete the corresponding column of A giving (together with (5.2)) the new system

$$A^*\dot{\underline{y}}^* = \underline{g}^*(\underline{y}^*) \tag{5.3}$$

corresponding to the MFE method with (from (5.2)) one fixed node. The solution of the system (5.3) will be a particular solution of the system (2.8), since the introduction of (5.2) maintains the consistency of the system.

A^* is the matrix $\{A^*_{jk}\}$ where $A^*_{jk} = A_{jk}$ except in a single row and column corresponding to $\dot{s}_i$, where all entries are zero except the diagonal entry, which is 1. If there is parallelism at only one node, so that A^* is non-singular, then by solving (5.3) for $\dot{\underline{y}}^*$ we may obtain the general solution of (2.8) by adding a multiple of the vector $\underline{u}_i$ of (5.1) spanning the null space, giving

$$\dot{\underline{y}} = \dot{\underline{y}}^* + c\underline{u}_i \ , \tag{5.4}$$

and choosing c to satisfy some external criterion, for example that s_i lies midway between s_{i-1} and s_{i+1}. Whatever c is chosen the residual (2.6) will be unchanged.

Another strategy is to remove the "parallel" node i altogether. Again the residual (2.6) is unaffected but the capacity of the method to follow developing features may be impaired. We have not followed up this strategy.

If there is parallelism at p nodes i_k (k=1,2,...,p), not necessarily consecutive, and if we replace the corresponding equations involving β_{i_k}'s by $\dot{s}_{i_k} = 0$ in each case, the remaining 2N-p equations will be linearly independent, since the rank of A is reduced by just one for each node which is parallel. Moreover, the null space is the union of p, disjoint, one-dimensional subspaces of the form

$$\left\{ c\underline{u}_{i_k} \ : \ \underline{u}_{i_k} = [0,0;0,0;\ldots;\overbrace{m_{i_k},1}^{i_k\text{'th block}};\ldots;0,0]^T \right\} \tag{5.5}$$

and if, as before, $\dot{\underline{y}}^*$ is the solution of the system (5.3), augmented to include the solutions $\dot{s}_{i_k} = 0$, then the general solution is

$$\dot{\underline{y}} = \dot{\underline{y}}^* + \sum_{k=1}^{p} c_k \underline{u}_{i_k} \quad . \tag{5.6}$$

In this approach we cannot exploit the decomposition (3.17) in solving (5.3) since the inclusion of the equation $\dot{s}=0$ as part of the system means that no such decomposition for A^* can be found. The replacement of (5.2) by an equation which does admit such a decomposition is possible but the relationship with the source problem is then less clear. In the present approach we can still use the powerful iteration procedure mentioned before to invert A^* since the spectral radius of the corresponding Jacobi iteration matrix is still $\frac{1}{2}$ (see Appendix).

Turning now to the other source of singularity of A, namely, $s_{i+1} = s_i$, we make use of the procedure for hyperbolic equations introduced in Wathen (1982). Supposing that the differential equation is hyperbolic and taking node overtaking to be indicative of the formation of a shock, we replace two of the equations with a shock condition

$$\dot{s}_{i+1} = \dot{s}_i = \lim \frac{\int_{s_1}^{s_2} L(u)dx}{a_2 - a_1} \tag{5.7}$$

as $a_1 \to a_i$, $s_1 \to s_i = s_{i+1}$ from the left, $a_2 \to a_{i+1}$, $s_2 \to s_i = s_{i+1}$ from the right, allowing a_i and a_{i+1} to take on different values even though $s_{i+1} = s_i$.

Since one of the diagonal submatrices of C now vanishes (see(4.6)), the MFE matrix decouples into two disjoint submatrices, each of which may be inverted provided that an internal boundary condition is imposed on the two points. This condition may be taken to be (5.7). If, after a time step, s_i exceeds s_{i+1}various procedures can be adopted to ensure that $s_{i+1} = s_i$, for example crude averaging or going back and taking a smaller time step.

When a second node attempts to overtake a pair of shocked nodes, such as s_i, s_{i+1} in the above, the same procedure can be used and a node deleted. Thus, for example, if the node s_{i-1} overtook the shocked pair s_i, s_{i+1} we would impose the condition

$$\dot{s}_{i+1} = \dot{s}_{i-1} = \lim \frac{\int_{s_1}^{s_2} L(u)dx}{a_2 - a_1}$$

as $a_1 \to a_{i-1}$, $s_1 \to s_{i-1} = s_{i+1}$ from the left, $a_2 \to a_{i+1}$, $s_2 \to s_{i+1} = s_{i-1}$ from the right, and delete the node s_i entirely. The deletion of the node i

is consistent with the concept of a loss of information in the presence of a shock: in the present MFE description a one-dimensional shock is always represented by precisely two nodes with the same nodal position (see figures 8.1, 8.2 and 8.3).

In non-hyperbolic problems where a shock is not expected, the above is not a feasible procedure: this case will be discussed in a later publication.

6. HIGHER DIMENSIONS

Considering now the MFE matrix A of (2.21) in a space of more than one dimension, we note first that A may still be written in the form (3.17) but that, although the matrix C is still square, the matrix M is now rectangular and does not possess an inverse. However, we can show that the sources of singularity A are still (i) column rank deficiency of M and (ii) singularity of C. For, suppose A is singular, i.e. there exists $\dot{\underline{y}} \neq 0$ such that $A\dot{\underline{y}} = 0$. Then

$$\dot{\underline{y}}^T A \dot{\underline{y}} = 0 \tag{6.1}$$

$$\Rightarrow \quad \dot{\underline{y}}^T M^T C M \dot{\underline{y}} = 0$$

$$\dot{\underline{w}}^T C \dot{\underline{w}} = 0 \tag{6.2}$$

where

$$M\dot{\underline{y}} = \dot{\underline{w}} \,. \tag{6.3}$$

It follows that either $\dot{\underline{w}} = M\dot{\underline{y}} = 0$, in which case M is column rank deficient or, since C is positive semi-definite, C is singular (see §3). These results are consistent with the one-dimensional behaviour in the previous two sections.

If A is non-singular, the system (2.21) may be solved by iteration, since A is symmetric and positive definite and the SOR iteration method, for example, converges. Moreover, if the conjecture in (3.6) is correct, the generalised conjugate gradient will work well: this has been seen in an exploratory example.

Now suppose that A is singular as a result of the matrix M being column rank deficient. We illustrate the form of M in the two-dimensional case; other cases are similar. In two dimensions we may write

$$v_t(\underset{\sim}{r},t) = \sum_{j=1}^{N}\left[\dot{a}_j\alpha_j\,(\underset{\sim}{r},\underset{\sim}{s}(t)) + \dot{X}_j\beta_j\,(\underset{\sim}{r},\underset{\sim}{a}(t),\underset{\sim}{s}(t)) + \dot{Y}_j\gamma_j(\underset{\sim}{r},\underset{\sim}{a},\underset{\sim}{s}(t))\right] \quad (6.4)$$

c.f. (2.16) where $\underset{\sim}{r} = (x,y)$, $\underset{\sim}{s} = \{X_1,Y_1;Y_2,Y_2;\ldots;X_N,Y_N\}$, and X_j,Y_j are the co-ordinates of node j. Also $(\beta_j,\gamma_j) = \underset{\sim}{\beta}_j$, where from (3.11)

$$\beta_j = -v_x\alpha_j, \qquad \gamma_j = -v_y\alpha_j \quad . \quad (6.5)$$

To find the form of M from

$$\dot{\underset{\sim}{w}} = M\dot{\underset{\sim}{y}} \quad (6.6)$$

consider a single element ℓ. At the vertex μ of the element, which is node i, say, we compare v_t in the two bases represented in the equations

$$v_t = \sum_{j=1}^{N}\left(\dot{a}_j\alpha_j + \dot{X}_j\beta_j + \dot{Y}_j\gamma_j\right) = \sum_{k=1}^{n}\sum_{\nu=1}^{3}\dot{w}_{k\nu}\,\phi_{k\nu} \quad (6.7)$$

to obtain

$$\begin{aligned}\dot{w}_{\ell\mu}\,\phi_{\ell\mu} &= \dot{a}_i\alpha_i + \dot{X}_i\beta_i + \dot{Y}_i\gamma_i \\ &= (\dot{a}_i - m_\ell\dot{X}_i - n_\ell\dot{Y}_i)\alpha_i, \end{aligned} \quad (6.8)$$

using (6.5), where $(m_\ell,n_\ell) = (v_x,v_y)$ on element ℓ. (For a boundary node both sides of (6.8) will be zero for assumed Dirichlet conditions). Thus, since $\phi_{\ell\mu} = \alpha_i$ on element ℓ, we have

$$\dot{w}_{\ell\mu} = \dot{a}_i - m_\ell\dot{X}_i - n_\ell\dot{Y}_i \quad (6.9)$$

so that in the row of M corresponding to $\dot{w}_{\ell\mu}$ there is just one 3-vector entry

$$(1, -m_\ell, -n_\ell) = \underset{\sim}{p}_\ell^T \quad , \quad (6.10)$$

say, in the i^{th} blocked column: note that the vector $\dot{\underset{\sim}{y}}$ is ordered as in (2.22), namely

$$\dot{\underset{\sim}{y}} = (\dot{a}_1,\dot{X}_1,\dot{Y}_1;\dot{a}_2,\dot{X}_2,\dot{Y}_2;\ldots;\dot{a}_N,\dot{X}_N,\dot{Y}_N) \quad .$$

For each element k surrounding node i there will be a similar 3-vector entry in the i^{th} column; thus there will be the same number of $\underset{\sim}{p}_k^T$'s in column i as there are elements around node i. Indeed if we order the vector $\dot{\underset{\sim}{w}}$ not

in the elementwise manner of (3.8)(which is shown in Fig. 6.1), but in a nodewise manner (as in Fig. 6.2), then the columns of the matrix M will be interchanged in such a way that the result, N say, is block diagonal. That is, there exists a permutation matrix Q such that

$$N = QM = \begin{bmatrix} \square & & 0 \\ & \ddots & \\ 0 & & \square \end{bmatrix} \qquad (6.11)$$

with rectangular blocks.

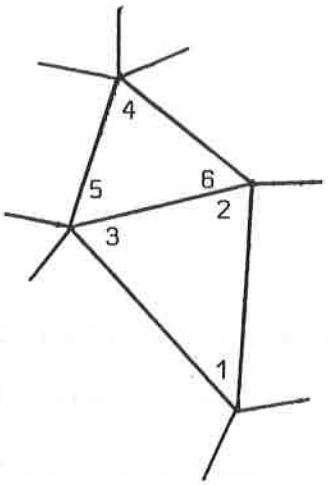

Fig. 6.1
Elementwise numbering

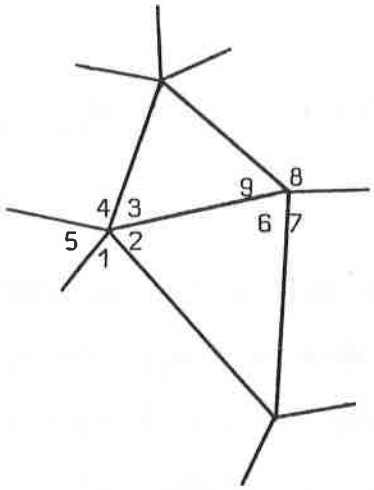

Fig. 6.2
Nodewise numbering

Each block consists of three columns of scalars (corresponding to the components $\underline{p}_K^T$) and the same number of rows as there are elements surrounding a node.

Column rank deficiency of M or N corresponds to row rank deficiency of the $\underline{p}_k^T$. Thus M will be column rank deficient if, either for all pairs of elements k and ℓ adjacent to the node j,

$$m_k = m_\ell \quad \text{and} \quad n_k = n_\ell , \tag{6.12}$$

or if there exist numbers λ , μ , not both zero, such that $\lambda m_k + \mu n_k = \lambda m_\ell + \mu n_\ell$. To deal with this type of degeneracy we return to the MFE matrix A whose elements A_{ij} are now given by (3.1), i.e.

$$A_{ij} = \begin{bmatrix} \langle \alpha_i, \alpha_j \rangle & \langle \alpha_i, \beta_j \rangle & \langle \alpha_i, \gamma_j \rangle \\ \langle \beta_i, \alpha_j \rangle & \langle \beta_i, \beta_j \rangle & \langle \beta_i, \gamma_j \rangle \\ \langle \gamma_i, \alpha_j \rangle & \langle \gamma_i, \beta_j \rangle & \langle \gamma_i, \gamma_j \rangle \end{bmatrix} \tag{6.13}$$

and (3.2) $$A_{ii} = \frac{1}{\delta} \sum_{j \neq i} A_{ij} , \tag{6.14}$$

where $\delta=1$ in the two dimensional case.

When $m_k = m_\ell$, and $n_k = n_\ell$, β_j and γ_j are parallel to α_j in all elements k and ℓ surrounding the node j. In this case there is a unique m and n for all k,ℓ and the solution in the whole patch of elements surrounding the node j is coplanar. The MFE equations $\langle v_t - L(v), \beta_j \rangle = \langle v_t - L(v), \gamma_j \rangle = 0$ are redundant and we omit them, striking out also the corresponding columns of the matrix A. The resulting matrix is non-singular (if this is the only parallelism) and is consistent with the equations (c.f. (5.2)).

$$\dot{X}_j = 0, \qquad \dot{Y}_j = 0. \tag{6.15}$$

The solution of the reduced system

$$A^*(\underline{y}^*)\dot{\underline{y}}^* = \underline{g}^*(\underline{y}^*) \tag{6.16}$$

(c.f. (5.3)), augmented to include (6.15), leads to a complete solution

$$\dot{\underline{y}} = \dot{\underline{y}}^* + \underline{y}_j c \tag{6.17}$$

where $\underset{\sim}{u}_j = (\underset{\sim}{u}_j^{(1)}, \underset{\sim}{u}_j^{(2)})$ is the null vector of the full system, which has components of the form

$$\begin{aligned} \underset{\sim}{u}_j^{(1)} &= \{0,0,0;0,0,0;\ldots;m_j,1,0;\ldots;0,0,0\}^T \\ \underset{\sim}{u}_j^{(2)} &= \{0,0,0;0,0,0;\ldots;n_j,0,1;\ldots;0,0,0\}^T \ , \end{aligned} \tag{6.18}$$

and $\underset{\sim}{c} = (c_1,c_2)^T$ can be chosen to satisfy some external criterion, for example that X_j,Y_j lies at the centroid of the patch of elements surrounding node j (c.f. the argument in Section 5).

When there exist λ, μ, not both zero, such that

$$\lambda m_k + \mu n_k = \lambda m_\ell + \mu n_\ell , \tag{6.19}$$

then $\lambda\beta_j + \mu\gamma_j$ is parallel to α_j in all elements k and ℓ surrounding the node j. The vectors $\underset{\sim}{p}_k = (1, -m_k, -n_k)^T$, for all k surrounding the node j, span a two-dimensional space and the null space is the orthogonal space. The null space may be spanned by

$$\underset{\sim}{v}_j = [0,0,0;0,0,0;\ldots;\underset{\sim}{n};\ldots;0,0,0] \tag{6.20}$$

where $$\underset{\sim}{n} = [m_k(n_k-n_\ell) + (m_k-m_\ell)n_k,\ n_k - n_\ell,\ m_k - m_\ell] \tag{6.21}$$

and k, ℓ are chosen such that $\underset{\sim}{p}_k \neq \underset{\sim}{p}_\ell$. This can always be done, for if not, the stronger parallelism above applies. Geometrically, this latter type of parallelism corresponds to the solution on a patch of elements surrounding a node consisting of just two planes, where two element edges coincide with the line of intersection.

As in the earlier case the solution for $\dot{\underset{\sim}{y}}$ can be written

$$\dot{\underset{\sim}{y}} = \dot{\underset{\sim}{y}}^* + c\underset{\sim}{v}_j \ , \tag{6.22}$$

where $\dot{\underset{\sim}{y}}^*$ is the solution of the modified system, the equation $\langle v_t - L(v), \beta_j \rangle = 0$ being replaced by $\dot{X}_j = 0$ or $\langle v_t - L(v), \gamma_j \rangle$ by $\dot{Y}_j = 0$. The constant c is again chosen to satisfy some external criterion.

Turning now to singularity of A as a result of singularity of C, we refer back to the form of C given by (3.16). Again we illustrate in the two-dimensional case.

Suppose that S_k is the area of the triangular element k and that we employ the elementwise numbering as in Fig. 6.1. Then performing the integrations in (3.16) we find that the k'th 3x3 block of C has the form

$$C_k = \frac{S_k}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \tag{6.23}$$

and singularity of C_k occurs when the area S_k of the element k vanishes. In one dimension this corresponded to node overtaking: in the present case the situation can best be described by saying that a plane section of the solution has become "vertical", i.e. parallel to the v-axis.

In hyperbolic problems we again take the situation as indicative of the formation of a shock. As in one dimension the piecewise linear solution has developed an infinitely large slope and the function v has become double valued. Our underlying procedure is to preserve this situation by introducing a shock condition as part of the algebraic system, deleting those equations superseded by such a condition. In two dimensions, assuming that the node i is attempting to cross the line j_1,j_2 (see Fig. 6.3), we impose the internal boundary conditions (shock conditions)

$$-(\dot{X}_i,\dot{Y}_i)\cdot\hat{\underset{\sim}{n}} = (\dot{X}_{j_1},\dot{Y}_{j_1})\cdot\hat{\underset{\sim}{n}} = (\dot{X}_{j_2},\dot{Y}_{j_2})\cdot\hat{\underset{\sim}{n}} = \lim \frac{\int_{\underset{\sim}{s}_1}^{\underset{\sim}{s}_2} L(u)\, dn}{a_2 - a_1} \tag{6.24}$$

as $a_1 \to a_i$, $\underset{\sim}{s}_1 \to \underset{\sim}{s}_i = \underset{\sim}{s}_{i_n}$ from the front, $a_2 \to a_{i_n}$, $\underset{\sim}{s}_2 \to \underset{\sim}{s}_{i_n} = s_i$ from the back where $\hat{\underset{\sim}{n}}$ is in the direction of $\underset{\sim}{s}_i$, (c.f. (5.7)). If node i runs into node j_1, the one-dimensional procedure of section 5 is applied in the direction of $\hat{\underset{\sim}{n}}$.

There is one significant difference in the geometrical behaviour of elements which shock in two dimensions as compared with one dimension. In one dimension an element is lost but the character of the solution in adjacent elements is unaffected In two dimensions, when node i meets the line joining nodes j_1 and j_2 (causing the vertical triangle (ij_1j_2)) an element is lost and in the process the piecewise linear solution in the triangles (ij_1j_3), (ij_2j_3) becomes adjacent to that in the triangle $(j_1j_2j_4)$. The triangulation of the plane is corrupted but it is easily restored

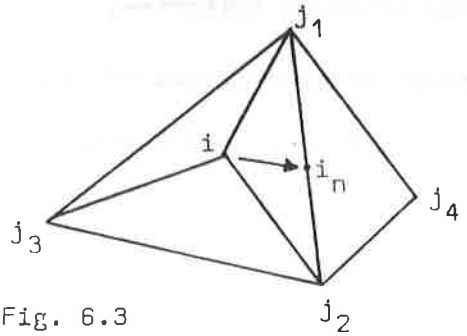

Fig. 6.3

by the device of renumbering the triangle $(j_1 j_2 j_4)$ as $(j_1 i_n j_4)$ and the (lost) triangle $(i j_1 j_2)$ as $(i_n j_4 j_2)$. Thus, in two dimensions, in the formation of a shock the number of triangles is preserved in general, as opposed to the situation in one dimension where the number of elements is reduced by one.

These ideas go over into three dimensions. The matrix C becomes singular when a node runs into a face, a side, or another node of the mesh tetrahedra. In all cases the system of equations may be modified by the introduction of a shock condition similar to (5.7) or (6.24) which supersedes that part of the algebraic system which is causing the singularity. In the case of a node running into a face of a tetrahedron one element is lost but three are created as a result of the application of an argument analogous to that for two dimensions above. The number of nodes however remains unchanged.

Questions concerning the introduction and deletion of nodes have not been discussed, since we believe that they constitute a separate issue. We merely note that there is no problem from an algebraic point of view in the insertion of a new node now that parallelism is no longer a problem. Criteria for altering the number of nodes or elements need a separate study.

We turn now to the application of the ideas in the paper to systems of equations.

7. SYSTEMS OF EQUATIONS

We confine the discussion to one dimension, although the extension to higher dimensions is straightforward.

The question arises as to whether to choose a common grid or individual grids for the different components of the system. Gelinas et al (loc. cit.) include examples which use a common grid, but since the use of individual grids preserves the advantageous structure of the matrices, we have preferred this option in what follows and in the associated example.

For the system of evolutionary equations

$$u_t^\ell - L^\ell(u^1,u^2,..,u^M) = 0 \qquad (\ell = 1,2,...,M), \tag{7.1}$$

we may generalise the procedure in Section 2, seeking semi-discrete solutions of the form

$$v^\ell(x,t) = \sum a_j^\ell(t)\alpha_j^\ell(x;\underline{s}^\ell(t)) \tag{7.2}$$

(in one dimension), where a_j^ℓ is the nodal amplitude at s_j^ℓ for each component of v. Note that we use a separate mesh $\{s_i^\ell : i=1,2,...,N_\ell\}$ for each component ℓ. The argument proceeds as in Section 1 with the addition of superfixes ℓ.

In place of (2.6) we minimise the L_2 norm

$$\sum_{\ell=1}^{M} \| v_t^\ell - L^\ell(v^1,v^2,..,v^M) \|_{L_2}^2 \tag{7.3}$$

and this leads to the set of MFE equations

$$\begin{aligned}
&\sum_{j=1}^{N_\ell} \dot{a}_j^\ell \langle\alpha_j^\ell,\alpha_i^\ell\rangle + \sum_{j=1}^{N_\ell} \dot{s}_j^\ell \langle\beta_j^\ell,\alpha_i^\ell\rangle = \langle L^\ell(u^1,u^2,...,u^M),\ \alpha_i^\ell\rangle \\
&\sum_{j=1}^{N_\ell} \dot{a}_j^\ell \langle\alpha_j^\ell,\beta_i^\ell\rangle + \sum_{j=1}^{N_\ell} \dot{s}_j^\ell \langle\beta_j^\ell,\alpha_i^\ell\rangle = \langle L^\ell(u^1,u^2,...,u^M),\ \beta_i^\ell\rangle
\end{aligned} \tag{7.4}$$

for $i=1,2,...,N_\ell$ and $\ell=1,2,...,M$.

If we now write

$$\underline{y}^\ell = \{a_1^\ell,s_1^\ell;a_2^\ell,s_2^\ell;...;a_{N_\ell}^\ell,s_{N_\ell}^\ell\} \tag{7.5}$$

the equations (7.4) can be written as M ordinary differential equation systems linked only by their right hand sides, namely,

$$A(\underline{y}^\ell)\dot{\underline{y}}^\ell = \underline{g}^\ell(\underline{y}^1,\underline{y}^2,...,\underline{y}^M) \tag{7.6}$$

for $\ell = 1,2,...,M$.

The structure of the $2N_\ell \times 2N_\ell$ matrix A of (7.6) is precisely the same as for the scalar case, with elements calculated using the nodal amplitudes and positions of the ℓ'th component only. The $2N_\ell$ - vector g^ℓ has elements given by

$$g^{\ell}_{2i-1} = \langle L^{\ell}(u^1,u^2,\ldots,u^M),\ \alpha^{\ell}_i\rangle$$
$$g^{\ell}_{2i} = \langle L^{\ell}(u^1,u^2,\ldots,u^M),\ \beta^{\ell}_i\rangle \qquad (7.7)$$

for $i = 1,2,\ldots,N_{\ell}$.

The important feature of this approach to systems is that all the results of the previous sections apply. In particular, each MFE matrix $A^{\ell} = A(\underline{y}^{\ell})$ can be decomposed in the form

$$A^{\ell} = (M^{\ell})^T C^{\ell} M^{\ell} \qquad (7.8)$$

(in any number of dimensions) and, although parallelism can occur within each component ℓ, the same procedures as in previous sections are always applicable.

The integrations in (7.7), which involve evaluation of components other than ℓ over elements corresponding to the ℓ- component, can be carried out using, for example, Gaussian quadrature, picking up values of all components at the Gauss points of each ℓ-element.

An example of a system is given in the next section.

8. DISCUSSION

In this paper we have shown that the Moving Finite Element Method, although intrinsically non-linear, leads to a matrix system which is relatively simple to analyse and easy to adapt in the two degenerate cases, namely, parallelism, where our technique is completely general and applies to all evolutionary equations, and element folding, or node-overtaking, where we propose a technique for hyperbolic problems.

As a result the MFE solution comes down to the solution, in time only, of the system of ordinary differential equations

$$\dot{\underline{y}} = A^{-1}(\underline{y})\underline{g}(\underline{y}) \qquad (8.1)$$

(c.f. (2.7) (or (5.3))), which can be solved numerically by some time-stepping scheme. Note that the degeneracy of A (which causes a modification of (8.1)) is dependent on $\underline{y}$, so that any implicit scheme will need to know the updated value of $\underline{y}$ before any degeneracy of A can be recognised and special action can be taken. For this reason explicit or Runge-Kutta schemes are to be preferred in the solution of (8.1). This is not a serious disadvantage

since the movement of the nodes to some extent pre-empts the necessity of the use of implicit schemes to prevent instability (see below).

To discuss the implications of the results we mention four examples, all in one-dimension where the theory is most developed. They are :-

(i) the inviscid Burger's equation

$$u_t + uu_x = 0$$

which has special significance for the MFE method discussed here in that the MFE solution is exact for piecewise linear data (see Wathen (1982): (ii) exactly the same as (i) but with some extra 'parallel nodes" added into the initial data. (This example shows the effect of the parallelism algorithm of section 5.)

(iii) the Buckley-Leverett equation in the form given by Concus and Proskurowski (1979)

$$u_t + \frac{\partial}{\partial x}\left(\frac{u^2}{u^2 + \frac{1}{2}(1-u)^2}\right) = 0$$

with the solution held fixed at the left hand boundary, where, although we do not expect an exact solution, we note a tendency for the nodes to follow the characteristics approximately, illustrating how explicit methods may be adequate when time-stepping is used to solve (8.1)

(iv) the wave equation, written as a system, where (avoiding characteristics) we solve the non-diagonalised system to illustrate how separate meshes can be used to provide an approximation to the exact solution.

The results for (i), (ii) and (iii) (the latter reproduced from Wathen (1982)) are given in Figs. 8.2, 8.2 and 8.3 respectively. In all three examples an arbitrarily large time step is possible subject only to the presence or development of shocks, i.e. the limitation on a time step is simply up to the instant where a node runs into a shock or a shock is formed. In example (i) the exact solution is reproduced for piecewise linear data but in example (iii) accuracy is lost because the "wave speed" is not piecewise linear and the piecewise linear character of the data is not preserved; accuracy can be improved by taking a number of smaller time

steps. In cases where the characteristics are not straight further accuracy will be lost. In addition accuracy is, of course, dependent on the number of nodes used.

Finally, we consider in more detail the wave equation, example (iv), written as a system in the form

$$u^1_t - u^2_x = 0, \quad u^2_t - u^1_x = 0 \tag{8.2}$$

(c.f. (7.1)). The corresponding set of MFE equations is

$$\begin{aligned} A(\underline{y}^1)\frac{d\underline{y}^1}{dt} &= \underline{G}^1(\underline{y}^2) \\ A(\underline{y}^2)\frac{d\underline{y}^2}{dt} &= \underline{G}^2(\underline{y}^1) \ , \end{aligned} \tag{8.3}$$

where $\underline{y}^1$, $\underline{y}^2$ are given by (7.5) for $\ell = 1,2$, and

$$\begin{aligned} G^1_{2i-1} &= \langle u^2_x, \alpha^1_i \rangle \ , \quad G^1_{2i} = \langle u^2_x, \beta^1_i \rangle \quad (i=1,2,\ldots,N_1) \\ G^2_{2i-1} &= \langle u^1_x, \alpha^2_i \rangle \ . \quad G^2_{2i} = \langle u^1_x, \beta^2_i \rangle \quad (i=1,2,\ldots,N_2) \end{aligned} \tag{8.4}$$

The MFE matrices can be decomposed into the form

$$A(\underline{y}^1) = M_1^T C_1 M_1, \qquad A(\underline{y}^2) = M_2^T C_2 M_2 \tag{8.5}$$

so that (8.3) becomes

$$M_1^T C_1 M_1 \frac{d\underline{y}^1}{dt} = \underline{G}^1(\underline{y}^2), \quad M_2^T C_2 M_2 \frac{d\underline{y}^2}{dt} = \underline{G}^2(\underline{y}^1) \tag{8.6}$$

and, provided that there is no degeneracy,

$$\frac{d}{dt}\underline{y}^1 = M_1^{-1} C_1^{-1} M_1^{-T} \underline{G}^1(\underline{y}^2), \quad \frac{d}{dt}\underline{y}^2 = M_2^{-1} C_2^{-1} M_2^{-T} \underline{G}^2(\underline{y}^1). \tag{8.7}$$

The solution is then found by explicit Euler time-stepping (as are the solutions in (i), (ii) and (iii). The results are shown in Fig. 8.4, the first two frames showing the initial data. Note that for component u^1, all the nodes in the initial data are parallel and the algorithm of section 5 is automatically used.

As the solution evolves, curvature is introduced, and the parallelism disappears.

The true solution of this problem is the superposition of right and left moving triangular pulses: both are positive in the case of the second component (see Fig.(8.4)) while for the first component they are of opposite sign.

9. SUMMARY

In conclusion we make a number of recommendations based on the work of the present paper which we hope may clarify aspects of the Moving Finite Element Method. Firstly we note the well conditioned nature of the matrix of (3.3) which is simply derived from $A(y)$. Because of the power and simplicity of the inversion procedures and for reasons given below, we advocate explicit (or Runge-Kutta) methods for the numerical integration of the differential equations (9.4) for hyperbolic systems. In this connection we note that for the case of the scalar inviscid Burger's equation, the movement of the nodes precisely follows the characteristics and that the Moving Finite Element method with Euler explicit time stepping gives the exact solution for piecewise linear initial data with an unlimited time step (Wathen, 1982). In general this will not be the case but we may nevertheless expect the nodes to move in such a way as to follow the characteristics approximately where possible. Hence we do not expect an explicit solver for the MFE equations to be subject to the usual limitations on explicit solvers on fixed meshes.

Secondly, where systems of equations are involved we recommend separate meshes for each component rather than a common mesh. This is because with this approach the stucture of the MFE equations is maintained together with the consequent advantages in solution procedure.

Finally we summarise the procedures described in this paper, taking the one-dimensional scalar problem as example. Form the set of equations

$$C\dot{\underline{w}} = \underline{b} \quad , \tag{9.1}$$

where $\dot{\underline{w}}$ is the vector (3.8) which provides the coefficients in the expansion (3.7) for $v_t(x,t)$, C is the matrix given by (3.16) which is square block 2x2 diagonal and depends on the nodal positions $\underline{s}$, and

$$b_k = \langle L(v), \phi_k \rangle \tag{9.2}$$

which depends on the vector of fundamental variables $\underline{y}$ (see (2.9)). Provided that no element is degenerate, C is non-singular and

$$\dot{\underline{w}} = C^{-1}\underline{b}. \tag{9.3}$$

To obtain $\dot{\underline{y}}$ from $\dot{\underline{w}}$ we use (3.9). In one dimension M is square block 2x2 diagonal depending on $\underline{y}$. Provided that there is no parallelism M is non-singular and

$$\dot{\underline{y}} = M^{-1}\dot{\underline{w}} = M^{-1}C^{-1}\underline{b}. \tag{9.4}$$

From (4.4) and (4.5) we have easy explicit inverses for C and M although the dependence of $\underline{b}$ on $\underline{y}$ (through $L(v)$ in (9.2)) involves quadrature.

If there is element degeneracy or parallelism (readily indicated by the singularity of C or M, respectively) proceed as follows. In the case of element degeneracy (C singular) the system degenerates into two disjoint systems; this necessitates an internal boundary condition which, in the case of shocks, can be taken to be the jump condition (5.7). In the case of parallelism (M singular) work with $\underline{y}$ alone, solving (2.8) directly by temporarily fixing any parallel node(s), solving the new system and adding multiple(s) of the vector(s) (5.5) spanning the null space of A. The latter can be done in such a way as to locate the parallel nodes in any convenient way, e.g. equally spaced. In solving the new system it is convenient to turn to iterative methods, the spectral radius of the Jacobi iteration matrix being $\frac{1}{2}$ and the SOR method with optimised parameter (3.5) or the generalised conjugate gradient method converging very rapidly.

In higher dimensions the procedure is very similar although some of the steps are more intricate, particularly those avoiding singularities. For details see the text.

ACKNOWLEDGEMENTS

We are greatly indebted to Professor K.W. Morton for extensive discussions and in particular for pointing out the advantages of the use of the base functions $\phi_{k\nu}$. We would also like to thank Professor G.H. Golub for drawing our attention to the method of the first reference, and Dr. B.M. Herbst for further discussions.

A.J. Wathen wishes to acknowledge the support of an SERC CASE Studentship and of the London Research Station of British Gas.

REFERENCES

Concus, P., Golub, G.H. & O'Leary, D.P., 1975. A Generalised Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations. Sparse Matrix Computations Ed. J.R. Bunch & D.J. Rose. Academic Press.

Concus, P. & Proskurowski, W., 1979. Numerical Solution of a Nonlinear Hyperbolic Equation by the Random Choice Method. J. Comput. Phys. 30, 153-166.

Dupont, T., 1982. Mesh Modification for Evolution Equations. Math. Comp. 39 No. 159. 85-107.

Gelinas, R.J., Doss, S.K. & Miller, K., 1981. The Moving Finite Element Method: Applications to General Partial Differential Equations with Multiple Large Gradients. J. Comput. Phys. 40, 202-249.

Herbst, B.M., 1982. Moving Finite Element Methods for the Solution of Evolution Equations. Ph.D. Thesis, University of the Orange Free State.

Lynch, D.R., 1982. Unified Approach to Simulation on Deforming Elements with Application to Phase Change Problems. J. Comput. Phys. 47, 387-411.

Miller, K. & Miller, R., 1981. Moving Finite Elements, Part I. SIAM J.N.A., 18, 1019-1032.

Miller, K., 1981. Moving Finite Elements, Part II. SIAM J.N.A., 18, 1033-1057.

Varga, R.S., 1962. Matrix Iterative Analysis. Prentice Hall.

Wathen, A.J., 1982. Moving Finite Elements and Applications to Some Problems in Oil Reservoir Modelling. University of Reading, Numerical Analysis Report 4/82.

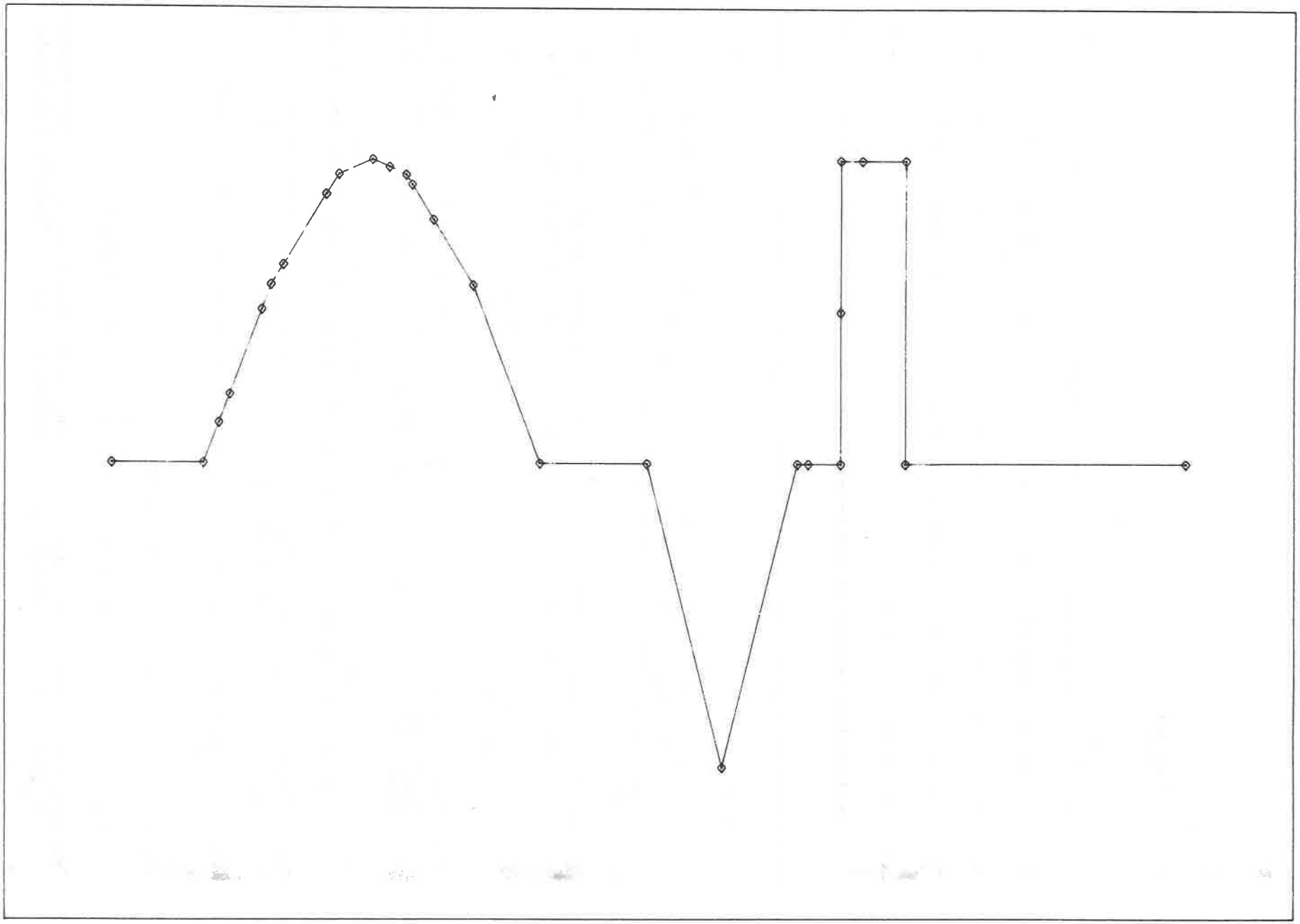

Figure 8.2 (1st frame)

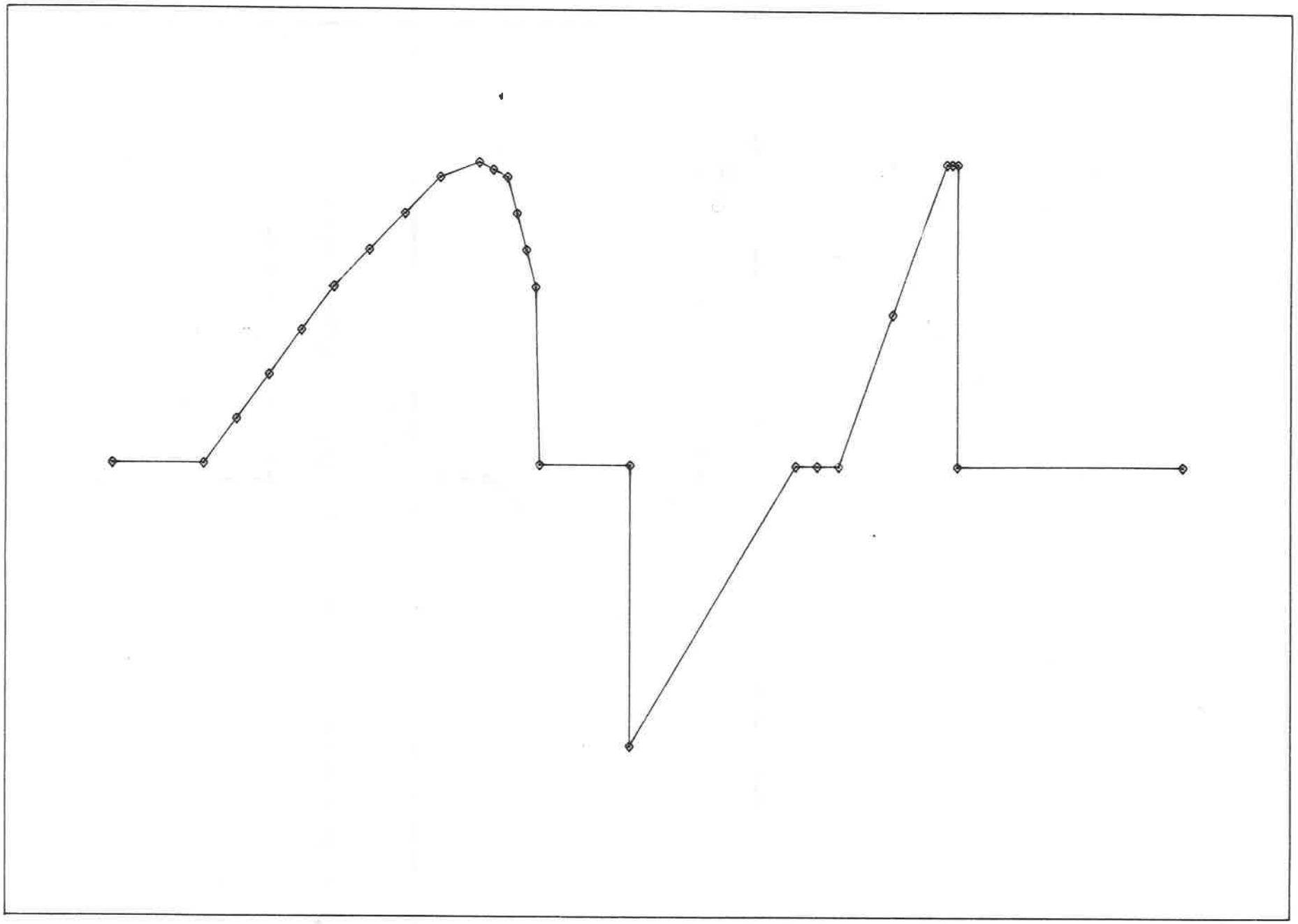

Figure 8.2 (2nd frame)

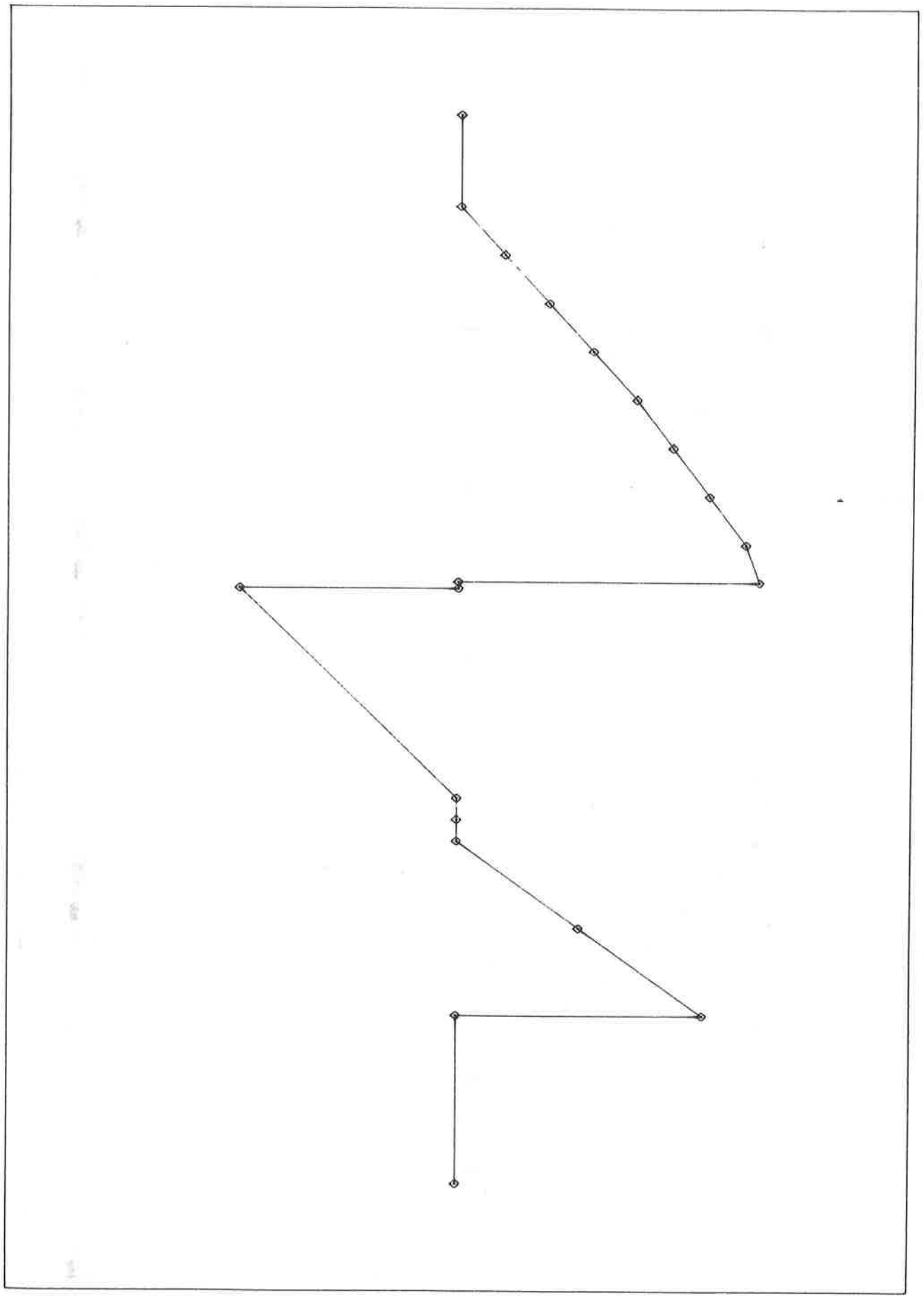

Figure 8.2 (3rd frame)

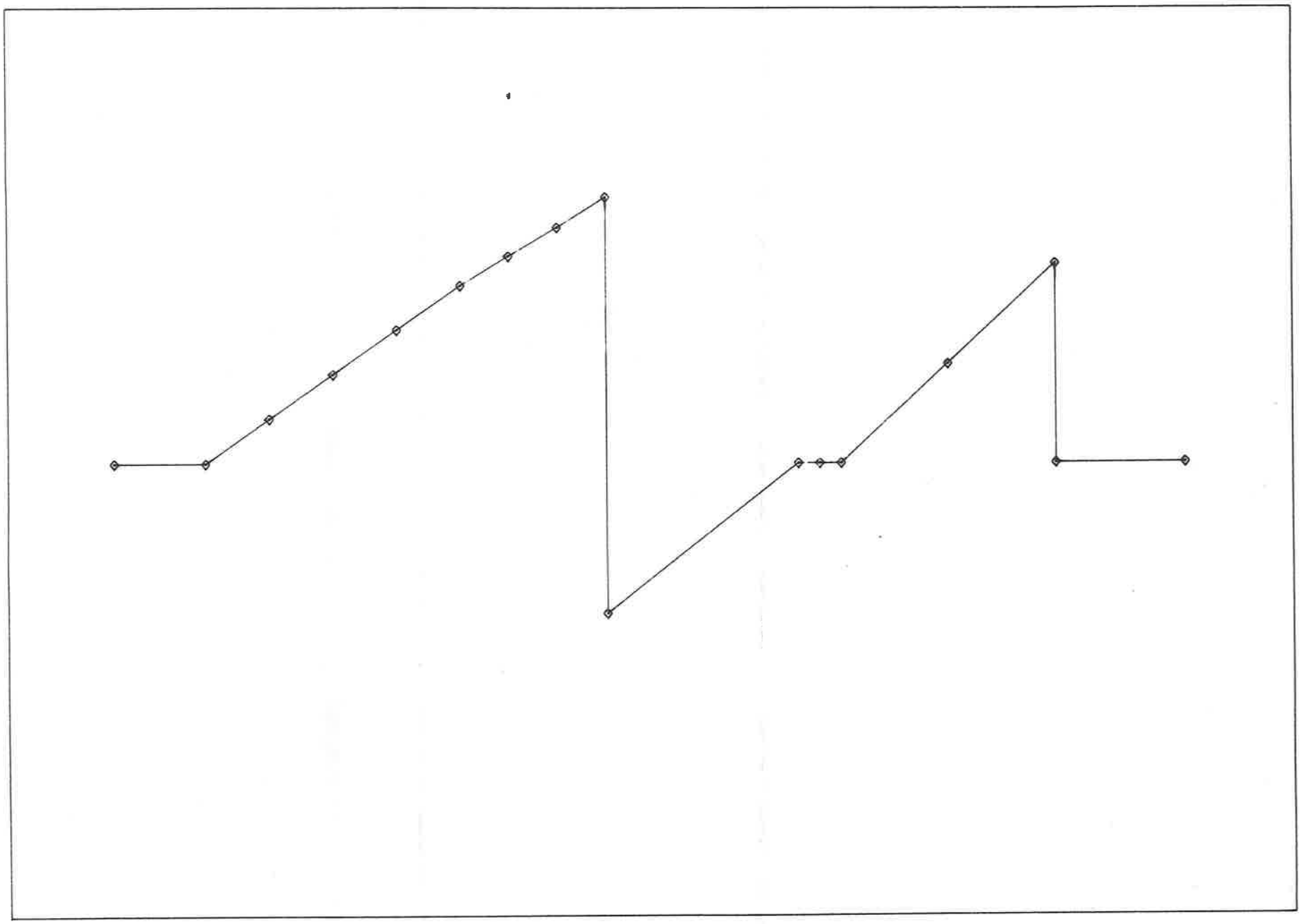

Figure 8.2 (4th frame)