JOURNAL OF COMPUTATIONAL PHYSICS 108, 59–72 (1993)

Capturing Contact Discontinuities in Steady-State Conservation Laws

JOHN E. LAVERY*

Board on Mathematical Sciences, National Research Council, Washington, DC 20418

Received July 20, 1990; revised November 12, 1992

Solution of the steady-state scalar conservation law $u_x + \tau u_y = 0$ (τ constant, $0 \leqslant \tau \leqslant 1$) on the unit square is considered. A grid of n^2 equal square cells with sides of length $h = 1/n$ is placed on the unit square. Numerical values $u_{i,j}$ are located at the cell vertices $(x_i, y_j) = (ih, jh)$. The conservation law is discretized on each cell by finite volumes with trapezoidal integration. When Dirichlet conditions are given on the bottom and left boundary segments, there results a system of n^2 equations for n^2 unknown values $u_{i,j}$. When the Dirichlet boundary conditions are discontinuous, the solution of this system of equations is an oscillatory approximation of the discontinuous continuum solution. If one adds outflow boundary conditions of linearity in the normal direction on the top and right boundary segments, the numerical system becomes overdetermined. The l_2 (least-squares) solution of this overdetermined system, which is called the "basic l_2 solution" to distinguish it from the "final l_2 solution" introduced later, has a smeared discontinuity but is less oscillatory than the solution of the nonoverdetermined system without the outflow boundary conditions. Equations for certain cells on which the numerical solution has large variation are then omitted from the system and other equations for newly created triangular cells are added to the system. The l_2 solution of this altered system, called the "final l_2 solution," is a much better approximation of the continuum solution, one in which the discontinuity is captured in a one-cell-wide path of cells. If, instead of using the l_2 method, one uses the l_1 method, that is, minimizes the sum of the absolute values of the residuals rather than the sum of squares of the residuals, there is further improvement in the numerical solution. The basic l_1 solution, that is, the l_1 solution of the system before equations for certain square cells are omitted and other equations for triangular cells are added, is nearly nonoscillatory and is less diffusive than the basic l_2 solution. The final l_1 solution, that is, the l_1 solution of the system after some equations are omitted and others are added, is also better than the final l_2 solution. The final l_1 solution is nonoscillatory and has a sharp discontinuity in a one-cell-wide path of cells that closely approximates the line on which the contact discontinuity in the continuum solution occurs. The conclusions of this paper are based on computational results for 14 values of τ between zero and one and for two different sets of discontinuous Dirichlet boundary conditions.

1. INTRODUCTION

Accurate calculation of two- and three-dimensional flows with contact discontinuities and shocks is one of the more difficult tasks in computational fluid dynamics. Two- and three-dimensional equations are often solved by shock-capturing methods with operator splitting, that is, by applying one-dimensional shock-capturing methods in each coordinate direction (see Chakravarthy [4], Harten [7], Leonard [13], Roe [18], Woodward and Collela [21]). These methods are not able to capture discontintuities in higher dimensions with the same high accuracy that can be achieved in one dimension. Even the algorithms of Davis [5], Hirsch *et al.* [8], Morton and Paisley [15], Powell and van Leer [17] and Roe [19], which are inherently higher-dimensional and not based on operator splitting, do not sharply capture discontinuities that are oblique to the grid. Adaptive gridding such as that used in Peraire *et al.* [16] alleviates but does not eliminate the problem: even though the cells near the discontinuity are smaller, the discontinuity is still smeared out over several cells. Shock-tracking methods (Glimm *et al.* [6]), which build discontinuities into the numerical solution, produce sharp discontinuities but the accuracy of the position of the discontinuity depends on the accuracy of the physics put into the jump conditions.

Computing contact discontinuities, which are noncompressive, is known to be more difficult than computing shocks, which are compressive (cf. [5, p. 80]). In the present paper, we develop and compare two shock-capturing procedures for obtaining nonoscillatory and nondiffusive numerical approximations of solutions with contact discontinuities. These procedures are based on l_2 (least-squares) minimization and l_1 minimization and do not use adaptive gridding. In previous papers, we have developed l_1 procedures for obtaining accurate shocks in solutions of one-dimensional Burgers' [9, 10] and Euler [11] equations and of two-dimensional Burgers' equations [12]. In the present paper, we will show that a variant of the l_1 procedure can produce equally accurate contact discontinuities.

We consider here only steady-state conservation laws. While most methods for solving steady-state conservation laws are time-dependent methods that yield convergence to the steady-state solution as $t \to \infty$, we will see in this paper that it is both natural and advantageous to solve steady-state equations by steady-state methods such as the l_1 and

* Original manuscript submitted when at Mathematical Sciences Division, Office of Naval Research, Arlington, Virginia.

0021-9991/93 $5.00

l_2 procedures. Only by using steady-state methods can the largest source of error in the numerical system, namely, the equations on cells that contain discontinuities, be eliminated from the system.

The continuum and discrete forms of the conservation law to be discussed in this paper are presented in the next section.

2. CONSERVATION LAW AND DISCRETIZATION

Consider the steady-state linear conservation law

$$u_x + \tau u_y = 0 \qquad \text{on} \quad D := (0, 1) \times (0, 1) \tag{2.1a}$$

(τ a constant, $0 \leqslant \tau \leqslant 1$) with the Dirichlet conditions

$$\begin{aligned} u(x, 0) &= g_0(x), \qquad 0 \leqslant x \leqslant 1, \\ u(0, y) &= g_1(y), \qquad 0 \leqslant y \leqslant 1, \end{aligned} \tag{2.1b}$$

on the bottom and left boundaries. Equation (2.1a) represents convection of a passive scalar u in a uniform flow along straight lines inclined at an angle of $\theta = \arctan(\tau)$ with respect to the x-axis. In what follows, g_0 is smooth, g_1 is smooth at all points except $y = h/2$ (h will be a mesh length, see (2.4a) below), where g_1 has a jump discontinuity, and $g_0(0) = g_1(0)$. The solution u of (2.1) then has a contact discontinuity along the line

$$y = \tau x + h/2. \tag{2.2}$$

This solution is the pointwise limit as $\varepsilon \to 0$ of the solution of the singularly perturbed equation

$$-\varepsilon(u_{xx} + u_{yy}) + u_x + \tau u_y = 0 = 0 \tag{2.3}$$

with boundary conditions (2.1b) and additional boundary conditions on the top and right boundary segments.

Place on D a uniform mesh of square cells $C_{i,j} := (x_{i-1}, x_i) \times (y_{j-1}, y_j)$ with n cells in the x-direction and n cells in the y-direction:

$$h := 1/n \tag{2.4a}$$

$$x_i := ih, \; 0 \leqslant i \leqslant n, \qquad y_j := jh, \; 0 \leqslant j \leqslant n. \tag{2.4b}$$

We seek numerical approximations $u_{i,j}$ of the values $u(x_i, y_j)$ of the continuum solution of (2.1) at the cell vertices.

The finite-volume discretization of (2.1a) is generated by integrating (2.1a) over each cell:

$$\begin{aligned} &-\tau \int_{x_{i-1}}^{x_i} u(x, y_{j-1})\, dx - \int_{y_{j-1}}^{y_j} u(x_{i-1}, y)\, dy \\ &\qquad + \tau \int_{x_{i-1}}^{x_i} u(x, y_j)\, dx + \int_{y_{j-1}}^{y_j} u(x_i, y)\, dy \\ &\qquad = \iint_{C_{i,j}} (u_x + \tau u_y)\, dA = 0. \end{aligned} \tag{2.5}$$

Discretizing each of the boundary integrals by the trapezoidal rule and multiplying by $2/h$, one obtains the "box" scheme

$$r_{i,j} := \left\{ \begin{array}{l} (-1+\tau)\, u_{i-1,j} + (1+\tau)\, u_{i,j} \\ +(-1-\tau)\, u_{i-1,j-1} + (1-\tau)\, u_{i,j-1} \end{array} \right\} = 0. \tag{2.6}$$

The good performance of the box scheme (2.6) when the solution u is smooth and the poor performance of the same scheme when the solution is discontinuous are well known and are recalled for the convenience of the reader in the next section. In that section and throughout the remainder of this paper, an example with

$$n = 14, \tag{2.7a}$$

$$\tau = \tan 35°, \tag{2.7b}$$

$$g_0(x) = 0, \qquad 0 \leqslant x \leqslant 1, \tag{2.7c}$$

$$g_1(y) = 0, \qquad 0 \leqslant y \leqslant h/2, \tag{2.7d}$$

$$g_1(y) = \cos(\pi y), \qquad h/2 < y \leqslant 1, \tag{2.7e}$$

is used to illustrate. In this example, the solution is

$$\begin{aligned} u(x, y) &= g_1(y - \tau x) = \cos(\pi(y - \tau x)) \\ &\qquad \text{to the left of the line} \quad y = \tau x + h/2, \end{aligned} \tag{2.8a}$$

$$\begin{aligned} u(x, y) &= 0 \qquad \text{on and to the right of the line} \\ &\qquad y = \tau x + h/2. \end{aligned} \tag{2.8b}$$

This solution u of (2.1), (2.7) is plotted in Fig. 1a at the mesh points (x_i, y_j), $0 \leqslant i \leqslant n$, $0 \leqslant j \leqslant n$. The straight-line segments in Fig. 1a are lines connecting adjacent values of $u(x_i, y_j)$. In this and all other perspective plots in this paper,

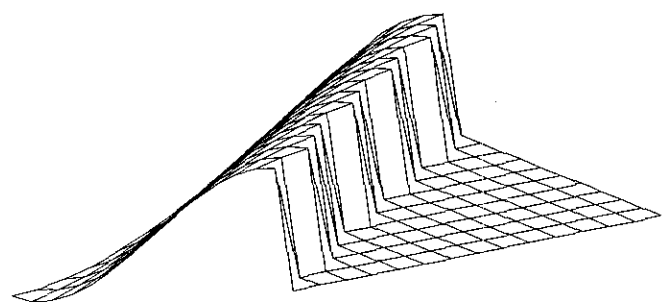

FIG. 1a. Continuum solution at node points.

-1.0000	-0.9877	-0.9510	-0.8909	-0.8089	-0.7069	-0.5876	-0.4537	-0.3087	-0.1560	0.0005	0.1569	0.3095	0.4545	0.5883
-0.9749	-0.9281	-0.8584	-0.7676	-0.6578	-0.5318	-0.3928	-0.2440	-0.0893	0.0677	0.2230	0.3728	0.5134	0.6413	0.7535
-0.9010	-0.8220	-0.7227	-0.6057	-0.4737	-0.3301	-0.1783	-0.0221	0.1346	0.2880	0.4343	0.5699	0.6915	0.7960	0.8809
-0.7818	-0.6746	-0.5508	-0.4134	-0.2659	-0.1117	0.0451	0.2009	0.3517	0.4939	0.6239	0.7385	0.8349	0.9107	0.9641
-0.6235	-0.4935	-0.3513	-0.2005	-0.0447	0.1122	0.2663	0.4139	0.5512	0.6750	0.7821	0.8700	0.9364	0.9798	0
-0.4339	-0.2876	-0.1341	0.0226	0.1787	0.3305	0.4741	0.6060	0.7231	0.8222	0.9012	0.9579	0.9910	0	0
-0.2225	-0.0672	0.0897	0.2445	0.3932	0.5322	0.6581	0.7679	0.8586	0.9283	0.9750	0	0	0	0
0	0.1565	0.3091	0.4541	0.5879	0.7073	0.8092	0.8912	0.9512	0.9878	0	0	0	0	0
0.2225	0.3723	0.5130	0.6410	0.7532	0.8468	0.9196	0.9698	0	0	0	0	0	0	0
0.4339	0.5695	0.6911	0.7957	0.8807	0.9440	0.9840	0	0	0	0	0	0	0	0
0.6235	0.7382	0.8346	0.9105	0.9640	0	0	0	0	0	0	0	0	0	0
0.7818	0.8698	0.9363	0.9797	0	0	0	0	0	0	0	0	0	0	0
0.9010	0.9578	0.9910	0	0	0	0	0	0	0	0	0	0	0	0
0.9749	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 1b. *Continuum solution at node points.*

the x-axis points toward the right, the y-axis points toward the left and the u-axis points upwards. The values of the $u(x_i, y_j)$ are given to four decimal places in Fig. 1b.

We introduce here the concept of "cell variation." The cell variation of a mesh function $z_{i,j}$ on the cell $C_{i,j} = [x_{i-1}, x_i] \times [y_{j-1}, y_j]$ is the sum of the absolute differences of the $z_{i,j}$ at adjacent vertices around the boundary of $C_{i,j}$:

$$\begin{aligned} &|z_{i-1,j-1} - z_{i-1,j}| + |z_{i-1,j} - z_{i,j}| \\ &\quad + |z_{i,j} - z_{i,j-1}| + |z_{i,j-1} - z_{i-1,j-1}|. \end{aligned} \tag{2.9}$$

The numerical total variation used by LeVeque and Goodman [14] in their analysis of two-dimensional TVD schemes is related to the cell variation (2.9). The cell variation is a truly two-dimensional measure of the steepness of z on cell $C_{i,j}$, one that is roughly independent of the orientation of the coordinate axes as long as the cells are small enough for the solution to be nearly monotonic on the cell. On cells on which z is smooth, the cell variation of z will be low to moderate. On cells on which z has a discontinuity of sufficient magnitude, the cell variation of z will be large. The cell variations of the values $u_{i,j}$ of Fig. 1a are plotted in Fig. 2a. Each node in Fig. 2a represents a cell variation for a cell of Fig. 1a. The discontinuity in Fig. 1a passes through 24 cells, the cells that make up the ridge of 24 nodes in Fig. 2a. Denoting each of these 24 cells by a symbol "S" (for "shear"), denoting each cell to the right of the "S" cells by a "0" (to indicate that these cells receive flow only from the parts of the boundary in (2.7c) and (2.7d)) and denoting each cell to the left of the "S" cells by a "1" (to indicate that these cells receive flow only from the part of the boundary in (2.7e)), one obtains the pattern in Fig. 2b.

In the following section, the disadvantages of solving (2.1) by the box scheme (2.6) are recalled.

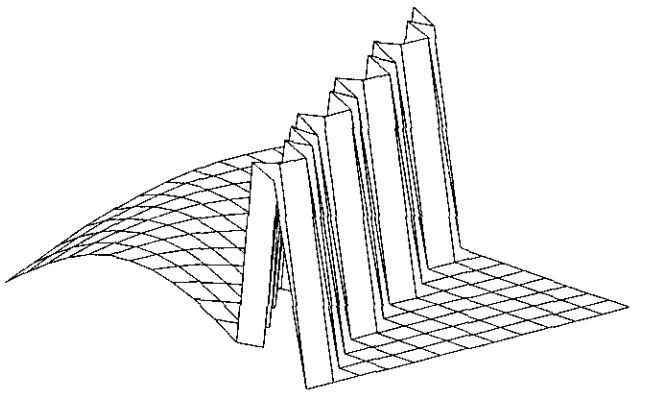

FIG. 2a. Cell variations of continuum solution.

3. THE NONOVERDETERMINED PROCEDURE

For each of the n^2 cells in the mesh, there is an equation (2.6). When boundary conditions (2.1b) are discretized as

$$\begin{aligned} u_{i,0} &= g_0(x_i), \qquad 0 \leqslant i \leqslant n, \\ u_{0,j} &= g_1(y_j), \qquad 1 \leqslant j \leqslant n, \end{aligned} \tag{3.1}$$

1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	S
1	1	1	1	1	1	1	1	1	1	1	1	S	S
1	1	1	1	1	1	1	1	1	1	S	S	S	0
1	1	1	1	1	1	1	1	1	S	S	0	0	0
1	1	1	1	1	1	1	S	S	S	0	0	0	0
1	1	1	1	1	1	S	S	0	0	0	0	0	0
1	1	1	1	S	S	S	0	0	0	0	0	0	0
1	1	1	S	S	0	0	0	0	0	0	0	0	0
1	1	S	S	0	0	0	0	0	0	0	0	0	0
S	S	S	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 2b. Cells of continuum solution through which discontinuity passes (24 "S" cells).

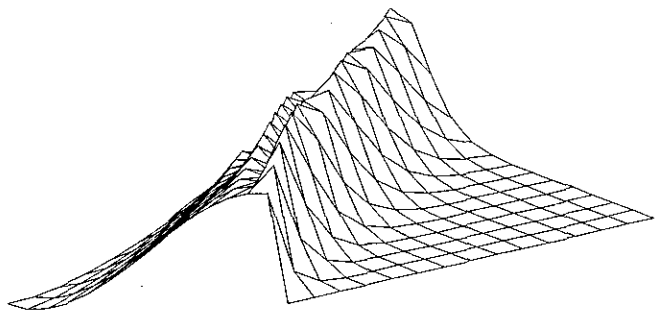

FIG. 3a. Nonoverdetermined solution.

and these values of $u_{i,j}$ are eliminated from Eqs. (2.6), there results a precisely determined (that is, neither underdetermined nor overdetermined) set of n^2 equations

$$r_{i,j}=0, \qquad 1\leqslant i\leqslant n, \quad 1\leqslant j\leqslant n, \tag{3.2}$$

for the n^2 unknowns $u_{i,j}$, $1\leqslant i\leqslant n$, $1\leqslant j\leqslant n$, which, in contrast to the overdetermined systems to be treated later in this paper, will be called the "nonoverdetermined" system. The process of solving (2.1) numerically using this nonoverdetermined system will be called the "nonoverdetermined procedure." The nonoverdetermined procedure is most efficiently carried out by marching in the x-direction.

The solution of system (3.2) with boundary conditions (3.1) was computed for the 28 cases

$$n=14, \tag{3.3a}$$

$$\tau=\tan\theta, \qquad \theta=0^\circ \text{ to } 45^\circ \text{ step } 5^\circ, \qquad \tau=\tfrac{1}{5},\tfrac{1}{4},\tfrac{1}{3},\tfrac{1}{2}, \tag{3.3b}$$

$$g_0(x)=0, \qquad 0\leqslant x\leqslant 1, \tag{3.3c}$$

$$g_1(y)=0, \qquad 0\leqslant y\leqslant h/2, \tag{3.3d}$$

$$g_1(y)=1, \quad g_1(y)=\cos(\pi y), \qquad h/2<y\leqslant 1 \tag{3.3e}$$

(14 different τ and 2 different functions g_1). The results for all of these cases were typified by the results for case (2.7) ($\tau=\tan 35^\circ$ and $g_1(y)=\cos(\pi y)$, $h/2<y\leqslant 1$), which are presented in Figs. 3a and 3b. These and all other computational results obtained by the author in the preparation of this paper were done in double-precision arithmetic on a VAX 8800 at the Center for Computational Statistics of George Mason University. As expected, the numerical solution of Figs. 3 is a good approximation of the continuum solution of Figs. 1 far from the discontinuity in the smooth regions but it is quite oscillatory near the discontinuity of the continuum solution at the line $y=\tau x+h/2$.

The oscillation is obvious also in the cell variations of the $u_{i,j}$. These cell variations, each represented by a point, are plotted in Fig. 4a. If no a priori information about the location of the discontinuity were available, one might attempt to identify the position of the discontinuity by identifying the cells in Fig. 4a with largest cell variation. If the 24 cells with largest variation are each denoted by "S" (24 cells for purposes of illustration because the discontinuity in the continuum solution passes through 24 cells; see Fig. 2b) and the remaining cells are each denoted by "C," one obtains the dispersed pattern in Fig. 4b. The "S" cells in Fig. 4b are not a connected set and their positions do not approximate the positions of the "S" cells in Fig. 2b well. The positions of the cells of the nonoverdetermined solution with largest cell variation are thus not a good guide to the position of the discontinuity in the continuum solution.

Instead of accepting the cells with largest variation as the cells that contain the discontinuity, one can build connectivity into the process in the following manner. First choose the cell with the largest cell variation. Label that cell an "S" cell and call it, for the time being, the "current 'S' cell." It is through this "S" cell and a band of additional "S" cells that the contact discontinuity (shear) will be assumed to pass. These additional "S" cells are identified as follows. Compare the cell variation in the cell to the right of the current "S" cell to the cell variation in the cell above the current "S" cell.

-1.0000	-0.9876	-0.9508	-0.8905	-0.8081	-0.7057	-0.5866	-0.4469	-0.3285	-0.0776	-0.1626	0.3563	0.3091	0.2362	0.5733
-0.9749	-0.9280	-0.8581	-0.7669	-0.6568	-0.5307	-0.3885	-0.2554	-0.0323	-0.0717	0.4380	0.3006	0.3080	0.7165	1.0480
-0.9010	-0.8218	-0.7223	-0.6049	-0.4726	-0.3274	-0.1839	0.0167	0.0227	0.5026	0.2949	0.4058	0.8550	1.0936	0.9751
-0.7818	-0.6744	-0.5503	-0.4125	-0.2641	-0.1139	0.0697	0.1170	0.5514	0.2990	0.5261	0.9771	1.0983	0.8818	0.5644
-0.6235	-0.4932	-0.3507	-0.1993	-0.0450	0.1265	0.2082	0.5869	0.3186	0.6623	1.0715	1.0613	0.7672	0.4485	0.2249
-0.4339	-0.2872	-0.1334	0.0230	0.1863	0.2940	0.6132	0.3571	0.8051	1.1284	0.9853	0.6402	0.3410	0.1573	0.0651
-0.2225	-0.0669	0.0902	0.2480	0.3730	0.6346	0.4153	0.9433	1.1410	0.8762	0.5099	0.2465	0.1042	0.0398	0.0140
0	0.1568	0.3105	0.4448	0.6554	0.4912	1.0650	1.1057	0.7429	0.3851	0.1680	0.0648	0.0227	0.0074	0.0023
0.2225	0.3728	0.5099	0.6792	0.5799	1.1588	1.0230	0.5961	0.2729	0.1066	0.0372	0.0120	0.0036	0.0010	0.0003
0.4339	0.5690	0.7082	0.6744	1.2148	0.8975	0.4472	0.1787	0.0619	0.0194	0.0057	0.0016	0.0004	0.0001	0.0000
0.6235	0.7429	0.7657	1.2259	0.7375	0.3076	0.1055	0.0320	0.0089	0.0023	0.0006	0.0001	0.0000	0.0000	0.0000
0.7818	0.8443	1.1887	0.5543	0.1870	0.0537	0.0140	0.0034	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
0.9010	1.1035	0.3611	0.0930	0.0216	0.0047	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.9749	0.1719	0.0303	0.0053	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 3b. Nonoverdetermined solution.

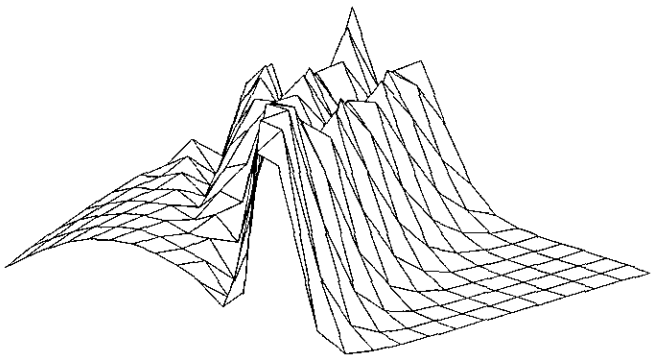

FIG. 4a. Cell variations of nonoverdetermined solution.

Choose the cell with the larger variation or, if the variations are equal, choose either cell and label this cell also an "S" cell. Now call this new "S" cell the current "S" cell and repeat the process in the previous three sentences, identifying on each step an additional "S" cell above or to the right of the current "S" cell, until the right or upper boundary is reached. Then do this whole procedure again, replacing "right" by "left" and "above" by "below," to identify a band of "S" cells to the left of and below the cell with the largest cell variation. There results a staircase band of "S" cells that is one cell wide and that divides the set of n^2 cells into two nontouching sets. The discontinuity is assumed to pass through the "S" cells. This process for identifying a path of cells that represents the discontinuity will be called "Process P" ("P" for "path" identification).

While Process P applied to a nonoscillatory numerical solution with a sharp layer near the line of discontinuity $y = \tau x + h/2$ could be expected to identify a path of cells that closely approximates the position of the line of discontinuity, the nonoverdetermined solution of Figs. 3 is so oscillatory that Process P is not successful in identifying a reasonable path. The "S" cells identified by Process P for the data of Figs. 3 are presented in Fig. 4c. The pattern in Fig. 4c is a poor approximation of the pattern in Fig. 2b.

```
C C C C C C C C C C C C C C
C C C C C C C C C C C C S C
C C C C C C C C S C C C C C
C C C C C C C C C S C C C C
C C C C C C C C S C C S C C
C C C C C C S S C C S C C C
C C C C C S C C S C C C C C
C C C C S C C S C C C C C C
C C C C S C S C C C C C C C
C C C S S S C C C C C C C C
C C S S C C C C C C C C C C
C S S C C C C C C C C C C C
S S C C C C C C C C C C C C
S C C C C C C C C C C C C C
```

FIG. 4b. Twenty-four "S" cells of nonoverdetermined solution with largest cell variation.

```
1 1 1 1 1 1 1 1 1 S 0 0 0 0
1 1 1 1 1 1 1 1 S S 0 0 0 0
1 1 1 1 1 1 1 1 S 0 0 0 0 0
1 1 1 1 1 1 1 1 S 0 0 0 0 0
1 1 1 1 1 1 1 S S 0 0 0 0 0
1 1 1 1 1 1 S S 0 0 0 0 0 0
1 1 1 1 1 S S 0 0 0 0 0 0 0
1 1 1 1 S S 0 0 0 0 0 0 0 0
1 1 1 1 S 0 0 0 0 0 0 0 0 0
1 1 1 S S 0 0 0 0 0 0 0 0 0
1 1 S S 0 0 0 0 0 0 0 0 0 0
1 S S 0 0 0 0 0 0 0 0 0 0 0
S S 0 0 0 0 0 0 0 0 0 0 0 0
S 0 0 0 0 0 0 0 0 0 0 0 0 0
```

FIG. 4c. "S" cells of nonoverdetermined solution determined by Process P.

Remark. By identifying "S" cells first in the direction "right and above" and then in the direction "left and below," Process P assumes that the contact discontinuity passes through the domain at an angle between 0° and 90° with respect to the x-axis. This is in consonance with the orientation of the contact discontinuity produced by a positive τ in (2.1a). More refined methods for identifying "S" cells, ones that do not make use of a priori knowledge of this type, will be used in future investigations.

The explanation for the erratic behavior of the nonoverdetermined procedure lies in the numerical error terms for (2.6). Equation (2.6) was obtained from (2.5) by trapezoidal integration (the fact that the trapezoidal integrals have been multiplied by a constant is not of concern here). The error expansions for the four trapezoidal integrations that produce the box scheme in (2.6) can be combined to produce an error expansion for the box scheme. This error expansion is dominated by the third-degree term

$$-\frac{h^3}{6}[u_{xyy} + \tau u_{xxy}]. \tag{3.4}$$

In regions where the solution u is smooth, this error term is a very small perturbation and the discretized conservation law (2.6) yields a numerical solution that is a good approximation of the continuum solution. However, near the discontinuity in the continuum solution at $y = \tau x + h/2$, the error term (3.4) is large and the discretized conservation law (2.6) yields a numerical solution that is, not surprisingly, a poor approximation of the continuum solution of (2.1). One could remedy this situation by attempting to do the numerical integration on cells with discontinuities in a more refined manner, one that takes into account the presence of the discontinuity. To do this, one would first have to identify the cells that contain the discontinuity and then determine the location of the discontinuity within these cells. However, precise knowledge of the cells that contain

the discontinuity and of the location of the discontinuity in those cells is not available in the nonoverdetermined solution (cf. Figs. 3). Even the solutions produced by the shock-capturing methods of [4, 5, 7, 8, 13, 16–19, 21], which are much more refined than the nonoverdetermined procedure discussed above, have discontinuities spread out over three or more cells, so that determination of the location of the discontinuities within cells is not possible.

The main source of error in system (3.2) is the erroneous numerical integration on the cells that contain the discontinuity. To eliminate this source of error from the system, we have two choices: (1) do accurate numerical integration on the cells that contain the discontinuity or (2) eliminate the equations for these cells from the system. The first choice involves not only finding the cells that contain the discontinuity but also determining the location of the discontinuity in these cells. This task is too large for the capabilities of current shock-capturing methods. We leave it for future work. The second choice merely involves identifying the cells that contain the discontinuity and does not require determining the location of the discontinuity within these cells. This smaller task, which is already large enough, will be treated in the present paper. If the goal of this paper were to treat time-dependent problems, we could not restrict our attention to this smaller task—determination of the precise location of the discontinuity within the cells would be required to proceed from time level to time level. We have, however, chosen to treat only steady-state equations here.

4. THE l_2 PROCEDURE

The reader will have noted that the strategy of eliminating from the system the equations for cells that contain a discontinuity will, in the framework of the nonoverdetermined procedure discussed in Section 3, lead to an underdetermined system. The way out of this apparent dilemma is to recall that the discretized conservation law (2.6) has higher-degree error terms. It is thus appropriate to solve system (3.2) with additional boundary conditions. We choose to pose "outflow" boundary conditions on the top and right boundary segments in addition to the "inflow" boundary conditions on the bottom and left boundary segments. Outflow boundary conditions (for example, vanishing first, second or third normal derivatives) are commonly used in the literature [4; 5, pp. 81ff; 13, pp. 1304ff; 21, p. 129]. Outflow boundary conditions create an overdetermination in system (3.2) that will allow the elimination of the main source of error in the system, namely, the equations for the cells that contain the discontinuity. The creation of this overdetermined system and its solution by a least-squares algorithm are carried out in the present section.

We choose here to impose outflow boundary conditions consisting of the vanishing of the (numerical) second derivative in the normal direction; that is, we require the solution at the boundary to be linear in the normal direction. To accomplish this, we introduce an additional layer of node points $(x_{n+1}, y_j) = (1+h, jh)$, $j = 0, 1, \ldots, n$, outside the right boundary segment and an additional layer of node points $(x_i, y_{n+1}) = (ih, 1+h)$, $i = 0, 1, \ldots, n+1$, outside the top boundary. At the right and top boundaries, we now impose the condition of linearity (vanishing second difference) in the normal direction:

$$u_{n-1,j} - 2u_{n,j} + u_{n+1,j} = 0, \qquad 0 \leqslant j \leqslant n+1, \tag{4.1a}$$

$$u_{i,n-1} - 2u_{i,n} + u_{i,n+1} = 0, \qquad 0 \leqslant i \leqslant n+1. \tag{4.1b}$$

At the same time, we add to system (3.2) the $2n+1$ equations (2.6) for i or $j = n+1$ so that there are a total of $(n+1)^2$ equations in the system:

$$r_{i,j} = 0, \qquad 1 \leqslant i \leqslant n+1, \quad 1 \leqslant j \leqslant n+1. \tag{4.2}$$

When the Dirichlet boundary conditions (3.1) are used to eliminate the $u_{i,j}$ on the bottom and left boundary segments ($j = 0$ and $i = 0$, respectively) and the outflow boundary conditions (4.1) are used to eliminate the $u_{i,j}$ outside the top and right boundary segments ($j = n+1$ and $i = n+1$, respectively), system (4.2) is a system of $(n+1)^2$ equations for the n^2 unknowns $u_{i,j}$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant n$.

In accordance with a long-established custom, one recently applied for fluid flows with shocks by Carey and Jiang [3], we first attempt to solve the overdetermined system of Eqs. (4.2) in the l_2 (least-squares) sense. The l_2 solution is the set $\{u_{i,j}\}$ that minimizes

$$\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} r_{i,j}^2. \tag{4.3}$$

The set $\{u_{i,j}\}$ that minimizes (4.3) will be called the "basic" l_2 solution ("basic" to distinguish it from the "final" l_2 solution introduced below).

Computational experiments were carried out for cases (3.3) using the IMSL subroutine DLSQRR, a least-squares solver based on a QR decomposition of the matrix. A

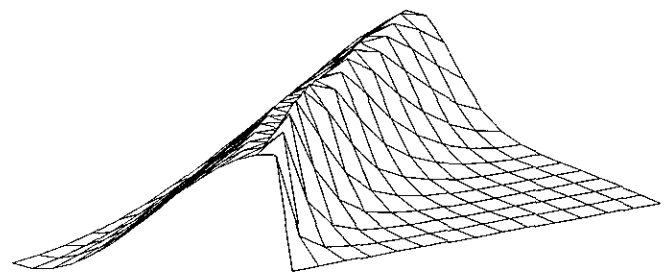

FIG. 5a. Basic l_2 solution.

-1.0000	-0.9747	-0.9309	-0.8674	-0.7842	-0.6828	-0.5649	-0.4336	-0.2905	-0.1436	0.0149	0.1659	0.3101	0.4588	0.5773
-0.9749	-0.9259	-0.8551	-0.7636	-0.6536	-0.5275	-0.3882	-0.2399	-0.0822	0.0670	0.2412	0.3745	0.5104	0.6635	0.7550
-0.9010	-0.8205	-0.7204	-0.6028	-0.4705	-0.3264	-0.1753	-0.0137	0.1263	0.3210	0.4214	0.5580	0.7303	0.8477	0.7825
-0.7818	-0.6736	-0.5490	-0.4112	-0.2631	-0.1093	0.0531	0.1891	0.3965	0.4591	0.6136	0.8000	0.8760	0.7753	0.6125
-0.6235	-0.4926	-0.3499	-0.1985	-0.0426	0.1187	0.2546	0.4652	0.4912	0.6800	0.8661	0.8814	0.7215	0.5274	0.3644
-0.4339	-0.2869	-0.1329	0.0242	0.1834	0.3212	0.5256	0.5224	0.7575	0.9205	0.8591	0.6464	0.4396	0.2683	0.1600
-0.2225	-0.0666	0.0909	0.2475	0.3873	0.5776	0.5576	0.8433	0.9537	0.8067	0.5549	0.3503	0.1975	0.0953	0.0439
0	0.1571	0.3109	0.4513	0.6227	0.6003	0.9313	0.9570	0.7251	0.4534	0.2651	0.1368	0.0582	0.0144	-0.0014
0.2225	0.3731	0.5122	0.6631	0.6526	1.0130	0.9229	0.6188	0.3495	0.1888	0.0879	0.0313	0.0025	-0.0090	-0.0099
0.4339	0.5696	0.7017	0.7137	1.0779	0.8475	0.4954	0.2511	0.1250	0.0510	0.0134	-0.0039	-0.0094	-0.0090	-0.0065
0.6235	0.7408	0.7801	1.1151	0.7305	0.3652	0.1648	0.0753	0.0254	0.0029	-0.0063	-0.0081	-0.0066	-0.0042	-0.0024
0.7818	0.8453	1.1146	0.5765	0.2396	0.0955	0.0398	0.0096	-0.0021	-0.0061	-0.0060	-0.0043	-0.0025	-0.0011	-0.0005
0.9010	1.0690	0.3943	0.1306	0.0456	0.0171	0.0015	-0.0032	-0.0045	-0.0037	-0.0024	-0.0012	-0.0005	-0.0001	0.0000
0.9749	0.1970	0.0482	0.0148	0.0048	-0.0011	-0.0021	-0.0022	-0.0016	-0.0009	-0.0004	-0.0001	0.0000	0.0001	0.0000
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 5b. Basic l_2 solution.

limiting factor in choosing the n in the numerical experiments was the full-matrix storage and large computing time required by DLSQRR (and by the IMSL subroutine DR2AV used for numerical experiments involving l_1 minimization reported in Section 5 below). In the future, solvers that require storage of and operations on only the nonzero blocks of the tightly banded matrix of system (4.2) will be used.

The basic l_2 solution for case (2.7), which is representative of the results for all of cases (3.3), is presented in Figs. 5a and 5b. The solution of Figs. 5 has a less oscillatory transition over the discontinuity than does the solution of the nonoverdetermined system of Figs. 3. Let us now investigate how this basic l_2 solution can be used to identify the cells that contain the discontinuity and how an improved l_2 solution can be obtained by eliminating the equations for those cells.

It has been an unwritten but common assumption in l_2 work for problems with discontinuities that the discontinuity is located in the cells that have equations with large absolute residuals. We show in Fig. 6a a plot of the absolute residuals of the basic l_2 solution of Figs. 5 (each vertex in Fig. 6a represents the absolute residual of an equation on a cell in Fig. 5a). The absolute residuals in Fig. 6a, while large in the general vicinity of the discontinuity, are oscillatory and are an unreliable guide to the exact location of the discontinuity. Recall that the discontinuity in the continuum solution passes through 24 cells, the 24 "S" cells of Fig. 2b. If the 24 cells of the basic l_2 solution with largest absolute residuals are each denoted by the symbol "S" and the other cells are each denoted by "C," one obtains the pattern in Fig. 6b. A comparison of the pattern of symbols "S" in Fig. 6b with that in Fig. 2b shows that the positions of the 24 largest absolute residuals of the basic l_2 solution are not a good guide to the location of the discontinuity in the continuum solution.

Let us now consider the cell variations. In a one-dimensional setting, Harten [7, pp. 158ff] identified cells with large absolute values of various derivatives as cells in which a discontinuity was presumed to exist. We investigate here in two dimensions the use of a related criterion, large cell variations. The cell variations of the basic l_2 solution of Figs. 5 are plotted in Fig. 7a (just as in Figs. 2a and 4a, each node in Fig. 7a represents a cell variation). If we label the 24 cells with largest cell variation "S" cells and the remaining cells "C" cells, we obtain the pattern in Fig. 7b. While this

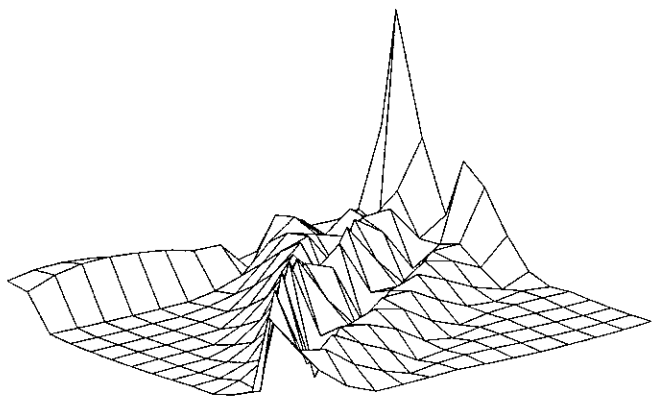

FIG. 6a. Absolute residuals of basic l_2 solution.

C	C	C	C	C	C	C	C	C	C	C	C	C	S
C	C	C	C	C	C	C	C	C	C	C	C	C	S
C	C	C	C	C	C	C	C	C	C	C	C	S	S
C	C	C	C	C	C	C	C	S	C	S	S	C	C
C	C	C	C	C	C	C	S	C	S	C	C	C	S
C	C	C	C	C	C	S	C	S	C	C	C	C	S
C	C	C	C	C	S	S	S	C	C	C	C	C	C
C	C	C	C	C	S	S	C	C	C	C	C	C	C
C	C	C	C	S	C	C	C	C	C	C	C	C	C
C	C	C	S	C	C	C	C	C	C	C	C	C	C
C	C	S	C	C	C	C	C	C	C	C	C	C	C
C	S	C	C	C	C	C	C	C	C	C	C	C	C
S	C	C	C	C	C	C	C	C	C	C	C	C	C
S	C	C	C	C	C	C	C	C	C	C	C	C	C

FIG. 6b. Twenty-four "S" cells of basic l_2 solution with largest absolute residuals.

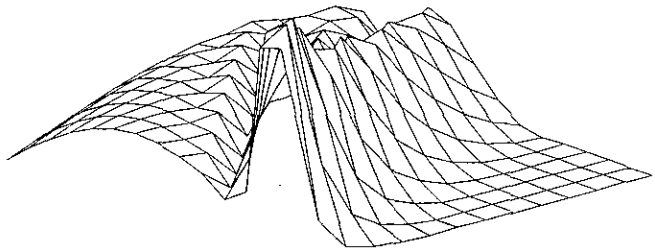

FIG. 7a. Cell variations of basic l_2 solution.

pattern of symbols "S" is a much better approximation of the pattern of Fig. 2b than the pattern for the nonoverdetermined solution in Fig. 4b, it still leaves something to be desired. The connection between the domain influenced by the boundary conditions (7c), (2.7d) and the domain influenced by boundary condition (2.7e) has not been completely broken (that is, the "S" cells do not divide the remaining cells into two nonoverlapping sets). If one eliminates the equations for the "S" cells of Fig. 7b from system (4.2), there will still remain in the system some cells (equations) through which the discontinuity must pass and these cells will be the source of smearing and oscillation. Also, note in Fig. 7b the wide swath of "S" cells in the lower left corner. There, the discontinuity is not captured in one but rather in two cells. When this is the case, there are nodes (x_i, y_j) that are surrounded by four "S" cells: nodes (x_3, y_3), (x_4, y_4), (x_4, y_5), and (x_5, y_5) of Figs. 5. If one omits the equations for the "S" cells of Fig. 7b from system (4.2), the resulting system will be ill-posed since the unknowns $u_{3,3}$, $u_{4,4}$, $u_{4,5}$, and $u_{5,5}$ will not occur in this new system. Thus, identifying the cells of the l_2 solution with largest cell variation as "S" cells and omitting them from the system is not a viable procedure.

If, instead of omitting the cells with largest cell variation, the path-identification algorithm, Process P, is used to identify a one-cell-wide path of "S" cells, the pattern in Fig. 7c

C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	S	C
C	C	C	C	C	C	C	C	C	C	S	S	C	C
C	C	C	C	C	C	C	C	S	S	C	C	C	C
C	C	C	C	C	C	C	S	S	C	C	C	C	C
C	C	C	S	S	S	S	S	C	C	C	C	C	C
C	C	C	S	S	S	C	C	C	C	C	C	C	C
C	C	S	S	S	C	C	C	C	C	C	C	C	C
C	S	S	S	C	C	C	C	C	C	C	C	C	C
S	S	C	C	C	C	C	C	C	C	C	C	C	C
S	C	C	C	C	C	C	C	C	C	C	C	C	C

FIG. 7b. Twenty-four "S" cells of basic l_2 solution with largest cell variation.

1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	S
1	1	1	1	1	1	1	1	1	1	1	S	S	S
1	1	1	1	1	1	1	1	1	1	S	S	0	0
1	1	1	1	1	1	1	1	S	S	S	0	0	0
1	1	1	1	1	1	1	S	S	0	0	0	0	0
1	1	1	1	1	S	S	S	0	0	0	0	0	0
1	1	1	S	S	S	0	0	0	0	0	0	0	0
1	1	S	S	0	0	0	0	0	0	0	0	0	0
1	S	S	0	0	0	0	0	0	0	0	0	0	0
S	S	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 7c. "S" cells of basic l_2 solution determined by Process P.

results. The pattern of "S" cells in Fig. 7c differs somewhat from the pattern of "S" cells in Fig. 2b, especially in the lower left corner, but it is a much better approximation of that pattern than is the pattern of Fig. 7b or the patterns of Figs. 6b, 4c, and 4b.

Accepting the "S" cells of Fig. 7c as cells through which the numerical discontinuity passes, we now propose to eliminate from system (4.2) the equations corresponding to these "S" cells, since it is these equations that are presumed to introduce most of the error into the system. This new system has, however, a disadvantage: it is ill-conditioned for τ near one and is ill-posed for $\tau = 1$. This is easily seen by noting that the coefficients $-1+\tau$ of $u_{i-1,j}$ and $1-\tau$ of $u_{i,j-1}$ in (2.6) become small as $\tau \to 0$ and the problem of calculating the $u_{i-1,j}$ and $u_{i,j-1}$ at nodes adjacent to three "S" cells becomes ill-conditioned. To illustrate, consider the nodes $(x_1, y_1) = (h, h)$ and $(x_1, y_2) = (h, 2h)$ in Fig. 7c. In system (4.2) with the equations for the "S" cells of Fig. 7c omitted, the unknown $u_{1,1}$ occurs in only one equation, Eq. (2.6) on cell $C_{2,1}$ (the other three adjoining cells are "S" cells and are omitted from the system). The coefficient of $u_{1,1}$, which is $-1+\tau$, is small for τ near one. Similarly, the unknown $u_{1,2}$ occurs only in the Eq. (2.6) on cell $C_{1,3}$ and has a coefficient $1-\tau$ that is small for τ near one. A brief glance at Fig. 7c shows that there are many unknowns $u_{i,j}$ occurring in only one equation with a coefficient of $-1+\tau$ or $1-\tau$. When $\tau = 1$, the coefficients of these unknowns are zero and the system is ill-posed.

This situation can be remedied in the following manner. All points that are surrounded by precisely three "S" cells are identified and divided into two groups: Group 0 consists of those nodes, like (x_1, y_1) in Fig. 7c, that are to the right of the path of "S" cells (these nodes have three "S" cells just to the left and above) and Group 1 consists of those nodes, like (x_1, y_2), that are to the left of the path of "S" cells (these nodes have three "S" cells just to the right and below). To Group 0 are added the boundary mesh point $(x_0, y_0) = (0, 0)$ just below the discontinuity in the

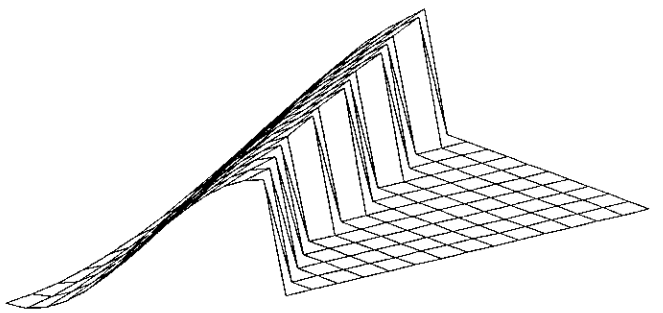

FIG. 8a. Final l_2 solution.

boundary conditions at $(0, h/2)$, any point on the right boundary between two "S" cells and any point on the top boundary with an "S" cell just to its left. To Group 1 are added the boundary mesh point $(x_0, y_1) = (0, h)$ just above the discontinuity in the boundary conditions at $(0, h/2)$, any point on the right boundary with an "S" cell just below it and any point on the top boundary between two "S" cells.

Now the points of Group 0 are connected by a linear spline (straight-line segments linking nearest points). For each segment from (x_{i_1}, y_{j_1}) to (x_{i_2}, y_{j_2}), $i_1 < i_2$, $j_1 < j_2$, of this linear spline, the finite-volume equation corresponding to the conservation law (2.1a) is created on the right-triangular cell T below this segment, which serves as the hypothenuse, that is, on the triangle consisting of the three segments connecting (x_{i_1}, y_{j_1}), (x_{i_2}, y_{j_1}), and (x_{i_2}, y_{j_2}). This finite-volume equation is

$$\int_{\text{hypothenuse}} (v_1 + v_2\tau)\, u\, ds - \tau \int_{x_{i_1}}^{x_{i_2}} u(x, y_{j_1})\, dx + \int_{y_{j_1}}^{y_{j_2}} u(x_{i_2}, y)\, dy = \iint_{\mathrm{T}} (u_x + \tau u_y)\, dA = 0, \quad (4.4)$$

where $(v_1, v_2) = (j_1 - j_2,\ i_2 - i_1)/\sqrt{(i_2 - i_1)^2 + (j_2 - j_1)^2}$ is the outward (upward) normal on the hypothenuse of the triangle. Analogously, the points of Group 1 are connected by a linear spline (straight-line segments linking nearest points). For each segment from (x_{i_1}, y_{j_1}) to (x_{i_2}, y_{j_2}), $i_1 < i_2$, $j_1 < j_2$, of this linear spline, the finite-volume equation corresponding to the conservation law (2.1a) is created on the right-triangular cell T above this segment, which serves as the hypothenuse, that is, on the triangle consisting of the three segments connecting (x_{i_1}, y_{j_1}), (x_{i_1}, y_{j_2}), and (x_{i_2}, y_{j_2}). This finite-volume equation is

$$\int_{\text{hypothenuse}} (v_1 + v_2\tau)\, u\, ds - \int_{y_{j_1}}^{y_{j_2}} u(x_{i_1}, y)\, dy + \tau \int_{x_{i_1}}^{x_{i_2}} u(x, y_{j_2})\, dx = \iint_{\mathrm{T}} (u_x + \tau u_y)\, dA = 0, \quad (4.5)$$

where $(v_1, v_2) = (j_2 - j_1,\ i_1 - i_2)/\sqrt{(i_2 - i_1)^2 + (j_2 - j_1)^2}$ is the outward (downward) normal on the hypothenuse of the triangle. The "T" cells of Eqs. (4.4) and (4.5) are triangular cells "along the discontinuity." They include portions of the "S" cells that were omitted from system (4.2).

Discretization of Eqs. (4.4) and (4.5) is accomplished by the trapezoidal rule in the following manner. The line integrals over the hypothenuses in (4.4) and (4.5) are discretized by the basic trapezoidal rule for one interval. The line integrals in x are discretized by the composite trapezoidal rule for $i_2 - i_1$ equal intervals. The line integrals in y are discretized by the composite trapezoidal rule for $j_2 - j_1$ equal intervals. The discretized equations are then multiplied by $1/h$ (so as to make the weighting of these equations correspond to the weighting of Eqs. (2.6)). The resulting equations for all of the "T" cells corresponding to the line segments in Groups 0 and 1 are now added to the system (4.2) from which the equations for the "S" cells have been omitted. This system will be called System L2. The l_2 solution of System L2 will be called the "final l_2 solution."

The final l_2 solution for each of cases (3.3) was calculated. Representative of these results was the final l_2 solution for case (2.7), which is presented in Figs. 8a and 8b. The final l_2 solution is a much better approximation of the continuum

-1.0000	-0.9747	-0.9308	-0.8672	-0.7840	-0.6825	-0.5650	-0.4343	-0.2938	-0.1469	0.0018	0.1499	0.2931	0.4281	0.5546
-0.9749	-0.9259	-0.8550	-0.7634	-0.6534	-0.5275	-0.3891	-0.2419	-0.0898	0.0637	0.2127	0.3585	0.4929	0.6164	0.7371
-0.9010	-0.8204	-0.7202	-0.6027	-0.4707	-0.3276	-0.1773	-0.0239	0.1299	0.2761	0.4216	0.5504	0.6690	0.7886	0.9026
-0.7818	-0.6735	-0.5490	-0.4115	-0.2645	-0.1117	0.0426	0.1956	0.3390	0.4835	0.6050	0.7194	0.8370	0.9502	1.0640
-0.6235	-0.4927	-0.3502	-0.1998	-0.0453	0.1091	0.2607	0.4014	0.5435	0.6564	0.7679	0.8847	0.9972	1.1104	0
-0.4339	-0.2872	-0.1340	0.0216	0.1756	0.3249	0.4627	0.6008	0.7047	0.8152	0.9320	1.0442	0	0	0
-0.2225	-0.0673	0.0887	0.2416	0.3881	0.5226	0.6551	0.7501	0.8621	0.9790	1.0895	0	0	0	0
0	0.1558	0.3068	0.4499	0.5805	0.7058	0.7934	0.9090	1.0261	0	0	0	0	0	0
0.2225	0.3711	0.5100	0.6358	0.7529	0.8353	0.9567	1.0730	0	0	0	0	0	0	0
0.4339	0.5680	0.6881	0.7968	0.8769	1.0058	0	0	0	0	0	0	0	0	0
0.6235	0.7371	0.8375	0.9169	0	0	0	0	0	0	0	0	0	0	0
0.7818	0.8719	0.9372	0	0	0	0	0	0	0	0	0	0	0	0
0.9010	0.9563	0	0	0	0	0	0	0	0	0	0	0	0	0
0.9749	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 8b. Final l_2 solution.

solution (Figs. 1) than is the basic l_2 solution (Figs. 5) or the nonoverdetermined solution (Figs. 3). This is not surprising, since the main source of error, the equations for the cells through which the discontinuity is presumed to pass, have been eliminated from the system. There is still room for improvement, however. The location of the discontinuity in Figs. 8 is somewhat in error, especially near the origin, and the solution at the points just to the left of the discontinuity is on the order of 10% larger than the continuum solution. The maximum absolute error of the $u_{i,j}$ to the right of the discontinuity versus the continuum solution $u(x_i, y_j) = 0$ is $0.171 * 10^{-15}$. The maximum absolute error of the $u_{i,j}$ to the left of the discontinuity versus the continuum solution $u(x_i, y_j) = \cos(\pi(y_j - \tau x_i))$ is 0.131.

The term "l_2 procedure" denotes the following algorithm: (1) l_2 solution of system (4.2); (2) identification of a band of "S" cells with large cell variation by Process P, (3) creation of System L2, which is system (4.2) with the equations for the "S" cells omitted and the equations for the "T" cells added; and (4) l_2 solution of System L2.

While the l_2 method is the long-established first choice for solving overdetermined systems, it is not the only alternative. A lesser known competitor will be investigated in the next section.

5. THE l_1 PROCEDURE

The l_1 method has been shown in [9–12] to perform well in capturing shocks in solutions of one- and two-dimensional conservation laws. We investigate in this section how the l_1 method can be used to capture contact discontinuities.

We start by solving the overdetermined system (4.2) in the l_1 sense. The l_1 solution is the set $\{u_{i,j}\}$ that minimizes

$$\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} |r_{i,j}|. \tag{5.1}$$

The set $\{u_{i,j}\}$ that minimizes (5.1) will be called the "basic" l_1 solution ("basic" to distinguish it from the "final" l_1 solution introduced below).

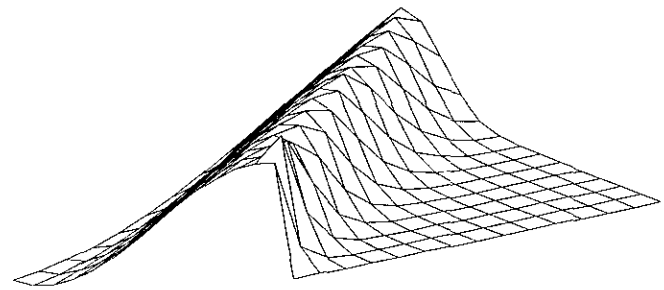

FIG. 9a. Basic l_1 solution.

To carry out l_1 minimization, the IMSL (Ed. 1.0) subroutine DR2AV, an implementation of the l_1 algorithm of Barrodale and Roberts [1; 2, pp. 186–201] was used. The Barrodale–Roberts algorithm was used previously in [9, 11] for solving one-dimensional problems with shocks. The Barrodale–Roberts algorithm is expensive—it requires at least $O(n^4)$ and perhaps as many as $O(n^6)$ operations [2, pp. 231–236]. More efficient l_1 algorithms such as the Bartels–Conn–Sinclair l_1 algorithm [2, pp. 202–211] and the Bloomfield–Steiger l_1 algorithm [2, pp. 212–218], both of which require $O(n^4)$ operations, were not used because they were not readily available in coded form. The l_1 algorithm of Seneta and Steiger [2, pp. 237–258; 20], which was used for one-dimensional flows in [10], was not used here because a basis for the null space of the matrix of the system of Eqs. (4.3) was not readily available. The l_1 algorithm of [12] was not used because that algorithm assumes that the cells with nonzero residuals form a connected set and, for contact discontinuities, that is not the case, as will be seen below.

Calculations were carried out for all of the cases (3.3). The basic l_1 solution for case (2.7), which is representative of these results, is presented in Figs. 9a and 9b. The basic l_1 solution has a transition over the discontinuity that is nearly nonoscillatory and is sharper than that of the basic l_2

-1.0000	-0.9689	-0.9213	-0.8555	-0.7714	-0.6700	-0.5527	-0.4240	-0.2817	-0.1367	0.0121	0.1600	0.3080	0.4560	0.6039
-0.9749	-0.9280	-0.8581	-0.7669	-0.6568	-0.5305	-0.3903	-0.2440	-0.0826	0.0690	0.2234	0.3713	0.5193	0.6673	0.8152
-0.9010	-0.8218	-0.7223	-0.6049	-0.4726	-0.3281	-0.1780	-0.0159	0.1343	0.2878	0.4346	0.5819	0.7262	0.8533	0.7642
-0.7818	-0.6744	-0.5503	-0.4125	-0.2644	-0.1112	0.0506	0.1991	0.3518	0.4969	0.6416	0.7733	0.8494	0.7038	0.5110
-0.6235	-0.4932	-0.3507	-0.1994	-0.0440	0.1167	0.2634	0.4150	0.5576	0.6960	0.8052	0.8224	0.6202	0.3755	0.2500
-0.4339	-0.2872	-0.1334	0.0233	0.1821	0.3269	0.4771	0.6153	0.7420	0.8188	0.7732	0.5264	0.2904	0.1376	0.0868
-0.2225	-0.0669	0.0903	0.2467	0.3896	0.5376	0.6703	0.7892	0.8828	0.7043	0.4278	0.2139	0.0928	0.0362	0.0220
0	0.1568	0.3103	0.4512	0.5959	0.7203	0.8210	0.8506	0.6068	0.3299	0.1488	0.0588	0.0210	0.0070	0.0041
0.2225	0.3728	0.5112	0.6508	0.7622	0.8339	0.7939	0.4991	0.2399	0.0968	0.0345	0.0113	0.0034	0.0010	0.0006
0.4339	0.5690	0.7005	0.7931	0.8256	0.7152	0.3877	0.1621	0.0577	0.0184	0.0054	0.0015	0.0004	0.0001	0.0001
0.6235	0.7429	0.8097	0.7949	0.6188	0.2791	0.0991	0.0307	0.0087	0.0023	0.0006	0.0001	0.0000	0.0000	0.0000
0.7818	0.8443	0.9386	0.5102	0.1792	0.0524	0.0138	0.0034	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
0.9010	1.1035	0.3611	0.0930	0.0216	0.0047	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.9749	0.1719	0.0303	0.0053	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 9b. Basic l_1 solution.

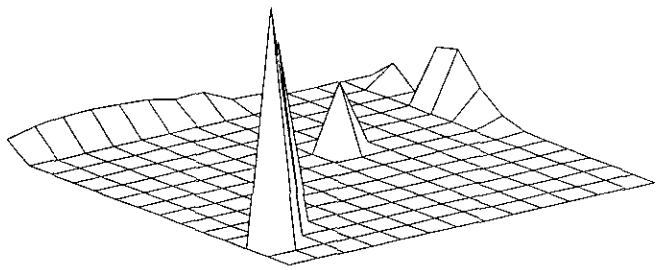

FIG. 10a. Absolute residuals of basic l_1 solution.

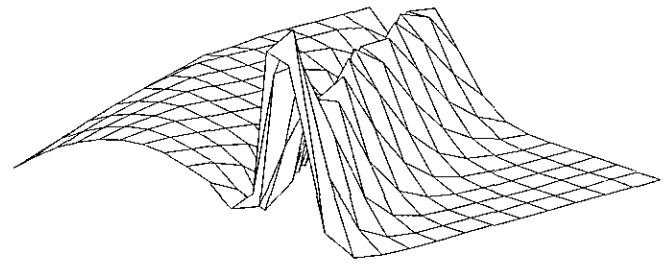

FIG. 11a. Cell variations of basic l_1 solution.

solution of Figs. 5. The nearly nonoscillatory capturing of the discontinuity in Figs. 9 in approximately three cells is comparable to the accuracy achieved by the best current methods. While exact comparisons are not readily available, Leonard's SHARP method for (2.1a) with $\tau = \tan 34°$, piecewise constant boundary conditions and a 25×25 mesh requires around four cells to capture the discontinuity [13, p. 1312, Fig. 18].

The basic l_1 solution will now be used to identify the cells that contain the discontinuity and to obtain an improved solution by eliminating the equations for these cells. We first address the question of whether the discontinuity is located in the cells that have equations with large absolute residuals, which is often assumed. In [12], the equations that had large absolute residuals were the equations that corresponded to the cells through which a shock passed. In the more difficult contact-discontinuity case being considered in the present paper, however, the equations that have large absolute residuals are not, in general, the equations of the cells through which the discontinuity passes. We show in Fig. 10a a plot of the absolute residuals of the basic l_1 solution of Figs. 9 (each node in Fig. 10a represents the absolute residual of an equation on a cell in Fig. 9a). The large absolute residuals in Fig. 10a are not clustered around the line of discontinuity. Recall that the discontinuity in the continuum solution passes through 24 cells, the 24 "S" cells of Fig. 2b. If the 24 cells of the basic l_1 solution with largest absolute residuals are each denoted by a symbol "S" and the other cells are each denoted by "C," one obtains the pattern in Fig. 10b. A comparison of Fig. 10b with Fig. 2b shows that the positions of the 24 largest residuals of the basic l_1 solution are not at all a good guide to the location of the discontinuity in the continuum solution.

The cell variations of the basic l_1 solution of Figs. 9 are plotted in Fig. 11a (each node represents a cell variation). If we label the 24 cells with largest cell variation "S" cells and the remaining cells "C" cells, we obtain the pattern in Fig. 11b. This pattern of symbols "S" is a better approximation of the pattern of Fig. 2b than is the pattern for the basic l_2 solution in Fig. 7b. Note in particular that the wide swath of "S" cells in the lower left corner of Fig. 7b (l_2 solution) has been replaced by a narrower path of "S" cells in Fig. 11b. Nevertheless, the connection between the domain influenced by boundary conditions (2.7c), (2.7d) and the domain influenced by boundary condition (2.7e) has still not been completely broken (that is, the "S" cells do not divide the remaining cells into two nonoverlapping sets). If one eliminates the equations for the "S" cells of Fig. 11b from system (4.2), there will still remain in the system some cells (equations) through which the discontinuity must pass and these cells will be the source of smearing and oscillation. Moreover, if one omits the equations for the "S" cells of Fig. 11b from system (4.2), the resulting system will be

```
S S S S S S S S S S C C C C
C C C C C C C C C C C C C S
C C C C C C C C C C C C C C
C C C C C C C C C C C C C S
C C C C C C C C C C C C C S
C C C C C C C C S C C C C S
C C C C C C C C C C C C C S
C C C C C C C C C C C C C S
C C C C C C C C C C C C C S
C C C C C C C C C C C C C S
C C S C C C C C C C C C C S
C S C C C C C C C C C C C S
C C C C C C C C C C C C C S
C C C C C C C C C C C C C C
```

FIG. 10b. Twenty-four "S" cells of basic l_1 solution with largest absolute residuals.

```
C C C C C C C C C C C C C C
C C C C C C C C C C C C C C
C C C C C C C C C C C C C C
C C C C C C C C C C C C S S
C C C C C C C C C C C C S C
C C C C C C C C C C C S S C
C C C C C C C C C C S S C C
C C C C C C C C C S S C C C
C C C C C C C C S S C C C C
C C C C C C C S S S C C C C
C C C C C S S C C C C C C C
C C S S S C C C C C C C C C
C S S S C C C C C C C C C C
S S C C C C C C C C C C C C
S C C C C C C C C C C C C C
```

FIG. 11b. Twenty-four "S" cells of basic l_1 solution with largest cell variation.

1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	S	S
1	1	1	1	1	1	1	1	1	1	1	S	S	0
1	1	1	1	1	1	1	1	1	S	S	S	0	0
1	1	1	1	1	1	1	1	S	S	0	0	0	0
1	1	1	1	1	1	1	S	S	0	0	0	0	0
1	1	1	1	1	S	S	S	0	0	0	0	0	0
1	1	1	1	S	S	0	0	0	0	0	0	0	0
1	1	1	S	S	0	0	0	0	0	0	0	0	0
1	S	S	S	0	0	0	0	0	0	0	0	0	0
S	S	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 11c. "S" cells of basic l_1 solution determined by Process P.

ill-posed since the unknown $u_{3,3}$ is surrounded by four "S" cells and will therefore not occur in this new system. Thus, identifying the cells of the l_1 solution with largest cell variation as "S" cells and omitting them from the system is not a viable procedure.

If, instead of omitting the cells with largest cell variation, the path-identification algorithm, Process P, is used to identify the "S" cells, the pattern in Fig. 11c results. The pattern of "S" cells in Fig. 11c differs only slightly from the pattern of "S" cells in Fig. 2b. It is a much better approximation of that pattern that is the pattern of Fig. 7c or 4c. Accepting the "S" cells of Fig. 11c as cells through which the numerical discontinuity passes, we now eliminate from system (4.2) the equations corresponding to the "S" cells of Fig. 11c. We proceed further as in Section 4. We identify two groups of nodes surrounded by precisely three "S" cells: Group 0 consists of such nodes to the right of the path of "S" cells along with certain boundary nodes; Group 1 consists of such nodes to the left of the path of "S" cells along with certain boundary nodes. The points of Group 0 and Group 1 are connected by linear splines and the finite-volume equations (4.4) and (4.5) are created on the triangular "T" cells with the segments of the linear splines as hypothenuses. Discretization of Eqs. (4.4) and (4.5) is accomplished by the trapezoidal rule. The discretized equations are multiplied by $1/h$. The resulting equations for all the "T" cells corresponding to the line segments in Groups 0 and 1 are added to the system (4.2) from which the equations for the "S" cells have been omitted. This system will be called System L1. The l_1 solution of System L1 is called the "final l_1 solution."

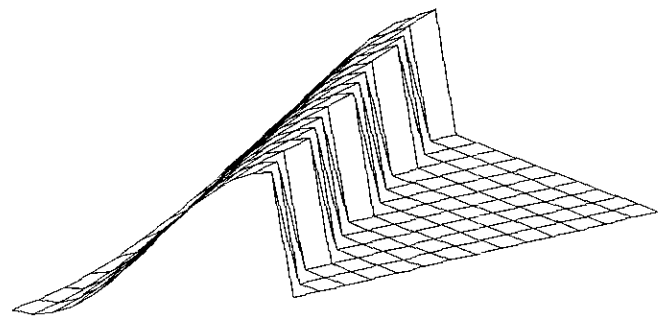

FIG. 12a. Final l_1 solution.

The final l_1 solution for each of cases (3.3) was calculated. Representative of these results was the final l_1 solution for case (2.7), which is presented in Figs. 12a and 12b. The final l_1 solution is an excellent approximation of the continuum solution (Figs. 1), one that is superior not only to the basic l_1 solution (Figs. 9) but also to the final l_2 solution (Figs. 8). In contrast to the final l_2 solution, the final l_1 solution has a more accurate path of cells representing the discontinuity and a smaller overshoot to the left of the discontinuity. To the right of the discontinuity, the maximum absolute error of the $u_{i,j}$ of the final l_1 solution versus the continuum solution $u(x_i, y_j) = 0$ is zero (for the final l_2 solution, it was $0.171 * 10^{-15}$). To the left of the discontinuity, the maximum absolute error of the $u_{i,j}$ of the final l_1 solution versus the continuum solution $u(x_i, y_j) = \cos(\pi(y_j - \tau x_i))$ is 0.0375 (for the final l_2 solution, it was 0.131).

The term "l_1 procedure" denotes the following algorithm:

-1.0000	-0.9689	-0.9213	-0.8555	-0.7714	-0.6699	-0.5531	-0.4230	-0.2837	-0.1355	0.0130	0.1611	0.3092	0.4524	0.5690
-0.9749	-0.9280	-0.8581	-0.7669	-0.6568	-0.5304	-0.3910	-0.2414	-0.0878	0.0733	0.2245	0.3726	0.5207	0.6582	0.7425
-0.9010	-0.8218	-0.7223	-0.6049	-0.4725	-0.3285	-0.1761	-0.0207	0.1396	0.2883	0.4351	0.5791	0.7036	0.7786	0.8630
-0.7818	-0.6744	-0.5503	-0.4125	-0.2646	-0.1099	0.0466	0.2052	0.3511	0.4960	0.6337	0.7449	0.8131	0.8991	0.9834
-0.6235	-0.4932	-0.3507	-0.1995	-0.0433	0.1135	0.2697	0.4128	0.5548	0.6842	0.7825	0.8467	0.9356	0	0
-0.4339	-0.2872	-0.1335	0.0236	0.1799	0.3331	0.4729	0.6108	0.7304	0.8170	0.8804	0.9736	0	0	0
-0.2225	-0.0669	0.0904	0.2454	0.3951	0.5312	0.6636	0.7725	0.8494	0.9139	0	0	0	0	0
0	0.1568	0.3097	0.4553	0.5872	0.7129	0.8105	0.8792	0.9415	0	0	0	0	0	0
0.2225	0.3726	0.5137	0.6406	0.7583	0.8448	0.9070	0.9682	0	0	0	0	0	0	0
0.4339	0.5698	0.6911	0.7997	0.8757	0.9334	0	0	0	0	0	0	0	0	0
0.6235	0.7382	0.8372	0.9036	0.9582	0	0	0	0	0	0	0	0	0	0
0.7818	0.8708	0.9291	0.9815	0	0	0	0	0	0	0	0	0	0	0
0.9010	0.9528	0	0	0	0	0	0	0	0	0	0	0	0	0
0.9749	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIG. 12b. Final l_1 solution.

(1) l_1 solution of system (4.2); (2) identification of a band of "S" cells with large cell variation by Process P; (3) creation of System L1, which is system (4.2) with the equations for the "S" cells omitted and the equations for the "T" cells added; and (4) l_1 solution of System L1. We note in passing that, in contrast to the nonoverdetermined procedure, neither the l_1 nor the l_2 procedure can be carried out in toto by marching. It is, however, possible to replace steps 4 of the l_1 and l_2 procedures by marching procedures if the outflow boundary conditions are partially relaxed.

6. DISCUSSION

Direct comparisons of the l_1 procedure and the l_2 procedure with other shock-capturing methods will be done in the future. These comparisons will be done on the basis of both accuracy and computing speed. We discuss here what is currently known.

The accuracy of the basic l_1 solution is already equivalent to that achieved by standard methods. This solution can, however, be improved by eliminating the main source of the error, namely, the cells through which the discontinuity passes. One might ask why one could not improve the results of standard time-dependent shock-capturing methods by eliminating the cells that contain the discontinuity. Unfortunately, most standard methods have fixed stencils and do not permit the elimination of cells. When cells are eliminated by shifting stencils (as in ENO), standard methods become unstable because of temporal waves passing through the eliminated cells. Since the cells through which the discontinuity passes cannot be eliminated, the numerical equations that have stencils extending over these cells must be "corrected." However, knowing how to correct the numerical equations for the discontinuity requires knowing the locus of the discontinuity. Neither operator-splitting methods nor the truly higher-dimensional algorithms mentioned in the Introduction provide information about the locus of the discontinuity with sufficient accuracy to obtain solutions without smearing or oscillation. Both the l_1 and the l_2 procedures discussed in this paper are inherently steady-state procedures that do not have temporal instability problems. In these procedures, the main source of error in the numerical solution, namely, the equations for the cells that contain the discontinuity, can be eliminated. Standard methods are forced to retain these equations in the system and "correct" them using inadequate information about the location of the discontinuity.

The box scheme (2.6) was shown in Section 3 to produce oscillation when used to compute solutions with discontinuities. It is significant that the basic l_1 solution of Figs. 9 is nearly nonoscillatory in spite of the fact that the underlying scheme is prone to oscillation. In contrast, the basic l_2 solution of Figs. 5 has some oscillation. This oscillation in the basic l_2 solution is not sufficiently pronounced to prevent the final l_2 solution from being a good approximation of the continuum solution for the cases treated in the present paper but it does suggest that, in more complicated cases, l_2 should be used only if there is evidence that the oscillations it produces will not be so severe as to prevent identification of nearly correct "S" cells.

The basic l_1 solution not only has much less oscillation than does the basic l_2 solution, it also has less diffusion. This comment may sound strange to those of us who are accustomed to the seemingly universal trade-off between oscillation and diffusion. However, the computational results (see Figs. 5 and 9) show this to be the case. The basic l_2 solution of Figs. 5 has some oscillation and requires four to five cells for transition over the contact discontinuity. The basic l_1 solution is nearly nonoscillatory and accomplishes the transition over the contact discontinuity in three cells.

The l_1 procedure is more accurate than the l_2 procedure and other shock-capturing procedures but it remains to be seen whether its computing speed will be competitive. A comparison of the computing speed of the l_1 procedure with the computing speeds of the l_2 procedure and standard methods was not carried out because the IMSL subroutines DR2AV and DLSQRR used to produce the l_1 and l_2 results of the present paper require storage for and computations with full matrices. Future implementations of the l_1 and l_2 methods will take into account the tightly banded structure of the matrix of system (4.2) to reduce storage and operation count. At present one can say only that it is expected that a band-matrix l_2 procedure analogous to the algorithm in [3] will require computing time on the order of the computing time required by standard methods and than a band-matrix l_1 procedure will be slower than a band-matrix l_2 procedure.

While the subject of the present paper is solution of steady-state equations, some comments on the use of l_1-based methods for time-dependent equations are in order. For time-dependent equations, merely identifying the "S" cells that contain the discontinuity will not be sufficient. To propagate the discontinuity correctly in time, one must determine the location of the discontinuity inside these cells. The location of the discontinuity could be determined by solving the conservation law locally on a fine grid on the path of "S" cells using boundary conditions provided by the final l_1 solution of the global problem. This approach has been successfully carried out for refining the position of a shock in the solution of one-dimensional Euler equations [11]. This is a topic for future investigation.

7. CONCLUSION

The l_1 procedure described in this paper captures steady-state contact discontinuities in bands of cells that are only

one cell wide and are located close to the position of the discontinuity in the continuum solution. The numerical values on each side of the discontinuity are nonoscillatory and highly accurate right up to the edge of the discontinuity. The less expensive l_2 procedure produces final results nearly as good as those of the l_1 procedure for the cases investigated in this paper. However, diffusive and oscillatory phenomena that occur in the basic l_2 solution suggest caution in replacing l_1 by l_2.

REFERENCES

1. I. Barrodale and F. D. K. Roberts, *SIAM J. Numer. Anal.* **10**, 839 (1973).
2. P. Bloomfield and W. L. Steiger, *Least Absolute Deviations* (Birkhäuser, Boston, 1983).
3. G. F. Carey and B. N. Jiang, *Int. J. Numer. Methods Eng.* **26**, 81 (1988).
4. S. R. Chakravarthy, *AIAA J.* **21**, 699 (1983).
5. S. F. Davis, *J. Comput. Phys.* **56**, 65 (1984).
6. J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv. *Adv. Appl. Math.* **6**, 259 (1985).
7. A. Harten, *J. Comput. Phys.* **83**, 148 (1989).
8. C. Hirsch, C. Lacor, and H. Deconinck, AIAA Paper 87-1163, 1987.
9. J. E. Lavery, *J. Comput. Phys.* **79**, 436 (1988).
10. J. E. Lavery, *SIAM J. Numer. Anal.* **26**, 1081 (1989).
11. J. E. Lavery, *J. Comput. Phys.* **86**, 1 (1990).
12. J. E. Lavery, *SIAM J. Numer. Anal.* **28**, 141 (1991).
13. B. P. Leonard, *Int. J. Numer. Methods Fluids* **8**, 1291 (1988).
14. R. J. LeVeque and J. B. Goodman, Lectures in Applied Mathematics, Vol. 22-2, edited by B. E. Engquist *et al.* (Am. Math. Soc., Providence, RI, 1985), p. 51.
15. K. W. Morton and M. F. Paisley, *J. Comput. Phys.* **80**, 168 (1989).
16. J. Peraire, M. Vahdati, K. Morgan, and O. C. Zienkiewicz, *J. Comput. Phys.* **72**, 449 (1987).
17. K. G. Powell and B. van Leer, NASA TM 102029, 1989 (unpublished).
18. P. L. Roe, *J. Comput. Phys.* **43**, 357 (1981).
19. P. L. Roe, *J. Comput. Phys.* **63**, 458 (1986).
20. E. Seneta and W. L. Steiger, *Discrete Appl. Math.* **7**, 79 (1984).
21. P. Woodward and P. Collela, *J. Comput. Phys.* **54**, 115 (1984).