INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 8, 977–987 (1988)

IMPROVING THE ACCURACY OF THE LEAST-SQUARES FINITE ELEMENT APPROXIMATION OF THE LINEARIZED STEADY EULER EQUATIONS BY AN EMBEDDING METHOD

P. WILDERS

Delft University of Technology, Department of Mathematics & Informatics, P.O. Box 356, 2600 AJ Delft, The Netherlands

SUMMARY

The standard least-squares finite element method for the linearized Euler equations turns out to be inaccurate. This method is studied in detail for a system of composite type, obtained by transformation of the linearized Euler equations. The shortcomings of the method are clarified and an embedding method is constructed. It is shown numerically that this new method is $O(h^2)$-accurate.

KEY WORDS Euler equations Least squares Finite elements Embedding methods

1. INTRODUCTION

We consider the system

$$\frac{\partial \mathbf{A}\mathbf{w}}{\partial x_1} + \frac{\partial \mathbf{B}\mathbf{w}}{\partial x_2} = \mathbf{f}, \tag{1}$$

where

$$\mathbf{w} = (p, u, v)^{\mathrm{T}},$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2\bar{u} & 0 \\ 0 & \bar{v} & \bar{u} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & \bar{v} & \bar{u} \\ 1 & 0 & 2\bar{v} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f^{(1)} \\ f^{(2)} \\ f^{(3)} \end{bmatrix}.$$

The functions $\bar{u}$, $\bar{v}$ and $f^{(i)}$ are given. System (1) is of composite type, i.e. one real and two complex characteristics occur. Equations (1) are obtained by applying Newton's method to the incompressible Euler equations. It is well known that numerical methods for (1) often lead to linear systems which are difficult to solve numerically. In order to avoid this problem, one may consider least-squares or embedding methods. Equation (1) is embedded in a system of second-order equations, with discrete approximations leading to linear systems that allow efficient solution methods.

The least-squares method has been applied to the numerical solution of the steady Euler

0271–2091/88/080977–11$05.50

Received 30 December 1986
Revised 24 November 1987

equations.[1, 2] For these equations embedding methods have been investigated by Johnson[3] and Chang and Johnson.[4] We have applied the method of Bruneau *et al.*[2] to obtain numerical solutions of the subcritical steady shallow-water equations. However, we have encountered some severe problems, which are most likely caused by local inaccuracies of the method near boundaries. None of the above authors considers the accuracy of the least-squares method. The method of Bruneau *et al.* is based on a least-squares solution of (1). We therefore propose to study (1) in more detail.

In Section 2 we present a brief account of the least-squares finite element method. In Section 3 a numerical test problem is defined and some numerical results are presented. In turns out that the method is inaccurate. In Section 4 we transform (1) by multiplying the system with the left eigenvector, corresponding to the real characteristic, and by introducing the linearized total pressure P as a new variable. The transformed system is more suitable for a detailed analysis and for accurate numerical computations. A drawback of the transformed system is the occurrence of terms which are not in conservation form. We are only interested in smooth solutions of (1) and therefore this is acceptable.

In Section 5 we investigate the least-squares solution of the transformed system in more detail. In particular, we look at the numerical approximation along boundaries using truncation error analysis. In Wilders[5] a single conservation equation has been studied in this manner and this work serves as our guide. This analysis provides a new embedding method for the numerical solution of the transformed system and application to the test problem of Section 3 shows that accurate results are obtained. In Section 6 the new embedding method is developed further in order to include curved boundaries and non-Cartesian grids.

The resulting linear system in the new embedding method is not symmetrical. In Section 7 the iterative solution of this system is discussed. We use the CGS method (conjugate gradients squared).[6] Following Meijerink and Van der Vorst,[7] a new variant of the incomplete decomposition with corrections only on the main diagonal is constructed and promising results are obtained.

2. THE LEAST-SQUARES APPROACH

We study (1) on $\Omega \subset \mathbb{R}^2$. We define

$$\mathrm{I} = \int_\Omega \left\| \frac{\partial \mathbf{A}\mathbf{w}}{\partial x_1} + \frac{\partial \mathbf{B}\mathbf{w}}{\partial x_2} - \mathbf{f} \right\|_2^2 \mathrm{d}\Omega. \tag{2}$$

Let $\mathbf{w}$ be a minimum of I and let τ be an arbitrary scalar test function. We set $\boldsymbol{\tau}_m = \tau \mathbf{e}_m$, $m = 1, 2, 3$, where $\mathbf{e}_m$ denotes the mth unit vector. We have

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathrm{I}(\mathbf{w} + \varepsilon \boldsymbol{\tau}_m)\Big|_{\varepsilon = 0} = 0, \quad m = 1, 2, 3,$$

or

$$\int_\Omega \left(\frac{\partial \mathbf{A}\mathbf{w}}{\partial x_1} + \frac{\partial \mathbf{B}\mathbf{w}}{\partial x_2} - \mathbf{f} \right)^{\mathrm{T}} \left(\frac{\partial \mathbf{A}\boldsymbol{\tau}_m}{\partial x_1} + \frac{\partial \mathbf{B}\boldsymbol{\tau}_m}{\partial x_2} \right) \mathrm{d}\Omega = 0, \quad m = 1, 2, 3. \tag{3}$$

Note that the inner product occurs in (3). In more detail (3) reads

$$\int_\Omega \sum_{k=1}^{3} \left(\sum_{l=1}^{3} \frac{\partial a_{kl} w_l}{\partial x_1} + \frac{\partial b_{kl} w_l}{\partial x_2} - f^{(k)} \right) \left(\frac{\partial a_{km} \tau}{\partial x_1} + \frac{\partial b_{km} \tau}{\partial x_2} \right) \mathrm{d}\Omega = 0, \quad m = 1, 2, 3. \tag{4}$$

Equations (4) are approximated in the finite-dimensional subspace spanned by isoparametric

bilinear quadrilateral elements ϕ_j. We apply the so-called product approximation or group formulation.[2,8-10] This means that terms like $a_{kl}w_l$ are approximated by $\Sigma_j(a_{kl}w_l)_j\phi_j$. For example, the term in (4) obtained by setting $m=k=2, l=3$ leads to the contribution

$$\sum_j (2\bar{u}_i\bar{u}_j s_{ij}^{(1,2)}+\bar{v}_i\bar{u}_j s_{ij}^{(2,2)})v_j-(2\bar{u}_i d_{ij}^{(1)}+\bar{v}_i d_{ij}^{(2)})f_j^{(2)}, \tag{5}$$

where

$$s_{ij}^{(\alpha,\beta)}=\int_\Omega \frac{\partial\phi_i}{\partial x_\alpha}\frac{\partial\phi_j}{\partial x_\beta}\mathrm{d}\Omega, \qquad d_{ij}^{(\alpha)}=\int_\Omega \frac{\partial\phi_i}{\partial x_\alpha}\phi_j\mathrm{d}\Omega, \quad \alpha,\beta=1,2.$$

For the computation of the integrals we use a four-point Gaussian quadrature formula.

Let $\mathbf{n}=(n_1, n_2)$ denote the outward unit normal on $\partial\Omega$. Integration by parts in (3) leads to

$$-\int_\Omega\left(\mathbf{A}^{\mathrm{T}}\frac{\partial}{\partial x_1}+\mathbf{B}^{\mathrm{T}}\frac{\partial}{\partial x_2}\right)\left(\frac{\partial\mathbf{A}\mathbf{w}}{\partial x_1}+\frac{\partial\mathbf{B}\mathbf{w}}{\partial x_2}-\mathbf{f}\right)^{\mathrm{T}}\boldsymbol{\tau}_m\mathrm{d}\Omega$$
$$+\int_{\partial\Omega}(n_1\mathbf{A}^{\mathrm{T}}+n_2\mathbf{B}^{\mathrm{T}})\left(\frac{\partial\mathbf{A}\mathbf{w}}{\partial x_1}+\frac{\partial\mathbf{B}\mathbf{w}}{\partial x_2}-\mathbf{f}\right)^{\mathrm{T}}\boldsymbol{\tau}_m\mathrm{d}s=0, \quad m=1,2,3. \tag{6}$$

We conclude that (3) are the Galerkin equations associated with the system of second-order equations

$$-\left(\mathbf{A}^{\mathrm{T}}\frac{\partial}{\partial x_1}+\mathbf{B}^{\mathrm{T}}\frac{\partial}{\partial x_2}\right)\left(\frac{\partial\mathbf{A}\mathbf{w}}{\partial x_1}+\frac{\partial\mathbf{B}\mathbf{w}}{\partial x_2}-\mathbf{f}\right)=0, \tag{7}$$

with boundary conditions that are either essential ($\boldsymbol{\tau}_m=0$) or natural. The latter imply that certain linear combinations of the original first-order equations should be zero on the boundary.

Formula (7) shows that the least-squares method is equivalent with a special embedding method. This embedding method has been studied previously by Johnson[3] and Chang and Johnson[4] who explored finite difference methods. We remark that the finite element least-squares method described in (4) and (5) leads to a nine-point approximation of (7) on Cartesian grids. Furthermore, the matrix is symmetric and positive definite. System (1) is of composite type. The characteristic eigenvalues are $\lambda=\pm i$ and $\lambda=\bar{v}/\bar{u}$. System (7) is of composite type as well, because written as a system of six first-order equations we find the same eigenvalues with multiplicity two. Inspection of (6) tells us that the least-squares method automatically generates the correct number of natural boundary conditions for (7). It should be remarked that on walls the second and third rows of the matrix $n_1\mathbf{A}^{\mathrm{T}}+n_2\mathbf{B}^{\mathrm{T}}$ are dependent and consequently only one natural condition is generated besides the essential condition of zero normal velocity.

3. THE NUMERICAL TEST PROBLEM

We consider a test problem on $\Omega=\langle 0,1\rangle\times\langle 0,1\rangle$. In (1) we take

$$\begin{cases} p=0 \\ u=2+\sin\pi x_1\cos\pi x_2, \quad v=-\cos\pi x_1\sin\pi x_2, \\ \bar{u}=u, \bar{v}=v. \end{cases} \tag{8}$$

The functions $f^{(i)}$ in the right-hand side of (1) are chosen such that the given $\mathbf{w}=(p,u,v)$ is the exact solution. Note that the velocity field is divergence-free and that the choice of the coefficients $\bar{u}$ and $\bar{v}$ is natural, because (1) is obtained by applying Newton's method.

According to Zajaczkowski[11] and Saxer *et al.*,[12] possible boundary conditions are

$$\begin{cases} u \text{ and } v \text{ given at the inflow boundary } x_1 = 0, \\ p \text{ given at the outflow boundary } x_1 = 1, \\ v = 0 \text{ at the walls } x_2 = 0, 1. \end{cases} \tag{9}$$

We have no intentions of going deeply into inflow and outflow conditions at this stage. The correct inclusion of the walls is much more vital, as will become clear in the following sections. For our purpose it is more appropriate to replace (9) by

$$\begin{cases} p \text{ and } u \text{ given at the inflow boundary } x_1 = 0, \\ u \text{ given at the outflow boundary } x_1 = 1, \\ v = 0 \text{ at the walls } x_2 = 0, 1. \end{cases} \tag{10}$$

A Cartesian equidistant grid is employed. The nodal points have indices $i, j = 0, \ldots, n$. We choose $n = 8, 16, 32, 64$. The discrete maximum-norm of the error is denoted by $e_\infty(n)$ and the discrete l_2-norm by $e_2(n)$. With $e_\bullet$, $\bullet = \infty, 2$, either of these norms is represented generically. We define

$$q_{2n} = {}^2\log \frac{e_\bullet(n)}{e_\bullet(2n)}. \tag{11}$$

In Table I the computed values of $e_\bullet$ (16) and q_{2n} are presented. Of course, the explored method is the least-squares finite element method from Section 2.

The conclusion is that the least-squares method is inaccurate. We notice that the nodal point with the largest error is situated on the characteristic boundary, which is in agreement with Wilders.[5] In this work a single conservation equation is studied and it is shown that the least-squares scheme is inaccurate with respect to the treatment of boundaries.

It should be remarked that the inaccuracies in Table I are by no means caused by the choice of the inflow and outflow conditions in (10). In the computations the conditions (10) and the often-employed conditions (9) have been found to differ little.

4. A FUNDAMENTAL TRANSFORMATION

System (1) is not very suitable for analysis. We prefer to make the hyperbolic and elliptic nature more transparent. The left eigenvector $\boldsymbol{l}$, corresponding to the real eigenvalues $\lambda = v/u$, reads

Table I. Values of $e_\bullet(16)$ and q_{2n} for (1), (8)

	e_∞	e_2
$e_\bullet(16)$	0·57e0	0·14e0
q_{16}	1·3	1·1
q_{32}	1·6	1·6
q_{64}	1·6	1·7

$\boldsymbol{l}=(-(\bar{u}^2+\bar{v}^2), \bar{u}, \bar{v})$. As in the theory of hyperbolic systems, (1) is premultiplied with $\boldsymbol{l}$ and the corresponding 'Riemann invariant' P is introduced, viz.

$$P=p+\bar{u}u+\bar{v}v. \tag{12}$$

The variable P is the linearized total pressure. The transformed system reads

where
$$\mathbf{K}\frac{\partial\tilde{\mathbf{A}}\mathbf{w}}{\partial x_1}+\mathbf{L}\frac{\partial\tilde{\mathbf{B}}\mathbf{w}}{\partial x_2}+\mathbf{C}\mathbf{w}=\mathbf{g}, \tag{13}$$

$\mathbf{w}=(P, u, v)^{\mathrm{T}}$,

$$\tilde{\mathbf{A}}=\begin{bmatrix}1 & 0 & 0\\ \alpha & \bar{u} & -\bar{v}\\ 0 & \bar{v} & \bar{u}\end{bmatrix}, \qquad \tilde{\mathbf{B}}=\begin{bmatrix}1 & 0 & 0\\ 0 & \bar{v} & \bar{u}\\ \alpha & -\bar{u} & \bar{v}\end{bmatrix}, \qquad \mathbf{C}=\begin{bmatrix}0 & a & b\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix},$$

$$\mathbf{K}=\begin{bmatrix}\bar{u} & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}, \qquad \mathbf{L}=\begin{bmatrix}\bar{v} & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix},$$

$\alpha=1$,

$$a=\bar{u}\bar{u}_{x_1}+\bar{v}\bar{v}_{x_1}+\bar{u}\bar{v}_{x_2}-\bar{v}\bar{u}_{x_2}, \qquad b=\bar{v}\bar{u}_{x_1}-\bar{u}\bar{v}_{x_1}+\bar{u}\bar{u}_{x_2}+\bar{v}\bar{v}_{x_2}.$$

The reason for the introduction of the parameter α will become clear in due course. We remark that in practical situations it is preferable to interchange the order of Newton linearization and transformation, because this simplifies the resulting expressions for a and b.

System (13) is special in the sense that for $\alpha=a=b=0$ a decoupling occurs. The first equation becomes a simple non-conservative convection equation in P, while the second and third equations give rise to an elliptic system in u and v. System (13) with $\alpha=a=b=0$ is therefore useful as a reference system. For the boundary conditions we take

$$\begin{cases} P, u \text{ given at the inflow boundary } x_1=0,\\ u \text{ given at the outflow boundary } x_1=1,\\ v=0 \text{ at the walls } x_2=0, 1. \end{cases} \tag{14}$$

If $\alpha=a=b=0$, then one may actually prove that these conditions lead to a well-posed problem, because the Lopatinski condition for the elliptic part[13] is satisfied. Note that (14) is only a slight modification of (10), which facilitates a comparison at a later stage. In fact this was the reason for the use of (10) instead of (9). All numerical computations for (13) have been done with the boundary conditions (14).

The least-squares method described in Section 2 can equally well be used to solve (13). The product approximation is explored with the exception of the terms corresponding to non-constant entries of $\mathbf{K}$ and $\mathbf{L}$. The second-order system associated with the least-squares method reads

$$-\left(\tilde{\mathbf{A}}^{\mathrm{T}}\frac{\partial}{\partial x_1}\mathbf{K}^{\mathrm{T}}+\tilde{\mathbf{B}}^{\mathrm{T}}\frac{\partial}{\partial x_2}\mathbf{L}^{\mathrm{T}}+\mathbf{C}^{\mathrm{T}}\right)\left(\mathbf{K}\frac{\partial\tilde{\mathbf{A}}\mathbf{w}}{\partial x_1}+\mathbf{L}\frac{\partial\tilde{\mathbf{B}}\mathbf{w}}{\partial x_2}+\mathbf{C}\mathbf{w}-\mathbf{g}\right)=0. \tag{15}$$

It is straightforward to apply this method to the test problem of Section 3. Substitution of (8) in (12) gives the corresponding 'total pressure'. In Table II we present the computed values of $e_{\cdot}$(16)

Table II. Values of $e_\cdot(16)$ and q_{2n} for (13)

	e_∞	e_2
$e_\cdot(16)$	0·98e−1	0·27e−1
q_{16}	1·7	1·6
q_{32}	1·8	1·8
q_{64}	1·9	1·9

and q_{2n}. Comparison with Table I tells us that the results have improved considerably. However, this is surprising. The first equation in (13) reads

$$\bar{u}\frac{\partial P}{\partial x}+\bar{v}\frac{\partial P}{\partial y}+Q=g^{(1)},$$

which means that the main part consists of a non-conservative convection term. In Wilders[5] it has been shown that the least-squares method is inaccurate for a single conservative convection equation and one may wonder why these inaccuracies are absent in Table II. In the next section we perform a truncation error analysis similar to that of Wilders.[5] This analysis explains some aspects of Table II and will guide us to a new embedding method.

5. A NEW EMBEDDING METHOD

We perform a truncation error analysis along characteristic boundaries. It has been shown in Wilders[5] that this approach is suitable. To avoid irrelevant details, we choose $\bar{u}=1$, $a=b=0$ and $g=0$ in (13). Furthermore, $\bar{v}$ is such that

$$\bar{v}|_{x_2=0}=0 \tag{16}$$

and therefore $x_2=0$ is a characteristic boundary. Only the first equation in the embedded system (13) is important at this stage. This equation reads

$$-\left(\frac{\partial}{\partial x_1}+\frac{\partial}{\partial x_2}\bar{v}\right)\left(\frac{\partial P}{\partial x_1}+\bar{v}\frac{\partial P}{\partial x_2}\right)-\frac{\partial}{\partial x_1}\left(\frac{\partial P}{\partial x_1}+\frac{\partial(u-\bar{v}v)}{\partial x_1}+\frac{\partial(\bar{v}u+v)}{\partial x_2}\right)$$
$$-\frac{\partial}{\partial x_2}\left(\frac{\partial P}{\partial x_2}+\frac{\partial(\bar{v}u+v)}{\partial x_1}-\frac{\partial(u-\bar{v}v)}{\partial x_2}\right)=0. \tag{17}$$

In order to prevent unnecessary complications, we assume that u and v are given functions in (17) and, to make the presentation more compound, we choose $u=v=0$. For the purpose of analysis a Newton–Cotes quadrature formula is used instead of a Gaussian formula. This facilitates the analysis. The molecule at the characteristic boundary is represented in Figure 1.

The finite element equation in the boundary point 2 reads

$$-\frac{h_2}{2h_1}(P_1-2P_2+P_3)-\frac{1}{4}[\bar{v}_3(P_6-P_3)-\bar{v}_1(P_4-P_1)]$$
$$-\frac{1}{4}[\bar{v}_2(P_3-P_1)+\bar{v}_5(P_6-P_4)]-\frac{h_1}{2h_2}(\bar{v}_2^2+\bar{v}_5^2)(P_5-P_2)$$
$$-\frac{h_2}{2h_1}(P_1-2P_2+P_3)-\frac{h_1}{2h_2}(P_5-P_2)=0. \tag{18}$$

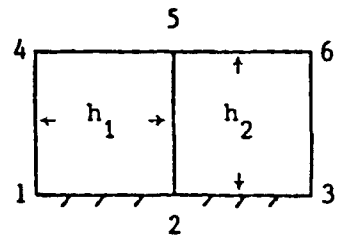

Figure 1. Molecule on boundary

Let P be a smooth solution of (17), let $\bar{v}_2 = 0$ (characteristic boundary) and let $h_1 = O(h_2)$; then the truncation error $e(P)$ of (18) is given by

$$e(P) = e_1(P) + e_2(P) + O(h_2^2), \tag{19}$$

$$e_1(P) = -\frac{1}{4} h_2 \frac{\partial}{\partial x_2^2} \bar{v} \left(\frac{\partial P}{\partial x_1} + \bar{v} \frac{\partial P}{\partial x_2} \right), \tag{20}$$

$$e_2(P) = -\frac{1}{6} h_2 \frac{\partial^3 P}{\partial x_2^3}. \tag{21}$$

For general solutions of (17), $e(P) = O(h_2)$ holds. For those solutions of (17) which are a solution of (13) as well, we automatically have

$$e_1(P) = 0. \tag{22}$$

The error e_1 is due to the convective term and (22) is surprising, because in Wilders[5] such a property had to be enforced. According to Wilders,[5] the relation (22) explains to some extent the better accuracy found in Table II. One can pursue the argument further and improve the method once more by avoiding the occurrence of the error e_2, which is easily done. Instead of (13) we define the following embedded system:

$$-\left(\bar{\mathbf{A}} \frac{\partial}{\partial x_1} \mathbf{K}^{\mathrm{T}} + \bar{\mathbf{B}} \frac{\partial}{\partial x_2} \mathbf{L}^{\mathrm{T}} \right) \left(\mathbf{K} \frac{\partial \tilde{\mathbf{A}} \mathbf{w}}{\partial x_1} + \mathbf{L} \frac{\partial \tilde{\mathbf{B}} \mathbf{w}}{\partial x_2} + \mathbf{C}\mathbf{w} - \mathbf{g} \right) = 0, \tag{23}$$

where

$$\bar{\mathbf{A}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{u} & \bar{v} \\ 0 & -\bar{v} & \bar{u} \end{bmatrix}, \qquad \bar{\mathbf{B}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{v} & -\bar{u} \\ 0 & \bar{u} & \bar{v} \end{bmatrix}.$$

The key to success in this method is that, along characteristic boundaries,

$$e(P) = O(h_2^2) \tag{24}$$

for those solutions of (17) which are a solution of (13) as well. Just as for (15), the system (23) is of the composite type, with characteristic eigenvalues $\lambda = \pm i$ and $\lambda = v/u$ of multiplicity two. The required natural boundary conditions for (23) are obtained in a straightforward way after partial integration of (23). It has been found numerically that the matrix $\mathbf{C}^{\mathrm{T}}$ in (15) is of little importance and therefore this matrix has been omitted in the new embedding method.

We now return to the computations of Sections 3 and 4. The new embedding method, called Emb, is explored. The details of the computations are exactly as in Section 4. In Table III the

Table III. Values of $e_*(16)$ and q_{2n} for (13), using Emb

	e_∞	e_2
$e_*(16)$	0·50e−1	0·13e−1
q_{16}	1·8	1·9
q_{32}	1·9	1·9
q_{64}	1·9	2·0

computed values of $e_*(16)$ and q_{2n} are presented. Comparison with Table II tells us that the results have improved once more. However, one may argue that this improvement is not vital. In the next section it turns out that the method Emb can easily be adapted to include an accurate treatment of curved boundaries as well. In the least-squares method leading to Table II this is an open question.

Comparison of Tables I and III leads to the conclusion that the new method Emb, which is based on the transformed system (13), improves the original least-squares method for (1) considerably. In fact on a moderate sized grid, say 16×16, the error in the new method is more than a factor of ten smaller.

Finally, we observe that a more efficient implementation of the embedding method is possible. Inspection of (23) shows that in the second and third equations four combinations of the form T occur, where

$$T = -\bar{u}\left(\frac{\partial^2 \eta}{\partial x_1 \partial x_2} - \frac{\partial^2 \eta}{\partial x_2 \partial x_1}\right), \qquad \eta = \bar{v}u + \bar{u}v.$$

For smooth solutions such terms vanish and may therefore be deleted. However, one must impose the original natural boundary conditions, which means that deleting such terms implies the occurrence of boundary integrals and consequently boundary elements.

6. ON THE INCLUSION OF CURVED WALLS

Until now our work has been concerned with straight walls. In this section the inclusion of curved walls is described. The relation (24) is only valid along straight walls and in fact it turns out that the accuracy of the embedding method Emb declines in some cases with curved walls. Therefore we have to develop the method further. In the new method, called EmbChb, inaccuracies along curved characteristic boundaries are absent. The method EmbChb consists of Emb plus a modification involving the implementation of a characteristic boundary scheme. For a single conservation equation this approach has already been considered in Wilders.[5]

Let Γ be a characteristic boundary with parametric representation $(x_1(t), x_2(t))$, where

$$\mathrm{d}x_1/\mathrm{d}t = \bar{u}, \qquad \mathrm{d}x_2/\mathrm{d}t = \bar{v}. \tag{25}$$

The characteristic form of the first equation of (13) along Γ reads

$$\mathrm{d}P/\mathrm{d}t + au + bv = g^{(1)}. \tag{26}$$

On Γ the first equation of (23) is replaced by

$$-\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}P}{\mathrm{d}t} + au + bv - g^{(1)}\right) = 0. \tag{27}$$

Further details of the implementation can be found in Wilders.[5] We only remark that the implementation is not very difficult, because Γ is a part of $\partial\Omega$, and that we have used a two-point Gaussian quadrature rule for the computation of integrals along Γ.

In order to evaluate the new method EmbChb, we define a new test problem. Once more the exact solution is given by (8) with the 'total pressure' P computed via (12). For the lower wall we now choose the streamline originating in the point with co-ordinates $x_1 = 0$, $x_2 = 0{\cdot}25$. Further details of the region and the grids remain the same. In Figure 2 the domain and grid for $n = 8$ is presented.

On the lower boundary the normal component of the velocity is zero. The other boundary conditions are as in (14). The implementation of the boundary condition on the lower wall can be done in several ways. For example, Bruneau *et al.*[2] use an iterative scheme. We choose instead a traditional implementation. The unknowns and equations are transformed on the boundary using local co-ordinates. Details can be found in Pinder and Gray[14] (pp. 275 and 281) and Engelman *et al.*[15] We only remark that we have computed the normal direction as in Pinder and Gray.

Further details of the implementation remain identical to those from Section 4. In Table IV the computed values of $e_\cdot(16)$ and q_{2n} are presented. Comparison with Table III tells us that the method EmbChb performs well and one may draw the conclusion that this method turns out to be promising.

7. AN ITERATIVE METHOD

The resulting linear system in the new embedding method is not symmetrical. For the iterative solution the CGS method, recently developed by Sonneveld,[6] has been implemented. We compare the number of necessary iterations for the original least-squares method LS (Section 2), for the least-squares method LST (transformed system, Section 4), for the new embedding method Emb (Section 5) and for the modified embedding method EmbChb (Section 6). The test problem

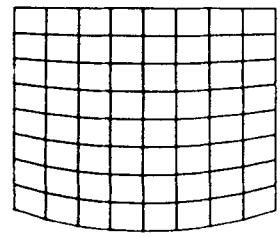

Figure 2. The domain and grid for $n = 8$

Table IV. Values of $e_\cdot(16)$ and q_{2n} for (13), using EmbChb

	e_∞	e_2
$e_\cdot(16)$	0·34e−1	0·84e−2
q_{16}	1·8	1·9
q_{32}	2·0	1·9
q_{64}	2·0	2·0

on the unit square (Sections 3 and 4) is used. For LS and LST the matrix is symmetrical and in this case the CGS method reduces to a variant of the conjugate gradients (CG) method (speaking roughly two CG iterations = one CGS iteration).

To start with, diagonal scaling is used as a preconditioner. More advanced preconditioners are difficult to obtain for all the above-mentioned methods simultaneously. For the methods based on the transformed system (13) we have succeeded in constructing a new preconditioner. This preconditioner is described in due course. In Table V the number of CGS iterations is presented (termination criterion: $\|\text{preconditioned residual}\|_2 \leqslant 1\text{e}-4$).

We may conclude that the application of the transformation from Section 4 improves the iterative properties. Furthermore, in the methods LST, Emb and EmbChb only minor differences are observed. The new preconditioner is only used here in conjunction with EmbChb, because in the previous sections this method has been found to be the most promising. It can be seen that the new preconditioner works well and in fact a fast solution method is available now. We remark that the results in Table IV are by no means special. In other test problems (we explicitly mention the one on the non-Cartesian grid of Section 6) similar results have been obtained.

Finally, we give the details of the new preconditioner. A very efficient preconditioner is the incomplete decomposition with corrections only on the main diagonal.[7,16,17] However, this decomposition failed to exist, even for method LS. It is well known that this situation may occur and, for example, in Meijerink and Van der Vorst[7] some suggestions are given to overcome this problem. Among other things, they suggest neglecting some of the Gaussian elimination corrections if the associated diagonal element becomes too small. We implemented this approach by neglecting a correction if the diagonal element became negative. However, this approach did not work well and in some cases divergence resulted. We have therefore constructed a new variant.

For the computation of the above-mentioned incomplete decomposition one uses the recurrence relation

$$d_i = a_{ii} - \sum_{j \in J_i} \frac{a_{ij} a_{ji}}{d_j}, \qquad J_i = \{ j < i : a_{ij} \neq 0 \}, \tag{28}$$

where the a_{ij} are the elements of the original matrix and where the d_i are associated with the corrected diagonal elements. For J_i we now propose

$$J_i = \{ i - i_0 \leqslant j < i : a_{ij} \neq 0 \}. \tag{29}$$

In the case of Emb and EmbChb good results are obtained with $i_0 = 1$. Enlarging i_0 did not improve the results. In the case of LS the choice $i_0 = 1$ led to an increase of the number of CGS iterations (with respect to diagonal scaling).

Table V. Number of iterations in the preconditioned CGS method

Number of unknowns	Diagonal scaling				New preconditioner
	LS	LST	Emb	EmbChb	EmbChb
198 ($n=8$)	118	60	54	55	18
782 ($n=16$)	258	119	113	136	40
3102 ($n=32$)	669	214	278	271	82

8. CONCLUSIONS

It has been shown that the accuracy of the numerical solution of the linearized Euler equations by a least-squares finite element method is disappointing, but that the method can be improved considerably by applying a transformation to the Euler equations and by considering a new embedding method, similar but not identical to the least-squares method. It has been found that this new solution method works well, even in the case of curved boundaries and non-Cartesian grids. A fast iterative solution of the resulting system has been provided by a preconditioned conjugate gradients type method, the so-called CGS method. The preconditioning method is very simple and it seems that further research might be valuable. The final goal of this study is the development of an accurate solution method of the steady shallow-water equations. We are only interested in smooth solutions of these equations and therefore it is acceptable that in the transformed system (13) the conservation form is destroyed. The next step in this study will be the implementation of the new embedding method as the linear solver for the solution of the full Euler and shallow-water equations by Newton iteration.

REFERENCES

1. C. A. J. Fletcher, 'A primitive variable finite element formulation for inviscid, compressible flow', *J. Comput. Phys.*, **33**, 301–312 (1979).
2. C. H. Bruneau, J. J. Chattot, J. Laminie and J. Guiu-Roux, 'Finite element least squares method for solving full steady Euler equations in a plane nozzle', in E. Krause (ed.), *Proc. 8*th Int. Conf. on Numerical Methods in Fluids, Springer, New York, 1982.
3. G. M. Johnson, 'Relaxation solution of the full Euler equations', in E. Krause (ed.), *Proc. 8*th Int. Conf. on Numerical Methods in Fluids, Springer, New York 1982.
4. S. Chang and G. M. Johnson, 'An embedding method for the steady Euler equations', *J. Comput. Phys.*, **63**, 191–200 (1986).
5. P. Wilders, 'On the accuracy of least squares finite elements for a first-order conservation equation', *Int. j. numer. methods fluids*, **8**, 957–964 (1988).
6. P. Sonneveld, CGS, a fast Lanczos-type solver for nonsymmetric linear systems', *Report 84-16*, Department of Mathematics, Delft University, 1984 (to be published in *Siam J. Sci. Stat. Comput.*).
7. J. A. Meijerink and H. A. Van der Vorst, 'Guidelines for the usage of incomplete decompositions in solving sets of linear equations as they occur in practical problems', *J. Comput. Phys.*, **44**, 134–155 (1981).
8. C. A. J. Fletcher, 'The group finite element formulation', *Comput. Methods Appl. Mech. Eng.*, **37**, 225–243 (1983).
9. J. J. Chattot, J. Guiu-Roux and J. Laminie, 'Numerical solution of a first-order conservation equation by a least squares method', *Int. j. numer. methods Fluids*, **2**, 209–219 (1982).
10. A. Segal and N. Praagman, 'A fast implementation of explicit time stepping algorithms with the finite element method for a class of non-linear evolution problems', *Int. j. numer. methods eng.*, **23**, 155–168 (1986).
11. W. M. Zajaczkowski, 'Solvability of an initial boundary value problem for the Euler equations in two-dimensional domain with corners', *Math. Methods Appl. Sci.*, **6**, 1–22 (1984).
12. A. Saxer, H. Felici, C. Neury and I. L. Ryhming, 'Euler flows in hydraulic turbines and ducts related to boundary conditions formulation, in M. Deville (ed.), *Proc. 7th GAMM Conf. on Numerical Methods in Fluid Mechanics*, Vieweg, Berlin, 1987 to be published by Vieweg, Berlin, 1988.
13. W. L. Wendland, *Elliptic systems in the Plane*, Pitman, 1979.
14. G. F. Pinder and W. G. Gray, *Finite Element Simulation in Surface and Subsurface Hydrology*, Academic Press, New York, 1977.
15. M. S. Engelman, R. L. Sani and P. M. Gresho, 'The implementation of normal and/or tangential boundary conditions in finite element codes for incompressible fluid flow', *Int. j. numer. methods Fluids*, **2**, 225–238 (1982).
16. S. C. Eisenstat, 'Efficient implementation of a class of preconditioned conjugate gradient methods', *Siam J. Sci. Stat. Comput.*, **2**, 1–4 (1981).
17. E. F. Kaasschieter, 'The solution of non-symmetric linear systems by bi-conjugate gradients or conjugate gradients squared', *Report 86-21*, Department of Mathematics, Delft University, 1986.