INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 8, 957–964 (1988)

ON THE ACCURACY OF LEAST-SQUARES FINITE ELEMENTS FOR A FIRST-ORDER CONSERVATION EQUATION

P. WILDERS

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

SUMMARY

The numerical solution of a single first-order conservation equation by a least-squares finite element method is considered. Isoparametric bilinear quadrilateral elements are used. The accuracy is studied numerically and it is shown that the discrete equations associated with nodal points on the boundaries should be modified in order to obtain an accurate numerical solution.

KEY WORDS Conservation equation Hyperbolic equation Least squares Finite elements

1. INTRODUCTION

We consider the equation

$$\frac{\partial u\omega}{\partial x_1}+\frac{\partial v\omega}{\partial x_2}=f, \tag{1}$$

where u, v and f are given functions. We apply least-squares finite elements for the numerical solution of (1). It is well known that numerical methods for (1) often lead to linear systems which are difficult to solve numerically. In the least-squares approach (1) is embedded in a second-order equation. In the discrete approximation of this equation linear systems occur with symmetrical and positive definite matrices and those systems are solved more easily. This is the main reason for investigating the least-squares approach.

The least-squares method has been applied for the numerical solution of the steady-state Euler equations.[1,2] For these equations related embedding methods have been investigated.[3,4] We have applied the method of Bruneau *et al.*[2] to obtain numerical solutions of the subcritical steady-state shallow water equations. However, we have encountered some severe problems, which most likely are caused by an inaccurate treatment of boundaries. None of the above authors refers to the question of the accuracy of the least-squares method. The steady-state shallow water equations are of the composite type (both real and complex characteristics occur) and we attribute the above mentioned inaccuracies to the occurrence of a real characteristic. We therefore propose to study the hyperbolic model problem (1).

Equation (1) has been studied previously by Chattot *et al.*[5] The authors apply the least-squares approach and compare finite element and finite difference methods for the associated second-order equation. We study the finite element method more thoroughly and shall show that this

0271–2091/88/080957–08$05.00

Received March 1986
Revised July 1986

method is $O(h)$ accurate owing to inaccuracy of the numerical approximation along characteristic boundaries. We propose a modified boundary scheme leading to more accurate results.

In Section 2 we give a brief account of the least squares finite-element method. We apply the so called product approximation or group formulation. This means that terms like $u\omega$ are approximated by $\Sigma(u\omega)_j\phi_j$. For us, the main reason is that the product approximation leads to efficient codes.[6,7] Chattot *et al.*[5] have shown that the product approximation is vital for obtaining conservative schemes. For more details and further references the reader may consult Fletcher.[6] In Section 3 we present some numerical results that illustrate the $O(h)$ accuracy of the finite elements. In Section 3 we also investigate the truncation error of the scheme along characteristic boundaries and present a method for correcting the truncation error. This method clearly illustrates the cause of the $O(h)$ accuracy of the original scheme. Finally in Section 4 we present a method in which the boundary scheme is based on characteristic co-ordinates. Numerical experiments show that this new boundary scheme is promising.

2. THE LEAST-SQUARES APPROACH

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

$$I(\omega)=\int_\Omega \left(\frac{\partial u\omega}{\partial x_1}+\frac{\partial v\omega}{\partial x_2}-f\right)^2 \mathrm{d}\Omega. \tag{2}$$

Let ω be a minimum of I and let τ be an arbitrary test function, then we have

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} I(\omega+\varepsilon\tau)|_{\varepsilon=0}=0$$

or

$$\int_\Omega \left(\frac{\partial u\omega}{\partial x_1}+\frac{\partial v\omega}{\partial x_2}-f\right)\left(\frac{\partial u\tau}{\partial x_1}+\frac{\partial v\tau}{\partial x_2}\right)\mathrm{d}\Omega=0. \tag{3}$$

Equation (3) is approximated in the finite-dimensional subspace spanned by isoparametric bilinear quadrilateral elements ϕ_i. As has been mentioned in Section 1, we adopt the product approximation and find

$$\sum_j [u_i(u_j s_{ij}^{(1,1)}+v_j s_{ij}^{(1,2)})+v_i(u_j s_{ij}^{(2,1)}+v_j s_{ij}^{(2,2)})]\,\omega_j=\sum_j u_i f_j d_{ij}^{(1)}+v_i f_j d_{ij}^{(2)}, \tag{4}$$

where

$$s_{ij}^{(k,l)}=\int_\Omega \frac{\partial\phi_i}{\partial x_k}\frac{\partial\phi_j}{\partial x_l}\mathrm{d}\Omega, \qquad d_{ij}^{(k)}=\int_\Omega \frac{\partial\phi_i}{\partial x_k}\phi_j\,\mathrm{d}\Omega; \quad k,l=1,2. \tag{5}$$

For the computation of the integrals in (5) we apply a Newton–Cotes integration formula.

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

$$-\int_\Omega \left(u\frac{\partial}{\partial x_1}+v\frac{\partial}{\partial x_2}\right)\left(\frac{\partial u\omega}{\partial x_1}+\frac{\partial v\omega}{\partial x_2}-f\right)\tau\,\mathrm{d}\Omega+\int_{\partial\Omega}\left(\frac{\partial u\omega}{\partial x_1}+\frac{\partial v\omega}{\partial x_2}-f\right)(un_1+vn_2)\tau\,\mathrm{d}s=0. \tag{6}$$

We conclude that equation (3) is the Galerkin equation associated with the second-order equation

$$-\left(u\frac{\partial}{\partial x_1}+v\frac{\partial}{\partial x_2}\right)\left(\frac{\partial u\omega}{\partial x_1}+\frac{\partial v\omega}{\partial x_2}-f\right)=0, \tag{7}$$

with boundary conditions that are either essential ($\tau=0$) or state that (1) should be satisfied on part of the boundary. At the end of this section we discuss the boundary conditions in more detail.

On a Cartesian grid (4) leads to a nine-point approximation of (7). This nine-point approximation is well known and appropriate for elliptic equations. However (7) is a parabolic equation with characteristic eigenvalue $\lambda=v/u$. Or, using modern terminology, (7) is hyperbolic, because written as a system of two first-order equations we find a single eigenvalue $\lambda=v/u$ with multiplicity two (both algebraically and geometrically). We remark that (7) is parabolic in a special way. As an example we choose $u=v=1$ and $f=0$. We introduce the canonical co-ordinates $\xi=x_1-x_2$, $\eta=x_1+x_2$ and (7) becomes

$$\frac{\partial^2\omega}{\partial\eta^2}=0. \tag{8}$$

This means that (7) is not of the 'heat conduction type', because no first-order derivatives in ξ occur.

In the case of (1) we have three possible types of boundary, namely inflow, outflow and characteristic. Only at inflow may one prescribe ω. Thus the only available essential boundary condition is at inflow. In the case of (7) we have the same possible types of boundary. However, both at inflow and outflow one should give a boundary condition. From (6) it follows that the least-squares approach automatically generates a natural boundary condition (or Neumann condition) at outflow. This condition states that (1) should be satisfied on the outflow boundary.

3. NUMERICAL RESULTS AND TRUNCATION ERROR ANALYSIS

We consider a test problem on $\Omega=\langle 0,1\rangle\times\langle 0,1\rangle$, i.e.

$$\begin{cases} u=(\alpha+1)e^{x_2}, & \alpha=0,1,\\ v=-x_2e^{x_2}, & f=0,\\ \omega=x_2^{\alpha}e^{x_1-x_2}. \end{cases} \tag{9}$$

We use a Cartesian equidistant grid. The nodal points have indices i and j, where $i,j=0,\ldots,n$. We choose $n=4, 8, 16, 32$. The numerical solution is denoted by $\bar{\omega}$. We set

$$e_\infty(n)=\max_{i,j=0,\ldots,n}|\omega_{i,j}-\bar{\omega}_{i,j}|, \tag{10}$$

$$e_2(n)^2=\frac{1}{(n+1)^2}\sum_{i,j=0}^{n}(\omega_{i,j}-\bar{\omega}_{i,j})^2. \tag{11}$$

We generically denote these norms by $e_\bullet$, $\bullet=\infty, 2$. We define

$$p_{2n}={}^2\log\left(\frac{e_\bullet(nn)}{e_\bullet(2n)}\right), \quad n=4, 8, 16, \quad \bullet=\infty, 2. \tag{12}$$

In the case of (9) we have two inflow, one outflow and one characteristic boundary. At inflow we prescribe ω. In Table I we present the computed values of $e_\bullet(4)$ and p_{2n}. From this table we see that for $\alpha=1$ the finite element method is only $O(h)$ accurate in the maximum norm. If we compare the cases $\alpha=0$ and $\alpha=1$, we see a completely different behaviour of the error. Inspection of (9) shows that this difference is possibly associated with the properties near the characteristic boundary.

Table I. Values of e. (4) and p_{2n} for (9)

	$\alpha=0$		$\alpha=1$	
	e_∞	e_2	e_∞	e_2
e. (4)	0·11e−2	0·44e−3	0·26e−1	0·78e−2
p_8	2·0	2·0	1·0	1·4
p_{16}	2·0	1·9	1·0	1·4
p_{32}	2·0	2·0	1·0	1·4

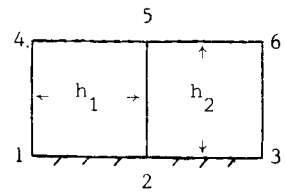

Figure 1. Molecule on boundary

The molecule near the characteristic boundary is represented in Figure 1. Equation (4) in the boundary point 2 reads

$$-\frac{h_2}{2h_1}u_2(u_1\omega_1-2u_2\omega_2+u_3\omega_3)-\frac{1}{4}u_2(v_1\omega_1-v_3\omega_3+v_6\omega_6-v_4\omega_4)$$
$$-\frac{1}{4}v_2(u_3\omega_3-u_1\omega_1+u_6\omega_6-u_4\omega_4)-\frac{h_1}{h_2}v_2(v_5\omega_5-v_2\omega_2)$$
$$=-\frac{1}{4}h_2u_2(f_3-f_1)-\frac{1}{2}h_1v_2(f_2+f_5). \tag{13}$$

Let ω be a smooth solution of (1) (or (7)), let $v_2=0$ (characteristic boundary) and let $h_1=O(h_2)$, then the truncation error $e(\omega)$ of (13) reads

$$e(\omega)=-\frac{1}{4}h_2u\frac{\partial^3 v\omega}{\partial x_1\partial x_2^2}+O(h_2^2). \tag{14}$$

In the case of (9) we have $e(\omega)=O(h_2^2)$ for $\alpha=0$ and $e(\omega)=O(h_1)$ for $\alpha=1$, which motivates a closer look.

For solutions of (1) it is possible to correct the truncation error on the molecule of Figure 1. We have to remark that this is not possible for general solutions of (7). However, we are only interested in those solutions of (7) which are a solution of (1) as well. From (1) it follows that

$$\frac{\partial^3 v\omega}{\partial x_1\partial x_2^2}=\frac{\partial^2 f}{\partial x_1\partial x_2}-\frac{\partial^3 u\omega}{\partial x_1^2\partial x_2}, \tag{15}$$

and a discrete version of the right-hand side may be used to correct the truncation error. We add to (13) two extra terms, α_l and α_r (l=left, r=right-hand side), given by

$$\alpha_l=\frac{1}{4}\frac{h_2}{h_1}u_2(u_1\omega_1-2u_2\omega_2+u_3\omega_3-u_4\omega_4+2u_5\omega_5-u_6\omega_6),$$

$$\alpha_r=\frac{1}{8}h_2u_2(f_3-f_1+f_4-f_6).$$

Table II. Corrected truncation error for (9) with $\alpha = 1$

	e_∞	e_2
$e.(4)$	0·81e−2	0·28e−2
p_8	1·9	1·9
p_{16}	1·9	2·0
p_{32}	2·0	2·0

We call this scheme the 'corrected truncation error' boundary scheme. In Table II we present the computed values of $e.(4)$ and p_{2n} for (9) with $\alpha = 1$.

If we compare Tables I and II, we see that the boundary correction improves the behaviour of the error considerably and we conclude that it is indeed the boundary treatment which triggers inaccuracies in the original scheme (4). The 'corrected truncation error' boundary scheme is not easily applied in more difficult situations, and in the next section we consider a different and more general approach.

4. A BOUNDARY SCHEME BASED ON CHARACTERISTIC CO-ORDINATES

In Section 3 it has been found that the original scheme (4) is inaccurate owing to its treatment of characteristic boundaries. In this section we investigate a new boundary scheme based on characteristic co-ordinates. It is not very difficult to implement this boundary scheme in a finite element code. The resulting linear system fails to be symmetrical. We therefore discuss the iterative solution of this system as well.

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

$$\frac{dx_1}{dt} = u, \qquad \frac{dx_2}{dt} = v. \tag{16}$$

The characteristic form of (7) along Γ reads

$$-\frac{d}{dt}\left(\frac{d\omega}{dt} + (u_{x_1} + v_{x_2})\omega - f\right) = 0. \tag{17}$$

Let s denote the arc length along Γ. We have

$$\frac{ds}{dt} = q, \qquad q^2 = u^2 + v^2. \tag{18}$$

The Galerkin approach applied to (17) gives

$$\int_\Gamma \left(q\frac{d\omega}{ds} + (u_{x_1} + v_{x_2})\omega - f\right)\frac{dq\tau}{ds}\,ds = 0, \tag{19}$$

where $\tau = \tau(s)$ is a test function. For the derivation of (19) we note that we have either an essential or a natural boundary condition (see Section 2). The original basis functions ϕ_i introduced in Section 2 are suitable for the discrete representation of (19), because Γ is a part of $\partial\Omega$. We obtain

$$\sum_j q_i(a_{ij} + b_{ij})\omega_j = \sum_j q_i c_{ij}, \tag{20}$$

with

$$a_{ij}=\int_\Gamma q\frac{d\phi_i}{ds}\frac{d\phi_j}{ds}\,ds, \tag{21}$$

$$b_{ij}=\int_\Gamma (u_{x_1}+v_{x_2})\frac{d\phi_i}{ds}\,\phi_j\,ds, \tag{22}$$

$$c_{ij}=\int_\Gamma f\frac{d\phi_i}{ds}\,\phi_j\,ds. \tag{23}$$

For the computation of a_{ij}, b_{ij} and c_{ij} we apply a Newton–Cotes integration formula. We recall that both q and f are given functions. Furthermore, for the computation of b_{ij} we use

$$b_{ij}=\int_\Gamma \sum_k \left(u_k\frac{\partial\phi_k}{\partial x_1}+v_k\frac{\partial\phi_k}{\partial x_2}\right)\frac{d\phi_i}{ds}\,\phi_j\,ds. \tag{24}$$

We remark that for the test problem, already considered in Table II, the characteristic boundary scheme gives (on the level of the chosen presentation in this table) the same results as the 'corrected truncation error' boundary scheme. We also consider a more difficult test problem. We define

$$r(x_1)=\beta(e^{x_1}-1), \qquad \beta=\frac{\alpha}{e-1}, \tag{25}$$

$$\Omega=\{(x_1,x_2)\colon 0\leqslant x_1\leqslant 1,\ r(x_1)\leqslant x_2\leqslant 1\}. \tag{26}$$

We take

$$\begin{cases} u=1, \\ \omega=\cos(\pi x_1)\sin(\pi x_2). \end{cases} \qquad v=(x_2+\beta)\frac{1-x_2}{1-r(x_1)}, \tag{27}$$

The function f is chosen such that the given ω is the exact solution of (1). We choose $\alpha=0$, 0·25. The mesh is equidistant in the x_1-direction, with $h_1=1/n$, where $n=4$, 8, 16, 32. In the x_2-direction we take n elements as well.

From (25)–(27) we see that $\Gamma_1=\{(x_1,x_2)\colon x_2=r(x_1)\}$ and $\Gamma_2=\{(x_1,x_2)\colon x_2=1\}$ are characteristic boundaries. On these boundaries we apply (20). We remark that for $\alpha=0{\cdot}25$ Γ_1 is curved. On the inflow boundary we prescribe ω. In Table III we present the computed values of $e.(4)$ and p_{2n}. We see that the asymptotic h dependence of the error has improved considerably. We also see

Table III. Computed values of $e.(4)$ and p_{2n}

	$\alpha=0$				$\alpha=0{\cdot}25$			
	Original scheme (4)		Characteristic scheme on Γ		Original scheme (4)		Characteristic scheme on Γ	
	e_∞	e_2	e_∞	e_2	e_∞	e_2	e_∞	e_2
$e.(4)$	0·19e0	0·98e−1	0·18e0	0·73e−1	0·19e0	0·10e0	0·18e0	0·71e−1
p_8	0·8	1·2	1·5	1·5	1·0	1·3	1·5	1·6
p_{16}	0·9	1·4	1·8	1·8	0·9	1·5	1·8	1·8
p_{32}	1·0	1·5	2·0	1·9	0·9	1·5	1·9	1·9

Table IV. Number of iterations in the preconditioned CGS method

Number of unknowns	Original scheme (4)	Characteristic scheme of Γ
272 ($n=16$)	20	17
1056 ($n=32$)	37	32

little difference between the cases $\alpha=0$ and $\alpha=0{\cdot}25$. This means that the method seems to work well on non-Cartesian grids.

Because we use a special boundary scheme, the resulting linear system fails to be symmetrical. This implies that the iterative solution deserves some attention. We apply the CGS method (CGS = conjugate gradients squared) recently developed by Sonneveld.[8] Preconditioning is done with the aid of an incomplete decomposition with corrections only on the main diagonal. For details the reader is referred to References 9–11. We also apply this method to solve the linear system resulting from the original scheme (4). In this case the matrix is symmetrical and positive definite and the CGS method reduces to a variant of the conjugate gradients method (roughly speaking, two CG iterations = one CGS iteration). In Table IV we present the number of iterations in the case of test problem (25) with $\alpha=0{\cdot}25$ for $n=16, 32$. For the termination criterion we have used $\|\text{preconditioned residual}\|_2 \leqslant 1{\cdot}0\text{e}-4$.

We conclude that the 'iterative properties' of the matrices do not differ significantly. Finally we report that in some cases the ordinary incomplete decomposition of the matrices failed to exist. The reason is that on the subdiagonals entries occur with both positive and negative values.

5. CONCLUSION

It has been shown that the accuracy of the numerical solution of a first-order conservation equation by a least-squares finite element method is disappointing, but that the method can be improved considerably by applying a special boundary scheme on characteristic boundaries. A general formulation of this boundary scheme has been given and it has been shown that this new scheme works well, even in the case of curved boundaries and non-Cartesian grids. It has been found that an iterative solution of the resulting system can be provided by a preconditioned conjugate gradients type of method, the so called CGS method. This study is the first step towards the development of an accurate solution method of the steady shallow water equations by means of a least-squares approach. Special boundary schemes for systems of equations of the composite type are in progress and will be the subject of future publications.

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. 8th Int. Conf. on Numerical Methods in Fluids*, Springer, 1982.
3. G. M. Johnson, 'Relaxation solution of the full Euler equations', in E. Krause (ed.), *Proc. 8th Int. Conf. on Numerical Methods in Fluids*, Springer, 1982.
4. S. Chang and G. M. Johnson, 'An embedding method for the steady Euler equations', *J. Comput. Phys.*, **63**, 191–200 (1986).
5. J. J. Chattot, J. Guiu-Roux and J. Laminie, 'Numerical solution of a first-order conservation equation by a least square method', *Int. j. numer. methods fluids*, **2**, 209–219 (1982).

6. C. A. J. Fletcher, 'The group finite element formulation', *Comput. Methods Appl. Mech. Eng.*, **37**, 225–243 (1983).
7. 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).
8. 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. Statist. Comput.*
9. S. C. Eisenstat, 'Efficient implementation of a class of preconditioned conjugate gradient methods', *SIAM J. Sci. Statist. Comput.*, **2**, 1–4 (1981).
10. 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).
11. E. F. Kaasschieter, 'The solution of non-symmetric linear systems by biconjugate gradients or conjugate gradients squared', *Report 86–21*, Department of Mathematics, Delft University, 1986.