M2AN - MODÉLISATION MATHÉMATIQUE ET ANALYSE NUMÉRIQUE

K. W. MORTON
M. STYNES

An analysis of the cell vertex method

M2AN - Modélisation mathématique et analyse numérique, tome 28, n° 6 (1994), p. 699-724

<http://www.numdam.org/item?id=M2AN_1994__28_6_699_0>

MATHEMATICAL MODELLING AND NUMERICAL ANALYSIS
MODÉLISATION MATHÉMATIQUE ET ANALYSE NUMÉRIQUE

(Vol 28, n° 6, 1994, p 699 à 724)

AN ANALYSIS OF THE CELL VERTEX METHOD (*)

by K. W. MORTON (1) and M. STYNES (2)

Communicated by R TEMAM

Abstract — *We present a new analysis of the cell vertex finite volume method, based on the construction of a mapping from the trial space to the test space of the method For a convection-diffusion problem in one dimension, we obtain an error estimate for our computed solution (in a weighted discrete H^1 norm) which depends only on the gradient nodal interpolation errors For pure convection in two dimensions, we use a new natural seminorm to prove local and global error estimates for the cases of flow transverse to the grid and flow parallel to the grid*

1. INTRODUCTION

The cell vertex finite volume method, together with its earlier form as the box difference scheme, has been widely used to approximate first order differential equations. In particular, its advantages for discretising the Euler equations of inviscid gas dynamics are now well recognised. Recently, in [4], [5] and [8], extensions have been proposed for the Navier-Stokes equations of viscous gas dynamics and impressive results obtained for model convection-diffusion problems. Some error analysis in one dimension was carried out by Mackenzie and Morton [5], mainly using finite difference techniques, and also by Morton and Süli [9] and Süli [12], [11] for the two-dimensional pure convection problem. These demonstrate some of the key features of the method, but it is clear from these papers that more general and more powerful methods of analysis are required.

(*) Manuscript received May 15, 1992, revised July 19, 1993

The work reported here forms part of the research programme of the Oxford-Reading Institute for Computational Fluid Dynamics

Partial financial support for the second author was provided by the Royal Society, London, the Royal Irish Academy, Dublin, and Oxford University Computing Laboratory

(1) Numerical Analysis Group, University of Oxford

(2) Mathematics Department, University College, Cork, Ireland

M² AN Modélisation mathématique et Analyse numérique 0764-583X/94/06/$ 4 00
Mathematical Modelling and Numerical Analysis

In the present paper we put forward an alternative analysis of the cell vertex scheme which seems to hold considerable promise It is based on mesh-dependent norms similar to those used by Suli [12], [11], within a nonconforming Petrov-Galerkin framework, but the key idea is to introduce a mapping between the trial and test spaces akin to the upwinded test functions of finite element methods or the upwinded control volumes of finite volume methods There is clearly a parallel with the approximate symmetrization technique of Barrett and Morton [1] , but instead of enhancing the symmetry of the bilinear form it aims merely to improve its positive-definiteness

In the next section the general approach is outlined and motivated Then in section 3 a particular mapping is used to analyse the one-dimensional convection-diffusion problem Section 4 shows how a similar mapping can materially sharpen the results given in [9] and also highlights the importance of a particular semi-norm in the analysis Finally, in section 5 it is shown how the need to choose a class of approximations for which this semi-norm becomes a norm leads to a natural solution of the convection problem with characteristic boundaries

2. OUTLINE OF ERROR ANALYSIS

We give here a brief sketch of the argument used in later sections to provide error analyses for the convection and convection-diffusion problems

We assume that the domain Ω of our differential equation $Lu = f$ is suitably divided by a given mesh (intervals in one dimension, quadrilaterals in two dimensions, etc) Our computed solution U will lie in the associated trial space S^h, which consists of piecewise linears in one dimension, piecewise isoparametric bilinears in two dimensions, etc This solution is defined by requiring U to satisfy

$$B(U, p) = (f, p) \quad \forall p \in T^h , \tag{2 1}$$

together with Dirichlet boundary conditions on U, where $B(\,.\,,\,.\,)$ $S^h \times T^h \to R$ is a bilinear form associated with the differential operator L, $(\,.\,,\,.\,)$ is the $L^2(\Omega)$ inner product, and T^h is the space of piecewise constants on the mesh

The key idea in the analysis is to construct a mapping $M \quad S^h \to T^h$ such that

$$B(V, MV) \geqslant C_1 \|V\|_h^2 \quad \forall V \in S^h , \tag{2 2}$$

where C_1 is some fixed positive constant (independent of the mesh and of the diffusion coefficient in L), and $\|\,.\,\|_h$ is some norm or semi-norm which is sufficiently strong to guarantee stability of the numerical method Then,

writing u^I for the interpolant to u from S^h, one has to extend the definition of $B(.,.)$ to a suitably smooth class of solutions u so that one can carry out the following argument :

$$\begin{aligned} C_1 \|u^I - U\|_h^2 &\leq B(u^I - U, M(u^I - U)) \\ &= B(u^I - u, M(u^I - U)) + B(u - U, M(u^I - U)) \\ &= B(u^I - u, M(u^I - U)), \end{aligned} \tag{2.3}$$

since $B(u, p) = (f, p) = B(U, p)$ for all $p \in T^h$.

Finally, the right-hand side of (2.3) can be expressed as a combination of $\|u^I - U\|_h$ and terms depending on $u^I - u$. The $\|u^I - U\|_h$ terms can be absorbed into the left-hand side of (2.3), leading to a bound on $\|u^I - U\|_h$ in terms of $u^I - u$.

This is an extension of the approach used in Stynes and O'Riordan [10] to analyse finite element methods for singularly perturbed two-point boundary value problems.

An insight into how one might construct a mapping M satisfying (2.2) above is provided by the following heuristic calculation. Consider the two-point boundary value problem

$$-\varepsilon u'' + au' = f \quad \text{on} \quad (0, 1), \tag{2.4}$$

$$u(0) = u(1) = 0, \tag{2.5}$$

where ε is a small positive parameter and a is a positive constant. Then the bilinear form $B(.,.)$ is essentially given by

$$B(v, w) = \int_0^1 (-\varepsilon v''(x) + av'(x))\, w(x)\, dx$$

$$\forall v \in W^{2,\infty}(0, 1), \quad \forall w \in L^1(0, 1). \tag{2.6}$$

In practice we shall only work with functions w which vanish on the interval $(1 - h, 1]$, where h is some positive constant (h will be the mesh diameter in our full analysis later). Suppose now that

$$Mu(x) = \begin{cases} u(x + h) & \text{for} \quad 0 \leq x \leq 1 - h, \\ 0 & \text{for} \quad 1 - h < x \leq 1, \end{cases} \tag{2.7}$$

so that the mapping M has the effect of translating the function u to the left, i.e. in the upwind direction since $a > 0$. Then we have

$$\begin{aligned} B(u, Mu) &= \int_0^{1-h} (-\varepsilon u''(x) + au'(x))\, u(x + h)\, dx \\ &= \varepsilon u'(0)\, u(h) + \int_0^{1-h} [\varepsilon u'(x)\, u'(x + h) + au'(x)\, u(x + h)]\, dx \end{aligned}$$

$$\approx \varepsilon u'(0)\, u(h) + \int_0^{1-h} [\varepsilon u'(u' + hu'') + au'(u + hu')]\, dx$$

$$= (\varepsilon + ah) \int_0^{1-h} (u')^2\, dx + \frac{1}{2} [\varepsilon h(u'(1-h))^2 + a(u(1-h))^2]$$

$$+ \varepsilon u'(0)\left[u(h) - \frac{1}{2} hu'(0)\right]. \qquad (2.8)$$

For $h > \varepsilon$, it turns out that the term

$$ah \int_0^{1-h} (u')^2\, dx \qquad (2.9)$$

guarantees stability (as is well-known in the analysis of the streamline diffusion method, for example). This indicates that, if we choose Mv to be a function which is essentially v upwinded, then we have grounds for expecting a satisfactory bound as in (2.2) above. Note that if in (2.7) we had used $u(x)$ instead of $u(x + h)$ — that is, if we had *not* upwinded u — then the lower bound obtained would have contained only

$$\varepsilon \int_0^{1-h} (u')^2\, dx\,, \qquad (2.10)$$

and this would be insufficient for stability.

The above calculation pertains directly to our convection-diffusion problem. In the case of pure convection, replacing $u(x + h)$ in (2.7) by $u'(x)$ leads more simply to a satisfactory result, as we shall see in section 4. However, irrespective of whether diffusion is present we follow essentially the same line of argument ; for further discussion on the use of upwinding in some form as a means of obtaining stability in finite element and finite volume methods, see also Morton [6], [7].

3. CONVECTION-DIFFUSION IN ONE DIMENSION

We shall in this section obtain an error bound for the solution obtained when the cell vertex finite volume method is applied to a singularly perturbed two-point boundary value problem. Our estimate is in a norm which is a discrete analogue of (2.8) above.

Consider the problem

$$-\varepsilon u'' + (au)' = f \quad \text{on} \quad (0, 1)\,, \qquad (3.1a)$$

$$u(0) = u_L\; u(1) = u_R\,, \qquad (3.1b)$$

where ε is a small positive parameter, and we assume that the function a is

smooth and satisfies

$$a(x) > \alpha > 0\,, \tag{3.2}$$

$$a'(x) \geq \beta \geq 0 \tag{3.3}$$

on [0, 1]. The condition (3.2) guarantees that the solution u of (3.1) can have a layer only at the boundary $x = 1$. Condition (3.3) will be needed later to ensure the stability of our numerical method ; it is the usual finite element condition « $b - a'/2 \geq 0$ » that would be applied to an equation in the form $-\varepsilon u'' + au' + bu = f$, and we note that (*cf.* Stynes and O'Riordan [10]) it can be deduced from (3.2) by making, if necessary, a change of dependent variable.

Place an arbitrary mesh $0 = x_0 < x_1 < \cdots < x_N = 1$ on [0, 1]. Set $h_i = x_i - x_{i-1}$ for each i, and $H = \max_i h_i$. For any function $g \in C[0, 1]$, we write g_i for $g(x_i)$.

Our finite volume scheme is Method B of Mackenzie [4] : find the piecewise linear function U such that

$$B_i(U) := -\varepsilon(U_i' - U_{i-1}') + (aU)_i - (aU)_{i-1} = \int_{x_{i-1}}^{x_i} f\,dx\,, \quad \text{for} \quad i = 1, \ldots, N-1\,, \tag{3.4}$$

$$U_0 = u_L \qquad U_N = u_R\,, \tag{3.5}$$

where for each piecewise linear V we set

$$V_j' = \frac{1}{(h_j + h_{j+1})}\,(h_{j+1} D_- V_j + h_j D_- V_{j+1}) \quad \text{for } j = 1, \ldots, N-1 \tag{3.6a}$$

and

$$V_0' = 2 D_- V_1 - V_1'\,. \tag{3.6b}$$

Here D_- is the backward divided difference operator. We assume that each integral $\int_{x_{i-1}}^{x_i} f\,dx$ is evaluated exactly.

Let S_0^h denote the $(N-1)$-dimensional space of continuous piecewise linear functions on the mesh which vanish at $x = 0$ and $x = 1$; and let S_E^h denote the corresponding set of piecewise linear functions satisfying the boundary conditions (3.1*b*). Let T^h denote the space, of the same dimension as S_0^h, consisting of functions which are piecewise constants on $[0, x_{N-1}]$ and which vanish on $(x_{N-1}, 1]$. Given $W \in T^h$, let W_i be the value of W on each interval (x_{i-1}, x_i). For each $V \in S_E^h$ and each $W \in T^h$, set

$$B(V, W) := \sum_{i=1}^{N-1} W_i B_i(V)\,. \tag{3.7}$$

Then (3.4) is equivalent to

$$B(U, W) = \sum_{i=1}^{N-1} W_i \int_{x_{i-1}}^{x_i} f \, dx = \int_0^1 Wf \, dx \quad \forall W \in T^h . \tag{3.8}$$

We also define $B(\,.\,,\,.\,) : C^1[0, 1] \times T^h \to R$ by

$$B(v, W) = \sum_{i=1}^{N-1} W_i \{ -\varepsilon(v'(x_i) - v'(x_{i-1})) + (av)_i - (av)_{i-1} \} . \tag{3.9}$$

Note that the two definitions (3.7) and (3.9) are consistent, being identical for $C^1[0, 1] \cap S_E^h$ which consists of the linear function $u_L(1 - x) + u_R x$, so there is no ambiguity in the definition of B. We thus have $B(\,.\,,\,.\,)$ defined on $(S_E^h \oplus C^1[0, 1]) \times T^h$ by $B(V + v, W) = B(V, W) + B(v, W)$.

We deduce from (3.1), (3.9) and (3.8) that

$$B(u - U, W) = 0 \quad \forall W \in T^h . \tag{3.10}$$

Also, using (3.7) and (3.9),

$$B(u - u^I, W) = -\varepsilon \sum_{j=1}^{N-1} W_j h_j D_- (u'(x_j) - (u^I)'_j) \tag{3.11}$$

since $(au)_j = (au^I)_j$ for each j.

We can now specify a suitable mapping M, as discussed in section 2. Given $V \in S_0^h$, define $MV \in T^h$ by

$$(MV)(x) = V_i \quad \text{on} \quad (x_{i-1}, x_i) \quad \text{for each} \quad i . \tag{3.12}$$

The function MV may be regarded as the limiting case, as the cell Péclet number tends to infinity, of the well-known Hemker test function [2], which is the local Green's function.

Using this M, we now prove a coercivity inequality which is a discrete analogue of (2.8) for our cell vertex method.

LEMMA 3.1 : *Let $V \in S_0^h$ be arbitrary. Assume that the mesh is arbitrarily graded, viz., $h_j \geq h_{j+1}$ for $j = 1, \ldots, N - 1$. Assume also that $h_2 \geq h_1/4$. Then*

$$B(V, MV) \geq (\alpha/2) \sum_{j=1}^{N} (V_j - V_{j-1})^2 + (\beta/2) \sum_{j=1}^{N} h_j V_j^2 . \tag{3.13}$$

Proof : By (3.7),

$$B(V, MV) = \varepsilon \sum_{j=2}^{N-1} (V'_{j-1} - V'_j) V_j - \frac{2 \varepsilon h_1 V_1}{h_1 + h_2} (D_- V_2 - D_- V_1) + \\ + \sum_{j=1}^{N-1} [(aV)_j - (aV)_{j-1}] V_j . \tag{3.14}$$

We begin by analysing the terms which are multiplied by ε. From summation by parts and $V_0 = V_N = 0$,

$$\sum_{j=2}^{N-1} (V'_{j-1} - V'_j) V_j = V'_1 V_2 + \sum_{j=2}^{N-1} V'_j (V_{j+1} - V_j)$$

$$= h_1 V'_1 D_- V_1 + \sum_{j=1}^{N-1} h_{j+1} V'_j D_- V_{j+1} . \quad (3.15)$$

Hence the « ε-terms » consist of

$$-\frac{2 h_1^2}{h_1 + h_2} D_- V_1 (D_- V_2 - D_- V_1) + h_1 V'_1 D_- V_1 + \sum_{j=1}^{N-1} h_{j+1} V'_j D_- V_{j+1}$$

and we show that on a graded mesh the definition of V'_j renders this sum positive. Substitution for V'_j from (3.6*a*) yields a sum of the form

$$\sum_{j=1}^{N} a_j (D_- V_j)^2 + \sum_{j=1}^{N-1} b_j D_- V_j D_- V_{j+1} , \quad (3.16)$$

where

$$a_1 = \frac{2 h_1^2 + h_1 h_2}{h_1 + h_2} , \quad a_j = \frac{h_{j-1} h_j}{h_{j-1} + h_j} \quad \text{for} \quad j = 2, \ldots, N \quad (3.17)$$

and

$$b_1 = h_2 - h_1 , \quad b_j = \frac{h_{j+1}^2}{h_j + h_{j+1}} \quad \text{for} \quad j = 2, \ldots, N - 1 .$$

Using the mesh grading, we have for $j = 1, \ldots, N - 1$

$$b_j = \frac{h_{j+1}^2}{h_j + h_{j+1}} \leqslant \frac{h_{j+1}}{2} \leqslant \frac{h_j}{2} \leqslant \frac{h_{j-1} h_j}{h_{j-1} + h_j} = a_j \quad (3.18)$$

and similarly

$$b_j \leqslant a_{j+1} . \quad (3.19)$$

Consequently

$$\sum_{j=2}^{N} a_j (D_- V_j)^2 + \sum_{j=2}^{N-1} b_j D_- V_j D_- V_{j+1} \geqslant$$

$$\geqslant \left(a_2 - \frac{b_2}{2} \right) (D_- V_2)^2 + \sum_{j=3}^{N-1} \left(a_j - \frac{1}{2} (b_{j-1} + b_j) \right) \times$$

$$\times (D_- V_j)^2 + \left(a_N - \frac{1}{2} b_{N-1} \right) (D_- V_N)^2$$

$$\geqslant \frac{1}{2} a_2 (D_- V_2)^2 + \frac{1}{2} a_N (D_- V_N)^2 , \quad (3.20)$$

by (3.18) and (3.19). Combining (3.16) and (3.20),

$$\sum_{j=1}^{N} a_j (D_- V_j)^2 + \sum_{j=1}^{N-1} b_j D_- V_j D_- V_{j+1} \geqslant$$

$$\geqslant a_1 (D_- V_1)^2 + \frac{1}{2} a_2 (D_- V_2)^2 + b_1 D_- V_1 D_- V_2 + \frac{1}{2} a_N (D_- V_N)^2 . \tag{3.21}$$

Now the first three terms here are a quadratic form in $D_- V_1$ and $D_- V_2$, which will be non-negative if and only if $b_1^2 \leqslant 4\, a_1 (a_2/2)$, that is,

$$(h_2 - h_1)^2 \leqslant 2 \left(\frac{2\, h_1^2 + h_1 h_2}{h_1 + h_2} \right) \left(\frac{h_1 h_2}{h_1 + h_2} \right) .$$

Setting $s := h_2/h_1$, this becomes

$$s^4 - 4\, s^2 - 4\, s + 1 \leqslant 0 ,$$

which clearly holds if $s \geqslant 1/4$, since $s^4 - 4\, s^2 < 0$ from the mesh grading.

Thus (3.21) becomes

$$\sum_{j=1}^{N} a_j (D_- V_j)^2 + \sum_{j=1}^{N-1} b_j D_- V_j D_- V_{j+1} \geqslant \frac{1}{2} a_N (D_- V_N)^2 , \tag{3.22}$$

so the ε-terms from (3.15) are non-negative.

Hence we obtain

$$B(V, MV) \geqslant \sum_{j=1}^{N-1} [(aV)_j - (aV)_{j-1}] V_j . \tag{3.23}$$

Using $V_0 = V_N = 0$,

$$\sum_{j=1}^{N-1} [(aV)_j - (aV)_{j-1}] V_j =$$

$$= (1/2) \sum_{j=1}^{N} [a_{j-1} (V_j - V_{j-1})^2 + (a_j - a_{j-1}) V_j^2]$$

$$\geqslant (\alpha/2) \sum_{j=1}^{N} (V_j - V_{j-1})^2 + (\beta/2) \sum_{j=1}^{N} h_j V_j^2 . \tag{3.24}$$

The inequalities (3.23) and (3.24) together imply that

$$B(V, MV) \geqslant (\alpha/2) \sum_{j=1}^{N} (V_j - V_{j-1})^2 + (\beta/2) \sum_{j=1}^{N} h_j V_j^2 , \tag{3.25}$$

as desired. □

We next obtain an expression for $B(u^I - U, M(u^I - U))$.

LEMMA 3.2 : *For* $R = u^I - U$,

$$B(R, MR) = \varepsilon \sum_{j=1}^{N} (R_j - R_{j-1})((u^I)'_{j-1} - u'(x_{j-1})), \tag{3.26}$$

where $(u^I)'_j$ *is given by* (3.6).

Proof : From (3.10) and (3.11), using the undivided backward difference operator Δ_-,

$$\begin{aligned} B(R, MR) &= B(u^I - u, MR) \\ &= \varepsilon \sum_{j=1}^{N-1} R_j \, \Delta_- (u'(x_j) - (u^I)'_j) . \end{aligned} \tag{3.27}$$

Summing (3.27) by parts and noting that $R_0 = R_N = 0$, we obtain (3.26). □

We can now obtain a convergence result in a discrete energy norm.

THEOREM 3.3 : *Assume that* $h_j \geq h_{j+1}$ *for* $j = 1, \ldots, N-1$ *and that* $h_2 \geq h_1/4$. *Then for* $R = u^I - U$

$$\alpha \sum_{j=1}^{N} (R_j - R_{j-1})^2 + \beta \sum_{j=1}^{N} h_j R_j^2 \leq (4 \varepsilon^2/\alpha) \sum_{j=0}^{N-1} (u'(x_j) - (u^I)'_j)^2 . \tag{3.28}$$

Proof : Take $V = R$ in Lemma 3.1 and invoke Lemma 3.2 to obtain

$$\begin{aligned} \alpha \sum_{j=1}^{N} (R_j - R_{j-1})^2 + \beta \sum_{j=1}^{N} h_j R_j^2 &\leq 2 B(R, MR) = \\ &= 2 \varepsilon \sum_{j=1}^{N} (R_j - R_{j-1})((u^I)'_{j-1} - u'(x_{j-1})) \\ &\leq 2 \left[\alpha \sum_{j=1}^{N} (R_j - R_{j-1})^2 \right]^{\frac{1}{2}} \left[\frac{\varepsilon^2}{\alpha} \sum_{j=1}^{N} ((u^I)'_{j-1} - u'(x_{j-1}))^2 \right]^{\frac{1}{2}} \end{aligned} \tag{3.29}$$

by the Cauchy-Schwarz inequality. The first term on the right can be cancelled against the left-hand side to give the required result. □

Remark : The assumptions made about the mesh in Theorem 3.3 are not restrictive in practice, in essence requiring only that the mesh is not coarsened in the boundary layer.

Moreover, provided that one has more information about the mesh, one

can replace the right hand side of (3.28) by a more explicit error bound. We show below how this is done in certain cases.

Towards this end, we first observe that Taylor expansions yield

$$(u^I)'_j - u'(x_j) = \frac{h_j\, h_{j+1}}{6(h_j + h_{j+1})} [h_j\, u'''(\eta_{1,j}) + h_{j+1}\, u'''(\eta_{2,j})] \quad (3.30a)$$

for $j = 1, \ldots, N-1$, where $x_{j-1} < \eta_{1,j} < x_j < \eta_{2,j} < x_{j+1}$, and

$$(u^I)'_0 - u'(x_0) = O(h_1^2)\,. \quad (3.30b)$$

We note also that Kellogg and Tsan [3] have shown, provided $a \in C^2[0, 1]$ and $f \in C^2[0, 1]$, that

$$|u^{(i)}(x)| \leq C\,[1 + \varepsilon^{-i} \exp(-\alpha(1-x)/\varepsilon)] \quad (3.31)$$

for $i = 0, 1, 2, 3$ and $0 < x < 1$, where we use C to denote a generic constant which depends only on a and f. Combining (3.30) and (3.31), we have

$$\begin{aligned}\left|(u^I)'_j - u'(x_j)\right| < {} & C h_j\, h_{j+1}[1 + \varepsilon^{-3}\exp(-\alpha(1-x_j)/\varepsilon)] + \\ & + C h_{j+1}^2[1 + \varepsilon^{-3}\exp(-\alpha(1-x_{j+1})/\varepsilon)] \quad (3.32a)\end{aligned}$$

for $j = 1, \ldots, N-1$ and

$$\left|(u^I)'_0 - u'(x_0)\right| \leq C h_1^2\,. \quad (3.32b)$$

We present two Corollaries of Theorem 3.3. The first Corollary deals with a mesh which resolves the layer near $x = 1$, while the other assumes that the mesh is coarse and does not resolve the layer. The cell vertex method exhibits different convergence properties in these two regimes, as is demonstrated numerically in Mackenzie and Morton [5].

COROLLARY 3.1 : *Assume the same hypotheses as in Theorem 3.3, and that* $a \in C^2[0, 1]$ *and* $f \in C^2[0, 1]$. *Set* $J = \max\{j : x_j \leq 1 - (3\varepsilon/\alpha)\ln(1/\varepsilon)\}$, *and assume that* $h_{J+1} \leq \varepsilon$. *Then, if* $H := \max\{h_j : j = 1, 2, \ldots, N\}$,

$$\alpha \sum_{j=1}^{N} (R_j - R_{j-1})^2 + \beta \sum_{j=1}^{N} h_j\, R_j^2 \leq C H^3\, \varepsilon^2 + C\,(h_{J+1}/\varepsilon)^3\,. \quad (3.33)$$

Proof : First, note that by (3.32) and the definition of J,

$$\left|u'(x_j) - (u^I)'_j\right| \leq C h_j\, h_{j+1} \quad (3.34)$$

for $j < J$. Thus from (3.32),

$$\sum_{j=0}^{N-1} (u'(x_j) - (u^I)'_j)^2 \leqslant$$

$$\leqslant CH^3 \sum_{j=0}^{J-1} h_{j+1} + C \sum_{j=J}^{N-1} [h_j^2 h_{j+1}^2 \varepsilon^{-6} \exp(-2\alpha(1-x_j)/\varepsilon) +$$

$$+ h_{j+1}^4 \varepsilon^{-6} \exp(-2\alpha(1-x_{j+1})/\varepsilon)]$$

$$\leqslant CH^3 + Ch_J^2 h_{J+1}^2 \varepsilon^{-6} \exp(-2\alpha(1-x_J)/\varepsilon) +$$

$$+ Ch_{J+1}^3 \varepsilon^{-6} \sum_{j=J}^{N-1} h_{j+1} \exp(-2\alpha(1-x_{j+1})/\varepsilon)$$

$$\leqslant CH^3 + Ch_J^2 h_{J+1}^2 + Ch_{J+1}^3 \varepsilon^{-6} \sum_{j=J}^{N-1} h_{j+1} \exp(-2\alpha(1-x_j)/\varepsilon), \qquad (3.35)$$

using the definition of J and

$$\exp(-2\alpha(1-x_{j+1})/\varepsilon) = \exp(-2\alpha(1-x_j)/\varepsilon) \exp(2\alpha h_{j+1}/\varepsilon)$$
$$\leqslant C \exp(-2\alpha(1-x_j)/\varepsilon) \qquad (3.36)$$

for $j \geqslant J$, because $h_{J+1} \leqslant \varepsilon$ and the mesh is graded. Regarded as a Riemann sum, we have

$$\sum_{j=J}^{N-1} h_{j+1} \exp(-2\alpha(1-x_j)/\varepsilon) <$$

$$< \int_0^1 \exp(-2\alpha(1-x)/\varepsilon)\, dx < \varepsilon/(2\alpha). \qquad (3.37)$$

From (3.35) and (3.37), we obtain

$$\sum_{j=0}^{N-1} (u'(x_j) - (u^I)'_j)^2 \leqslant CH^3 + Ch_{J+1}^3 \varepsilon^{-5}, \qquad (3.38)$$

and now an appeal to Theorem 3.3 completes the proof. □

COROLLARY 3.2 : *Assume the same hypotheses as in Theorem 3.3, and that* $a \in C^2[0, 1]$ *and* $f \in C^2[0, 1]$. *Assume also that*

$$h_N \geqslant (\varepsilon/\alpha) \ln (1/\varepsilon). \qquad (3.39)$$

Then

$$\alpha \sum_{j=1}^{N} (R_j - R_{j-1})^2 + \beta \sum_{j=1}^{N} h_j R_j^2 \leqslant C (\varepsilon/h_N)^2. \qquad (3.40)$$

Proof : Note that $h_N \geqslant (\varepsilon/\alpha) \ln (1/\varepsilon)$ and (3.31) together imply that

$$|u'(x_j)| \leqslant C \quad \forall j < N. \qquad (3.41)$$

Let J be defined as in Corollary 3.1. Then by the mesh grading,

$$x_{N-3} \leqslant 1 - 3\,h_N \leqslant 1 - 3(\varepsilon/\alpha)\ln(1/\varepsilon), \tag{3.42}$$

so $J \geqslant N - 3$.

As in the proof of Corollary 3.1,

$$\begin{aligned}\sum_{j=0}^{N-1} (u'(x_j) - (u^I)'_j)^2 &\leqslant CH^3 + \sum_{j=J}^{N-1} (u'(x_j) - (u^I)'_j)^2 \\ &\leqslant CH^3 + 2\sum_{j=J}^{N-1} (|u'(x_j)|^2 + |(u^I)'_j|^2) \\ &\leqslant CH^3 + C\sum_{j=J}^{N-1} (1 + h_{j+1}^{-2}),\end{aligned} \tag{3.43}$$

from (3.41), $|u(x)| \leqslant C$ on $[0, 1]$, and (3.6). But the last sum contains at most three terms because $J \geqslant N - 3$, so

$$\sum_{j=0}^{N-1} (u'(x_j) - (u^I)'_j)^2 \leqslant CH^3 + Ch_N^{-2} \leqslant Ch_N^{-2}. \tag{3.44}$$

The result now follows immediately from Theorem 3.3. □

Remark : Corollary 3.2 essentially states that the error in the computed solution, measured in a discrete energy norm, is $O(\varepsilon/h_N)$ when the mesh is coarse. Numerical results in Mackenzie and Morton [5] for the error in the discrete L^∞ norm also exhibit $O(\varepsilon/h_N)$ behaviour when (3.18) is satisfied.

4. CONVECTION IN TWO DIMENSIONS WITH NON-CHARACTERISTIC BOUNDARIES

In this section, we present convergence results for the cell vertex finite volume method when applied to a scalar first-order hyperbolic equation in two independent variables. Morton and Süli [9] have considered this problem using a nonuniform tensor product mesh (see also Süli [11] for a fuller study of the convection problem on more general quadrilateral meshes). Here we obtain some new local and global results for the method on various quadrilateral meshes, and also strengthen the original Morton and Süli result, by further exploiting the use of a mapping $M : S^h \to T^h$.

For the most part, we shall use the same notation as Morton and Süli [9]. Let Ω be a nonempty open convex set in R^2 with piecewise smooth boundary $\partial\Omega$. Let $\mathbf{a} = (a_1\ a_2) : \bar{\Omega} \to R^2$ be a given smooth function with $a_1^2 + a_2^2 > 0$ on $\bar{\Omega}$. Set

$$\begin{aligned}\partial_- \Omega &= \{\mathbf{x} \in \partial\Omega : \mathbf{a}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}, \\ \partial_+ \Omega &= \{\mathbf{x} \in \partial\Omega : \mathbf{a}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \geqslant 0\},\end{aligned} \tag{4.1}$$

where $\mathbf{n}(\mathbf{x})$ denotes the unit outward normal to $\partial\Omega$ at $\mathbf{x} \in \partial\Omega$.

We consider the boundary value problem

$$\nabla \cdot (\mathbf{a}u) = f \quad \text{in} \quad \Omega\,, \tag{4.2a}$$

$$u = 0 \quad \text{on} \quad \partial_- \Omega\,, \tag{4.2b}$$

where $f : \Omega \to R$ lies in $L^2(\Omega)$.

To discretize (4.2), we assume that we have a partition $K = \{K_i : i \in I\}$ of Ω, where I is some index set. Each element K_i of the partition is a convex quadrilateral, and we denote by F_i the affine function which maps the reference square $\hat{K} = (0, 1)^2$ onto K_i. We set $h = \max_i \{\text{diameter } (K_i)\}$. We write $Q_1(\hat{K})$ for the space of bilinear functions on $\hat{K}$, and $Q_0(\hat{K})$ for the space of constant functions on $\hat{K}$. For $\Omega_1 \subseteq \Omega$, let $H^1_-(\Omega_1)$ denote the space of all $v \in H^1(\Omega_1)$ whose trace on $\bar{\Omega}_1 \cap \partial_- \Omega$ is zero. Define

$$\mathscr{U}^h = \left\{v \in H^1(\Omega) : v = \hat{v} \circ F_i^{-1},\ \hat{v} \in Q_1(\hat{K}),\ i \in I\right\}\,,$$

$$\mathscr{U}^h_- = \mathscr{U}^h \cap H^1_-(\Omega)\,,$$

$$\mathscr{M}^h = \left\{p \in L^2(\Omega) : p = \hat{p} \circ F_i^{-1},\ \hat{p} \in Q_0(\hat{K}),\ i \in I\right\}\,.$$

Let $P^h : L^2(\Omega) \to \mathscr{M}^h$ be the orthogonal projector from $L^2(\Omega)$ onto $\mathscr{M}^h$, and let

$$I^h : (H^1_-(\Omega) \cap C(\bar{\Omega}))^2 \to (\mathscr{U}^h_-)^2 \tag{4.3}$$

be the interpolation projector onto $(\mathscr{U}^h_-)^2$.

Define the bilinear form $B(.,.) : H^1_-(\Omega) \cap C(\bar{\Omega}) \times \mathscr{M}^h \to R$ by

$$B(v, p) = (\nabla \cdot I^h(\mathbf{a}v), p)\,, \tag{4.4}$$

where $(.,.)$ is the $L^2(\Omega)$ inner product. We assume for the present that a finite volume approximation U to u exists such that $U \in \mathscr{U}^h_-$ and

$$B(U, p) = (f, p) \quad \forall p \in \mathscr{M}^h. \tag{4.5}$$

(Existence and uniqueness of U will be discussed later in this section.) From (4.2a), (4.4) and (4.5), it follows that

$$(\nabla \cdot (\mathbf{a}u - I^h(\mathbf{a}U)), p) = 0 \quad \forall p \in \mathscr{M}^h. \tag{4.6}$$

Our convergence results will be expressed in terms of certain seminorms which we now define.

For each $K_i \in K$, let $m(K_i)$ denote the area of K_i. For $\tilde{I}$ any nonempty subset of I, let $\tilde{\Omega} = \bigcup_{i \in \tilde{I}} K_i$. Set

$$|v|_{l_2(\tilde{\Omega})} = \left\{\sum_{i \in \tilde{I}} m(K_i) \left| \frac{1}{m(K_i)} \int_{K_i} v\, d\mathbf{x} \right|^2 \right\}^{1/2} \quad \forall v \in L^1(\Omega)\,. \tag{4.7}$$

We note that $|\,.\,|_{l_2(\Omega)}$ is a seminorm on $L^2(\Omega)$. It was first introduced by Süli [11], who considered the error $|u-U|_{l_2(\Omega)}$. We shall instead consider $|\nabla . I^h(\mathbf{a}(u-U))|_{l_2(\Omega)}$, which turns out to be a stronger and more natural seminorm for the analysis of $u-U$ (see Theorem 4.2 and Corollary 4.2 below). Note however that both $|\,.\,|_{l_2(\Omega)}$ and $|\nabla . I^h(\mathbf{a}(\,.\,))|_{l_2(\Omega)}$ are incapable of detecting chequerboard oscillations, in the computed solution U in the first case and in the computed flux $\mathbf{a}U$ in the second case.

Let $u^I \in \mathscr{U}^h$ be the interpolant to u from $\mathscr{U}^h$. We begin with a projection result for U.

THEOREM 4.1 : *Let* $\tilde{\Omega} = \bigcup_{i \in \tilde{I}} K_i$ *be arbitrary. Then*

$$|\nabla . I^h(\mathbf{a}(u^I - U))|_{l_2(\tilde{\Omega})} = |\nabla . I^h(\mathbf{a}(u-U))|_{l_2(\tilde{\Omega})} = |\nabla . (I^h(\mathbf{a}u) - \mathbf{a}u)|_{l_2(\tilde{\Omega})} . \tag{4.8}$$

Proof : Clearly $I^h(\mathbf{a}u^I) = I^h(\mathbf{a}u)$, so the first equality of the theorem holds. Next, for each $p \in \mathscr{M}^h$, by (4.6) we have

$$B(u-U, p) = (\nabla . (I^h(\mathbf{a}u) - \mathbf{a}u), p) = 0 .$$

Fix $i \in \tilde{I}$. Take p to be the characteristic function of K_i. This yields

$$\int_{K_i} \nabla . I^h(\mathbf{a}(u-U))\, d\mathbf{x} = \int_{K_i} \nabla . (I^h(\mathbf{a}u) - \mathbf{a}u)\, d\mathbf{x} . \tag{4.9}$$

Since $i \in \tilde{I}$ is arbitrary, the second equality of the theorem now follows from (4.9) and the definition of $|\,.\,|_{l_2(\Omega)}$. □

Theorem 4.1 expresses a local projection of the error $u-U$ in terms of a local truncation error. This can be made more specific in certain cases.

If the quadrilaterals K_i are sufficiently regular (in a precise sense due to Süli [11]), we can quantify the order of convergence, as follows. Let h_i denote the diameter of K_i, and let ρ_i denote the maximum diameter of circles contained in K_i. Denote by P_i and Q_i the midpoints of the diagonals of K_i.

COROLLARY 4.1 : *Let* $\tilde{\Omega} = \bigcup_{i \in \tilde{I}} K_i$ *be arbitrary. Assume that* $\mathbf{a}$ *is constant on* $\tilde{\Omega}$. *Assume also that there exist two constants* $c_0 \geq 0$ *and* $c_1 > 0$ *such that for all* $i \in \tilde{I}$,

$$\text{dist}\,(P_i, Q_i) \leq c_0\, m(K_i) , \tag{4.10a}$$

and

$$h_\iota \leq c_1 \rho_\iota . \tag{4.10b}$$

Then

$$|\nabla . (\mathbf{a}(u^I - U))|_{l_2(\tilde{\Omega})} \leq C h^2 |u|_{H^3(\tilde{\Omega})} \tag{4.11}$$

and

$$|\nabla . (\mathbf{a}(u - U))|_{l_2(\tilde{\Omega})} \leq C h^2 |u|_{H^3(\tilde{\Omega})} , \tag{4.12}$$

where $C = C(c_0, c_1, \mathbf{a})$.

Proof : From the proof of Theorem 4 of Süli [11], we obtain

$$|\nabla . (\mathbf{a}(u^I - u))|_{l_2(\tilde{\Omega})} \leq C h^2 |u|_{H^3(\tilde{\Omega})} \tag{4.13}$$

under the given hypotheses (4.10) on the K_ι. Since $\mathbf{a}$ is constant, (4.11) and (4.12) follow immediately from Theorem 4.1. □

Remark : If K_ι is a rectangle with edges parallel to the coordinate axes, then it is easy to see that

$$\frac{1}{m(K_\iota)} \int_{K_\iota} \nabla . I^h z \, d\mathbf{x} = (\nabla . I^h z)(q_\iota) \quad \forall z \in C(\bar{K}_\iota) , \tag{4.14}$$

where q_ι denotes the centroid of K_ι. Thus on tensor product meshes, $|\nabla . I^h(\mathbf{a}(u - U))|_{l_2(\tilde{\Omega})}$ is the discrete L^2 seminorm of the cell *centre* divergence error, viz.,

$$|\nabla . I^h(\mathbf{a}(u - U))|_{l_2(\tilde{\Omega})} = \left\{ \sum_{\iota \in \tilde{I}} m(K_\iota)(\nabla . I^h(\mathbf{a}(u - U))(q_\iota))^2 \right\}^{1/2} . \tag{4.15}$$

While $|\nabla . I^h(\mathbf{a} .)|_{l_2(\tilde{\Omega})}$ is generally only a seminorm on $L^2(\tilde{\Omega})$, it may be a norm on a smaller class of functions ; in particular, we need to ask under what circumstances it will be a norm on $\mathscr{U}_-^h|_{\tilde{\Omega}}$? That is, when will the proposition

« $V \in \mathscr{U}_-^h$ and $|\nabla . I^h(\mathbf{a}V)|_{l_2(\tilde{\Omega})} = 0$ together imply $V|_{\tilde{\Omega}} = 0$ » (4.16)

be true ? What is needed is the ability to show that cell-by-cell $V = 0$, starting from cells adjoining $\partial_- \Omega$ and eventually encompassing all of $\tilde{\Omega}$.

For simplicity we shall consider only the case where $\Omega = (0, 1)^2$,

$$a_i(\,.\,) > 0 \quad \text{on} \quad \bar{\Omega} \quad \text{for} \quad i = 1, 2\,, \tag{4.17}$$

and we have a tensor product mesh on Ω. Suppose that $V \in \mathscr{U}_-^h$ and $|\nabla \,.\, I^h(\mathbf{a}V)|_{l_2(\tilde{\Omega})} = 0$, where $\tilde{\Omega} = \bigcup_{i \in \tilde{I}} K_i$. By definition of $|\,.\,|_{l_2(\tilde{\Omega})}$, we consequently have

$$\int_{K_i} \nabla \,.\, I^h(\mathbf{a}V)\, d\mathbf{x} = 0 \quad \forall i \in \tilde{I}\,. \tag{4.18}$$

Evaluating $\int_{K_i} \nabla \,.\, I^h(\mathbf{a}V)\, d\mathbf{x}$ in terms of the nodal values of $\mathbf{a}$ and V, one sees easily that, if V is zero at the northwest, southwest and southeast corners of K_i, then (4.18) forces V to be zero at the northeast corner also. (This observation relies on the property (4.17).)

Thus suppose that $\tilde{\Omega}$ has the following property :

$$\partial_- K_i \subseteq \partial_- \Omega \cup \left(\bigcup_{i \in \tilde{I}} \partial_+ K_i \right) \quad \forall i \in \tilde{I}\,, \tag{4.19}$$

where

$$\begin{aligned} \partial_- K_i &= \{\mathbf{x} \in \partial K_i : \mathbf{a}(\mathbf{x}) \,.\, \mathbf{n}(\mathbf{x}) < 0\}\,, \\ \partial_+ K_i &= \{\mathbf{x} \in \partial K_i : \mathbf{a}(\mathbf{x}) \,.\, \mathbf{n}(\mathbf{x}) \geqslant 0\}\,, \end{aligned} \tag{4.20}$$

and $\mathbf{n}(\mathbf{x})$ denotes the unit outward normal to ∂K_i at $\mathbf{x} \in \partial K_i$. (In particular, if $\tilde{\Omega} = \Omega$, then $\tilde{\Omega}$ has property (4.19).) Then it is clear that $\tilde{\Omega}$ must contain the unique cell in K which has $(0, 0)$ as its southwest corner. Furthermore, using the observation of the previous paragraph one can then step cell by cell from left to right and bottom to top, to conclude that $V = 0$ on all of $\tilde{\Omega}$.

In more general situations (e.g., mesh not a tensor product, $\mathbf{a}$ not satisfying (4.17), etc.) one attempts to mimic this argument to conclude that $V = 0$ on $\tilde{\Omega}$. One still needs property (4.19), but the mesh geometry relative to the direction of $\mathbf{a}$ must also be taken into account.

We have deferred up to now the questions of existence and uniqueness of the cell vertex solution U. Assuming equality between the number of equations and unknowns, we show that existence and uniqueness of U depend precisely on whether $|\nabla \,.\, I^h(\mathbf{a}\,.\,)|_{l_2(\tilde{\Omega})}$ is a norm.

THEOREM 4.2 : *Let* $\tilde{\Omega} = \bigcup_{i \in \tilde{I}} K_i$ *be arbitrary. Set*

$$\mathscr{U}_-^h(\tilde{\Omega}) = \left\{ v \in H_-^1(\Omega) : v = \hat{v} \circ F_i^{-1},\ \hat{v} \in Q_1(\hat{K}),\ i \in \tilde{I} \right\}\,. \tag{4.21}$$

Write $(\,.\,,\,.\,)_{\tilde{\Omega}}$ *for the* $L^2(\tilde{\Omega})$ *inner product and consider the linear system of equations*

$$(\nabla \,.\, I^h(\mathbf{a}U),\, p)_{\tilde{\Omega}} = (f,\, p)_{\tilde{\Omega}} \quad \forall p \in \mathcal{M}^h \tag{4.22}$$

in the unknown function $U \in \mathcal{U}_-^h(\tilde{\Omega})$. *Assume that the number of nodes in* $\bar{\tilde{\Omega}} \backslash \partial_- \Omega$ *equals the cardinality of* $\tilde{I}$. *Then existence and uniqueness of* $U \in \mathcal{U}_-^h(\tilde{\Omega})$ *satisfying* (4.22) *are guaranteed if and only if* $|\nabla \,.\, I^h(\mathbf{a}\,.\,)|_{l_2(\tilde{\Omega})}$ *is a norm on* $\mathcal{U}_-^h(\tilde{\Omega})$.

Proof : First, observe that the nodes in $\bar{\tilde{\Omega}} \backslash \partial_- \Omega$ are precisely the nodes at which the value of U is not known a priori, while the cardinality of $\tilde{I}$ equals the number of linearly independent equations in (4.22). Hence our hypotheses guarantee that (4.22) can be expressed as a linear system of equations with the number of equations equal to the number of unknowns. Thus existence of U is guaranteed if and only if uniqueness of U is guaranteed by (4.22).

To examine uniqueness of U, suppose that

$$(\nabla \,.\, I^h(\mathbf{a}U),\, p)_{\tilde{\Omega}} = 0 \quad \forall p \in \mathcal{M}^h. \tag{4.23}$$

Take $p = P^h(\nabla \,.\, I^h(\mathbf{a}U))$ in (4.23). This yields

$$0 = \sum_{i \in \tilde{I}} \left(\int_{K_i} \nabla \,.\, I^h(\mathbf{a}U)\, d\mathbf{x} \right) \left(\frac{1}{m(K_i)} \int_{K_i} \nabla \,.\, I^h(\mathbf{a}U)\, d\mathbf{x} \right) = \\ = |\nabla \,.\, I^h(\mathbf{a}U)|^2_{l_2(\tilde{\Omega})}. \tag{4.24}$$

Thus $U = 0$ if and only if $|\nabla \,.\, I^h(\mathbf{a}\,.\,)|_{C_2(\tilde{\Omega})}$ is a norm on $\mathcal{U}_-^h(\tilde{\Omega})$. □

Remark : In the next section we shall discuss a particular case where the equality in the numbers of equations and unknowns assumed in Theorem 4.2 does not hold.

We now examine the relationship between Theorem 4.1 and the error estimates obtained previously by Morton and Süli [9], by further characterising $|\nabla \,.\, I^h(\mathbf{a}\,.\,)|_{l_2(\Omega)}$. Defining $\partial_{\Omega +} K_i = \partial_+ \Omega \cap \bar{K}_i$, set

$$|v|_{l_2(\partial_+ \Omega)} = \\ = \left\{ \sum_{K_i \in K,\, \partial_{\Omega +} K_i \neq \emptyset} m(\partial_{\Omega +} K_i) \left| \frac{1}{m(\partial_{\Omega +} K_i)} \int_{\partial_{\Omega +} K_i} v\, ds \right|^2 \right\}^{1/2}. \tag{4.25}$$

In [9], Morton and Süli consider the problem

$$\nabla \cdot (\mathbf{a}w) + bw = f \quad \text{in} \quad \Omega \tag{4.26a}$$

$$w = 0 \quad \text{on} \quad \partial_- \Omega\,, \tag{4.26b}$$

where $\Omega = (0, 1)^2$, $\mathbf{a} = (a_1, a_2)$ is constant with $a_1 > 0$ and $a_2 > 0$ on $\bar{\Omega}$, $b \in C(\bar{\Omega})$ with $b > 0$ on $\bar{\Omega}$, and $f \in L^2(\Omega)$. From the condition $b > 0$, a simplified version of the discrete Gårding inequality given in Süli [11] is used to prove stability of the cell vertex method, and show that on a tensor product mesh

$$|w - W|_{l_2(\Omega)} + |w - W|_{l_2(\partial_+ \Omega)} \leq Ch^2 |w|_{H^3(\Omega)}\,, \tag{4.27}$$

where W is the computed solution to (4.26b) and C is a generic constant. We use a mapping $M : S^h \to T^h$ to obtain the result for $b = 0$.

THEOREM 4.3 : *Assume that* $\Omega = (0, 1)^2$, *that we have a tensor product mesh, and that* $\mathbf{a}$ *is constant with* $a_1 > 0$ *and* $a_2 > 0$ *on* $\bar{\Omega}$. *Then* (4.5) *has a unique solution* $U \in \mathcal{U}_-^h$ *and*

(i) $$|U|_{l_2(\Omega)} + |U|_{l_2(\partial_+ \Omega)} \leq C\,|f|_{l^2(\Omega)} \tag{4.28}$$

(ii) $$|u - U|_{l_2(\Omega)} + |u - U|_{l_2(\partial_+ \Omega)} \leq Ch^2 |u|_{H^3(\Omega)}\,. \tag{4.29}$$

Proof : These are the same results as Theorems 3 and 4 of Morton and Süli [9], except that here $b \equiv 0$. To circumvent the requirement that $b > 0$ necessitates only a certain change in the argument which led to Theorem 2 of [9], giving instead the lemma proved below. □

LEMMA 4.4 : *Under the hypotheses of Theorem 4.3, there is a mapping* $M : \mathcal{U}_-^h \to \mathcal{M}^h$ *such that for* $\forall U \in \mathcal{U}_-^h$

$$B(U, MU) \geq \frac{1}{2} \max\,(a_1, a_2)\, \sigma\, e^{-\sigma} |U|^2_{l_2(\Omega)} + \\ + \frac{1}{2} e^{-\sigma} \min\,(a_1, a_2) |U|^2_{l_2(\partial_+ \Omega)} \tag{4.30}$$

for any $\sigma > 0$.

Proof : For $\mathbf{a}$ constant and on a tensor product mesh, and using the averaging operators $\mu_1 U_{ij} := \frac{1}{2}(U_{i-1,j} + U_{ij})$, $\mu_2 U_{ij} := \frac{1}{2}(U_{i,j-1} + U_{ij})$ and backward difference operators $\Delta_{-1} U_{ij} := U_{ij} - U_{i-1,j}$, $\Delta_{-2} U_{ij} := U_{ij} - U_{i,j-1}$ and with $\mathbf{x} = (x^1, x^2)$, we have

$$B(U, V) = \int_\Omega V \nabla \cdot (\mathbf{a}U)\, d\mathbf{x} \quad \forall U \in \mathcal{U}_-^h\,, \quad \forall V \in \mathcal{M}^h$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{N} V_{ij} [a_1 (x_j^2 - x_{j-1}^2)\, \Delta_{-1}\, \mu_2\, U_{ij} + a_2 (x_i^1 - x_{i-1}^1)\, \Delta_{-2}\, \mu_1\, U_{ij}]$$

$$= S_1 + S_2\,, \quad \text{say}\,. \tag{4.31}$$

Suppose now that $V(=MU) \equiv m_{ij}\,\mu_1\,\mu_2\,U_{ij}$ on $(x^1_{i-1}, x^1_i) \times (x^2_{j-1}, x^2_j)$. Then by summation by parts we obtain

$$S_1 = \sum_{j=1}^{N} a_1 (x^2_j - x^2_{j-1}) \sum_{i=1}^{M} \frac{1}{2} m_{ij} [(\mu_2 U_{ij})^2 - (\mu_2 U_{i-1,j})^2]$$

$$= \frac{1}{2} a_1 \sum_{j=1}^{N} (x^2_j - x^2_{j-1}) \times \left\{ m_{Mj} (\mu_2 U_{Mj})^2 - \sum_{i}^{M-1} (\mu_2 U_{ij})^2 (m_{i+1,j} - m_{ij}) \right\} \tag{4.32a}$$

$$S_2 = \frac{1}{2} a_2 \sum_{i=1}^{M} (x^1_i - x^1_{i-1}) \times \left\{ m_{iN} (\mu_1 U_{iN})^2 - \sum_{j=1}^{N-1} (\mu_1 U_{ij})^2 (m_{i,j+1} - m_{ij}) \right\}. \tag{4.32b}$$

Also we have

$$|U|^2_{l_2(\Omega)} \equiv \sum_{i=1}^{M} \sum_{j=1}^{N} (x^1_i - x^1_{i-1})(x^2_j - x^2_{j-1})(\mu_1\,\mu_2\,U_{ij})^2\,, \tag{4.33}$$

for which

$$(\mu_1\,\mu_2\,U_{ij})^2 \leqslant \frac{1}{2} [(\mu_1 U_{ij})^2 + (\mu_1 U_{i,j-1})^2]$$

$$\text{or} \quad \frac{1}{2} [(\mu_2 U_{ij})^2 + (\mu_2 U_{i-1,j})^2]\,, \tag{4.34}$$

and

$$|U|^2_{l_2(\partial_+\,\Omega)} \equiv \sum_{i=1}^{M} (x^1_i - x^1_{i-1})(\mu_1 U_{iN})^2 +$$

$$+ \sum_{j=1}^{N} (x^2_j - x^2_{j-1})(\mu_2 U_{Mj})^2\,. \tag{4.35}$$

Of the many possible choices for m_{ij}, let us suppose $a_1 \geqslant a_2$ and set $m_{ij} = e^{-\frac{1}{2}\sigma(x^1_{i-1} + x^1_i)}$ Then it is clear that

$$m_{ij} - m_{i+1,j} = \frac{1}{2} \sigma (x^1_{i+1} - x^1_{i-1})\, e^{-\sigma\xi}\,, \quad x^1_{i-1} < \xi < x^1_{i+1}$$

$$\geqslant \frac{1}{2} \sigma\, e^{-\sigma} (x^1_{i+1} - x^1_{i-1})\,, \tag{4.36}$$

and

$$m_{iN} \geqslant m_{Mj} = e^{-\sigma} \cdot e^{\frac{1}{2}\sigma(x^1_M - x^1_{M-1})} \geqslant e^{-\sigma} \left[1 + \frac{1}{2} \sigma (x^1_M - x^1_{M-1}) \right]. \tag{4.37}$$

Applying these bounds in (4.32), splitting the terms $x^1_{i+1} - x^1_{i-1} = (x^1_{i+1} - x^1_i) + (x^1_i - x^1_{i-1})$ and regrouping the sums gives

$$S_1 \geqslant \frac{1}{2} a_1 e^{-\sigma} \sum_{j=1}^{N} (x^2_j - x^2_{j-1})(\mu_2 U_{Mj})^2 + \frac{1}{2} a_1 \sigma e^{-\sigma} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{1}{2} (x^1_i - x^1_{i-1})(x^2_j - x^2_{j-1}) \times \times [(\mu_2 U_{ij})^2 + (\mu_2 U_{i-1,j})^2] \quad (4.38a)$$

$$S_2 \geqslant \frac{1}{2} a_2 e^{-\sigma} \sum_{i=1}^{M} (x^1_i - x^1_{i-1})(\mu_1 U_{iN})^2 , \quad (4.38b)$$

and hence

$$S_1 + S_2 \geqslant \frac{1}{2} a_1 \sigma e^{-\sigma} |U|^2_{l_2(\Omega)} + \frac{1}{2} a_2 e^{-\sigma} |U|^2_{l_2(\partial_+ \Omega)} .$$

The required result follows for general **a**. □

We can deduce an inequality of Poincaré-Friedrichs type for $|\cdot|_{l_2(\Omega)}$ from the above results.

COROLLARY 4.2 : *Under the hypotheses of Theorem* 4.3, *there exists a constant C such that*

$$|U|_{l_2(\Omega)} \leqslant C |\nabla \cdot (\mathbf{a}U)|_{l_2(\Omega)} \quad \forall U \in \mathscr{U}^h_- . \quad (4.39)$$

Proof : Let $f = \nabla \cdot (\mathbf{a}U)$ in (4.2) and apply the result (i) of Theorem 4.3. □

Remark : Corollary 4.2 shows that (4.11) is stronger than the inequality

$$|u^I - U|_{l_2(\Omega)} \leqslant C h^2 |u|_{H^3(\Omega)} \quad (4.40)$$

obtained by Morton and Süli [9].

Remark : Consider the nonlinear problem

$$\nabla \cdot \mathbf{F}(u) = f \quad \text{in} \quad \Omega \quad (4.41)$$

with suitable boundary conditions, where $\mathbf{F} : R \to R^2$ is a smooth function and $f \in L^2(\Omega)$. The cell vertex solution U satisfies

$$(\nabla \cdot I^h \mathbf{F}(U), p)_{\tilde{\Omega}} = (f, p)_{\tilde{\Omega}} \quad \forall p \in \mathscr{M}^h , \quad (4.42)$$

where $\tilde{\Omega} = \bigcup_{i \in \tilde{I}} K_i$ is some appropriate subset of Ω. On inspection it is clear

that the proof of Theorem 4.1 goes through as before, leading to the results

$$\left|\nabla . I^h(\mathbf{F}(u^I) - \mathbf{F}(U))\right|_{l_2(\tilde{\Omega})} = \left|\nabla . I^h(\mathbf{F}(u^I) - \mathbf{F}(u))\right|_{l_2(\tilde{\Omega})} \quad (4.43)$$

and

$$\left|\nabla . I^h(\mathbf{F}(u) - \mathbf{F}(U))\right|_{l_2(\tilde{\Omega})} = \left|\nabla . (I^h \mathbf{F}(u) - \mathbf{F}(u))\right|_{l_2(\tilde{\Omega})} . \quad (4.44)$$

Remark : The arguments presented in this section do not rely in any intrinsic way on the two-dimensional nature of the problem. Analogous results will hold for analogous n-dimensional problems with $n \neq 2$; one merely needs to alter the concepts of quadrilateral and isoparametric bilinear interpolant in the appropriate way.

5. CONVECTION IN TWO DIMENSIONS WITH CHARACTERISTIC BOUNDARIES

We now turn our attention to a particular situation in which the number of equations provided by the basic cell vertex method is a priori less than the number of unknowns. The requisite extra equations may, for example, be obtained by a « cell-splitting » approach suggested by Morton [6]. We analyse this problem and show that, if the method used to generate the extra equations has a certain property, then this will ensure optimal order of convergence of the computed nodal values. The cell-splitting method is shown to possess this property.

Let $\Omega = (0, 1)^2 \subset R^2$. Let $\mathbf{a} = (0, a_2) : \Omega \to R^2$ be a given smooth function with $a_2 > 0$ on $\bar{\Omega}$. Then, in the notation of section 4,

$$\partial_- \Omega = \{(x^1, 0) : 0 < x^1 < 1\} . \quad (5.1)$$

Consider the boundary value problem

$$\nabla . (\mathbf{a}u) = f \quad \text{on} \quad \Omega , \quad (5.2a)$$

$$u = 0 \quad \text{on} \quad \partial_- \Omega , \quad (5.2b)$$

where for simplicity we assume that $f \in C^3(\bar{\Omega})$.

We assume that we have a uniform tensor product mesh on Ω. Suppose that M and N are positive integers with $x_i^1 = i/M$ for $i = 0, \ldots, M$ and $x_j^2 = j/N$ for $j = 0, \ldots, N$, and set

$$K_{ij} = (x_i^1, x_{i+1}^1) \times (x_j^2, x_{j+1}^2) , \quad \text{for} \quad 0 \leq i \leq M - 1 , \quad 0 \leq j \leq N - 1 , \quad (5.3)$$

with $\Delta x = 1/M$, $\Delta y = 1/N$, $h_{ij} = \text{diameter}\,(K_{ij})$, $h = \max_{i,j}\{h_{ij}\}$. Let $H_-^1(\Omega)$, $\mathscr{U}_-^h$, $\mathscr{M}^h$ and I^h be defined analogously to section 4. Note that each

function $v \in \mathscr{U}_-^h$ is now piecewise bilinear since each quadrilateral K_{ij} is a rectangle.

We require our computed solution $U \in \mathscr{U}_-^h$ to satisfy

$$(\nabla \cdot I^h(\mathbf{a}U), p) = (\tilde{f}, p) \quad \forall p \in \mathscr{M}^h, \tag{5.4}$$

where $(.,.)$ is the $L^2(\Omega)$ inner product and $\tilde{f}$ is the bilinear interpolant to f on the mesh. We shall write U_i^j for $U(x_i^1, x_j^2)$, for all i and j. However, (5.4) by itself does not determine U uniquely. For suppose that we have computed U_i^j for $i = 0, \ldots, M$ and $j = 0, \ldots, n$ where $n \geq 0$ is fixed. We wish, as in the continuous problem (5.2), to proceed in the direction of the positive x^2-axis and now compute U_i^{n+1} for $i = 0, \ldots, M$. But to compute these $M + 1$ unknowns, (5.4) provides only M linearly independent equations (obtained by taking p to be the characteristic function of K_{in} for $i = 0, \ldots, M - 1$) which relate the U_i^{n+1} to the previously computed values of U. A further equation is needed here.

One resolution of this difficulty is the cell-splitting idea of Morton [6]. In the present case, this approach divides the cell K_{0n} into two halves by the line $x^1 = x_{1/2}^1 := (x_0^1 + x_1^1)/2$, applies the cell vertex method on each half, and finally requires that the values of $a_2 U$ at $(x_{1/2}^1, x_j^2)$ be linear interpolants of the values of $a_2 U$ at (x_0^1, x_j^2) and (x_1^1, x_j^2) for $j = n$ and $j = n + 1$ respectively. Written out explicitly, the above cell-splitting equations are

$$\frac{\Delta x}{4}\,[V_0^{n+1} + V_{1/2}^{n+1} - V_0^n - V_{1/2}^n] = \int_{x^2 = x_n^2}^{x_{n+1}^2} \int_{x^1 = x_0^1}^{x_{1/2}^1} \tilde{f}\, dx^1\, dx^2\,, \tag{5.5a}$$

$$\frac{\Delta x}{4}\,[V_{1/2}^{n+1} + V_1^{n+1} - V_{1/2}^n - V_1^n] = \int_{x^2 = x_n^2}^{x_{n+1}^2} \int_{x^1 = x_{1/2}^1}^{x_1^1} \tilde{f}\, dx^1\, dx^2\,, \tag{5.5b}$$

$$2\, V_{1/2}^n = V_0^n + V_1^n\,, \tag{5.5c}$$

$$2\, V_{1/2}^{n+1} = V_0^{n+1} + V_1^{n+1}\,, \tag{5.5d}$$

where for notational convenience we have set $V_i^j := (a_2 U)(x_i^1, x_j^2)$ for all i and j.

The equations (5.5) can easily be solved, yielding

$$\begin{aligned} V_0^{n+1} &= V_0^n + \frac{1}{\Delta x}\left[3 \int_{x^2 = x_n^2}^{x_{n+1}^2} \int_{x^1 = x_0^1}^{x_{1/2}^1} \tilde{f}\, dx^1\, dx^2 - \int_{x^2 = x_n^2}^{x_{n+1}^2} \int_{x^1 = x_{1/2}^1}^{x_1^1} \tilde{f}\, dx^1\, dx^2\right] \\ &= V_0^n + \int_{t = x_n^2}^{x_{n+1}^2} \tilde{f}(x_0^1, t)\, dt\,, \end{aligned} \tag{5.6a}$$

$$V_1^{n+1} = V_1^n + \frac{1}{\Delta x}\left[- \int_{x^2 = x_n^2}^{x_{n+1}^2} \int_{x^1 = x_0^1}^{x_{1/2}^1} \tilde{f}\, dx^1\, dx^2 + 3 \int_{x^2 = x_n^2}^{x_{n+1}^2} \int_{x^1 = x_{1/2}^1}^{x_1^1} \tilde{f}\, dx^1\, dx^2 \right]$$

$$= V_1^n + \int_{t = x_n^2}^{x_{n+1}^2} \tilde{f}(x_1^1, t)\, dt\,, \tag{5.6b}$$

where we have used the bilinearity of $\tilde{f}$ to simplify the expressions.

On the other hand, integrating (5.2*a*), we obtain

$$(a_2\, u)(x_0^1, x_{n+1}^2) = (a_2\, u)(x_0^1, x_n^2) + \int_{t = x_n^2}^{x_{n+1}^2} f(x_0^1, t)\, dt$$

$$= (a_2\, u)(x_0^1, x_n^2) + \int_{t = x_n^2}^{x_{n+1}^2} \tilde{f}(x_0^1, t)\, dt + O((\Delta y)^3)\,. \tag{5.7}$$

Subtracting (5.6*a*) from (5.7), we have

$$e_0^{n+1} = e_0^n + O((\Delta y)^3)\,, \tag{5.8}$$

where we set the nodal error

$$e_i^j = (a_2\, u)(x_i^1, x_j^2) - V_i^j \quad \forall i, j\,. \tag{5.9}$$

Similarly

$$e_1^{n+1} = e_1^n + O((\Delta y)^3)\,. \tag{5.10}$$

Clearly (5.8) and (5.10), together with $e_0^0 = e_1^0 = 0$, imply that

$$\left|e_j^n\right| \leq C\,(\Delta y)^2 \quad \text{for} \quad j = 0,\ 1 \quad \text{and} \quad n = 0, \ldots, N\,, \tag{5.11}$$

where C is a generic constant which is independent of the mesh.

We can now return to the issue raised earlier, namely how to compute U_i^{n+1} (or equivalently V_i^{n+1}) for $i = 0, \ldots, M$ from the U_i^n. Cell-splitting, as described above, yields V_0^{n+1} and V_1^{n+1}. Now taking p in (5.4) to be the characteristic function of K_{in} yields

$$V_i^{n+1} + V_{i+1}^{n+1} - V_i^n - V_{i+1}^n = \frac{2}{\Delta x} \int_{K_{in}} \tilde{f}\, dx^1\, dx^2\,,$$

$$\text{for} \quad i = 1, \ldots, M-1\,. \tag{5.12}$$

We can compute in order V_2^{n+1}, V_3^{n+1}, ..., V_M^{n+1} from (5.12).

Since $\tilde{f}$ is bilinear on K_{in},

$$\frac{2}{\Delta x}\int_{K_{in}} \tilde{f}\, dx^1\, dx^2 = \frac{\Delta y}{2}\,\big|f(x_i^1, x_n^2) + f(x_{i+1}^1, x_n^2) + \\ + f(x_i^1, x_{n+1}^2) + f(x_{i+1}^1, x_{n+1}^2)\big]\,. \quad (5.13)$$

We also have

$$\begin{aligned}
&(a_2 u)(x_i^1, x_{n+1}^2) + \\
&\quad + (a_2 u)(x_{i+1}^1, x_{n+1}^2) - (a_2 u)(x_i^1, x_n^2) - (a_2 u)(x_{i+1}^1, x_n^2) \\
&\quad = \int_{x_n^2}^{x_{n+1}^2} [(a_2 u)_y (x_i^1, t) + (a_2 u)_y (x_{i+1}^1, t)]\, dt \\
&\quad = \int_{x_n^2}^{x_{n+1}^2} [f(x_i^1, t) + f(x_{i+1}^1, t)]\, dt \\
&\quad = \frac{\Delta y}{2}\,[f(x_i^1, x_n^2) + f(x_{i+1}^1, x_n^2) + f(x_i^1, x_{n+1}^2) + f(x_{i+1}^1, x_{n+1}^2)] \\
&\quad + \int_{x_n^2}^{x_{n+1}^2} [f(x_i^1, t) - \tilde{f}(x_i^1, t) + f(x_{i+1}^1, t) - \tilde{f}(x_{i+1}^1, t)]\, dt\,. \qquad (5.14)
\end{aligned}$$

Combining (5.12), (5.13) and (5.14), we obtain

$$\begin{aligned}
&e_i^{n+1} + e_{i+1}^{n+1} - e_i^n - e_{i+1}^n = \\
&\quad = \int_{x_n^2}^{x_{n+1}^2} [f(x_i^1, t) - \tilde{f}(x_i^1, t) + f(x_{i+1}^1, t) - \tilde{f}(x_{i+1}^1, t)]\, dt \\
&\quad = -\frac{(\Delta y)^3}{6} f_{yy}(P_{in}) + O((\Delta x + \Delta y)^4)\,, \qquad (5.15)
\end{aligned}$$

for $i = 1, \ldots, M-1$, on using Taylor expansions, where P_{in} denotes the centroid of K_{in}. Now for $k = 2, \ldots, M$,

$$\begin{aligned}
&e_k^{n+1} - e_k^n + (-1)^k (e_1^{n+1} - e_1^n) = \\
&\quad = (-1)^{k+1} \sum_{i=1}^{k-1} (-1)^i [(e_{i+1}^{n+1} - e_{i+1}^n) + (e_i^{n+1} - e_i^n)] \\
&\quad = (-1)^{k+1} \sum_{i=1}^{k-1} (-1)^i \left[-\frac{(\Delta y)^3}{6} f_{yy}(P_{in}) + O((\Delta x + \Delta y)^4)\right] \qquad (5.16)
\end{aligned}$$

from (5.15). By combining terms in pairs, one sees that

$$\left| \sum_{i=1}^{k-1} (-1)^i f_{yy}(P_{in}) \right| \leq C \quad \text{for each} \quad k . \tag{5.17}$$

Hence (5.16) yields

$$\left| e_k^{n+1} - e_k^n \right| \leq \left| e_1^{n+1} - e_1^n \right| + C (\Delta x + \Delta y)^3 \tag{5.18}$$

for $k = 2, \ldots, M$, on assuming that $\dfrac{\Delta x}{\Delta y} \leq C$. Recalling (5.10), we deduce from (5.18) and $e_k^0 = 0$ that

$$\left| e_k^{n+1} \right| \leq C (\Delta x + \Delta y)^2 \text{ for } k = 2, \ldots, M \text{ and } n = 0, \ldots, N-1 , \tag{5.19}$$

on assuming that $\dfrac{\Delta y}{\Delta x} \leq C$. Since $a_2 > 0$, this shows that we have second order nodal convergence of U to u, which is best possible for the scheme (5.12).

Note that when proving convergence of (5.12), the only properties of the cell-splitting approach which we needed were (5.8) and (5.10). We formally state the results of this section below.

THEOREM 5.1 : *Assume that* $\dfrac{\Delta x}{\Delta y} \leq C$ *and* $\dfrac{\Delta y}{\Delta x} \leq C$. *Suppose that the cell vertex scheme* (5.12) *is used to solve* (5.2), *with* $(aU)_0^n$ *and* $(aU)_1^n$ *computed for each n by some method which yields*

$$\left| e_0^{n+1} - e_0^n \right| + \left| e_1^{n+1} - e_1^n \right| \leq C (\Delta y)^3 \quad \textit{for} \quad n = 0, \ldots, N-1 . \tag{5.20}$$

Then

$$\left| u(x_i^1, x_j^2) - U_i^j \right| \leq C (\Delta x + \Delta y)^2 \quad \forall i, j . \tag{5.21}$$

COROLLARY 5.1 : *Assume the hypotheses of Theorem* 5.1, *and suppose that cell-splitting is used to compute each* U_0^n *and* U_1^n. *Then*

$$\left| u(x_i^1, x_j^2) - U_i^j \right| \leq C (\Delta x + \Delta y)^2 \quad \forall i, j . \tag{5.22}$$

REFERENCES

[1] J. W. BARRETT and K. W. MORTON, 1984, Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems, *Computer Methods in Applied Mechanics and Engineering*, **45**, 97-122.

[2] P. W. HEMKER, 1977, *A numerical study of stiff two-point boundary problems,* PhD thesis, Mathematisch Centrum, Amsterdam.

[3] R. B. KELLOGG and A. TSAN, 1978, Analysis of some difference approximations for a singular perturbation problem without turning points, *Mathematics of Computation,* **32,** 1025-1039.

[4] J. A. MACKENZIE, 1991, *Cell vertex finite volume methods for the solution of the compressible Navier-Stokes equations,* PhD thesis, Oxford University Computing Laboratory, 11 Keble Road, Oxford, OX1 3QD.

[5] J. A. MACKENZIE and K. W. MORTON, 1992, Finite volume solutions of convection-diffusion test problems, *Mathematics of Computation,* **60**(201), 189-220.

[6] K. W. MORTON, 1991, Finite volume methods and their analysis, in J. R. Whiteman, editor, *The Mathematics of Finite Elements and Applications VII MAFELAP 1990,* Academic Press, 189-214.

[7] K. W. MORTON, 1992, Upwinded test functions for finite element and finite volume methods, in D. F. Griffiths and G. A. Watson, editors, *Numerical analysis 1991 Proceedings of the 14th Dundee Conference, June 1991,* number 260 in Pitman Research Notes in Mathematics Series, pp. 128-141, Longman Scientific and Technical.

[8] K. W. MORTON, P. I. CRUMPTON and J. A. MACKENZIE, 1993, Cell vertex methods for inviscid and viscous flows, *Computers Fluids,* **22**(2/3), 91-102.

[9] K. W. MORTON and E. SÜLI, 1991, Finite volume methods and their analysis, *IMA Journal of Numerical Analysis,* **11,** 241-260.

[10] M. STYNES and E. O'RIORDAN, 1991, An analysis of a singularly perturbed two-point boundary value problem using only finite element techniques, *Mathematics of Computation,* **56,** 663-676.

[11] E. SÜLI, 1991, The accuracy of finite volume methods on distorted partitions. In J. R. Whiteman, editor. *The Proceedings of the Conference on The Mathematics of Finite Elements and Applications VII MAFELAP,* pp. 253-260. Academic Press.

[12] E. SÜLI, 1992, The accuracy of cell vertex finite volume methods on quadrilateral meshes, *Mathematics of Computation,* **59**(200), 359-382.