JOURNAL OF COMPUTATIONAL PHYSICS 113, 304–308 (1994)

On the Numerical Solution of Conservation Laws by Minimizing Residuals

R. B. LOWRIE AND P. L. ROE

Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan 48109

Received April 8, 1993

The numerical solution of conservation laws by minimizing the residuals of an overdetermined set of discrete equations is studied. Previous research has shown that for certain formulations, minimizing the residuals in the L_1 norm will yield solutions that resolve discontinuities that are very sharp and correctly placed. In this study, we analyze a previously proposed method that numerically solves the 2D advection equation with discontinuous data. The method is able to resolve the discontinuity over one mesh cell, without generating spurious oscillations. However, we have found that incorrect solutions are generated for some data. This had led us to formulate and prove two theorems concerning these results. We also provide an analysis of the solution procedure, along with suggestions for developing future schemes that are more applicable to a wide range of problems.

1. INTRODUCTION

Recently Lavery [1, 2] put forward a radical new concept for shock capturing. By adding to the inviscid equations a dissipative term required only to be larger than the rounding error he formed an over-determined set of discrete equations and sought the solution that minimized the residuals in the L_1 norm. The tolerance shown by this norm for isolated large errors allowed the solution to generate shock discontinuities that were very sharp and correctly placed.

It is tempting to hypothesize that good schemes can be designed more generally by over-determining the solution "somehow," and then minimizing in some norm having a similar tolerance of large errors. Consider a 2D advection equation, given by

$$u_x + \tan(\alpha)\, u_y = 0 \qquad \text{in } \Omega, \tag{1a}$$

where $\Omega = \{(x, y) \in R^2 : 0 < x < 1, \quad 0 < y < 1\}$ with boundary Γ. For $0 < \alpha < 90°$, specify the inflow boundary conditions as

$$u = u_L \qquad \text{on } \Gamma_1 = \{(x, y) \in \Gamma : x = 0\}, \tag{1b}$$

$$u = u_R \qquad \text{on } \Gamma_2 = \{(x, y) \in \Gamma : x > 0 \text{ and } y = 0\}, \tag{1c}$$

with $u_L > u_R$. The exact solution is a jump discontinuity along the line $y = x \tan(\alpha)$, with the constant state $u = u_L$ above the line and $u = u_R$ below. Jiang [3, 4] has overdetermined the solution of (1) by subdividing a square mesh of spacing h into triangles, and reports impressive results. However, we have found that incorrect solutions are generated for some data. This had led us to formulate and prove two theorems concerning these results and to provide analysis of the solution procedure described in [4]. In this way we show that the new concepts, although still attractive, must be formulated with great caution.

2. ANALYSIS

To solve (1) numerically, Jiang [4] proposed subdividing a square mesh into triangles. For a mesh with N^2 cells, the divided mesh gives $2N^2$ residuals with N^2 unknowns. The system is over-determined, and some minimization procedure is employed. Linear triangles are chosen as the basis for u, with the discrete projection of u given by

$$u_h(x, y) = \sum_i u_i \psi_i(x, y), \tag{2}$$

where $\psi_i(x, y)$ is the shape function, which is the "tent function" for linear triangles. u_i is the value of $u(x, y)$ at $x = x_i$, $y = y_i$. The residual of Eq. (1a) for triangle T, with boundary ∂T, is written as

$$R_T = \oint_{\partial T} \{u_h\, dy - \tan(\alpha)\, u_h\, dx\}. \tag{3}$$

For the subdivided square mesh, the residuals may then be expressed as (see Fig. 1)

$$R_{\text{upper}} = u_{sw} + (v - 1)\, u_{nw} - v u_{ne}, \tag{4a}$$

$$R_{\text{lower}} = u_{ne} + (v - 1)\, u_{se} - v u_{sw}, \tag{4b}$$

where $v = \cot(\alpha)$.

0021-9991/94 $6.00

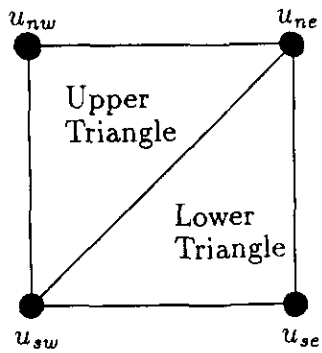

FIG. 1. A subdivided square mesh cell, with mesh values referenced in Eq. (4).

Define a weighted residual norm as

$$Lw_p = \sum_{i=1}^{2N^2} w_i |R_i|^p, \tag{5}$$

where R_i is the ith triangle residual and w_i is its corresponding weight factor. In order that the norm maintains its "distance" properties, we require that

$$0 < w_i < \infty. \tag{6}$$

For all $w_i = 1$, (5) reduces to the L_p norm,

$$L_p = \sum_{i=1}^{2N^2} |R_i|^p, \tag{7}$$

Throughout the remainder of this note, the following notation will be used:

- Let $\{S_h\}$ be the class of schemes whose solutions to (1) minimize (5). Note that this definition permits solutions at local minima, and therefore the solution may depend on the initial data.
- Let $\{u_h^1\}$ be the class of piecewise linear functions given by (2), that separate two constants states (u_L, u_R) by a discontinuity spanning only one mesh interval. Figure 2 is an example of one such function, to be discussed later.
- Let $\{S_h^1\} \subset \{S_h\}$ be the class of schemes that generate solutions in $\{u_h^1\}$.

An "ideal" scheme for solving (1) would be in $\{S_h^1\}$, *and* the solution would propagate the discontinuity at the correct angle. In other words, the hope is that a scheme exists in $\{S_h^1\}$ that gives the weak solution to (1) as $h \to 0$.

To generate solutions in $\{u_h^1\}$, the minimization procedure proposed in [4] is to minimize initially in L_2 and then to refine the solution using a weighted L_2 procedure to be described later. Unfortunately, we have found that the subset of $\{u_h^1\}$ *that propagates the discontinuity at the correct* angle is not at a local Lw_p minimum. The remainder of this note will prove two theorems concerning the feasibility of solutions in $\{u_h^1\}$, along with a discussion of the properties of the minimization procedure in [4].

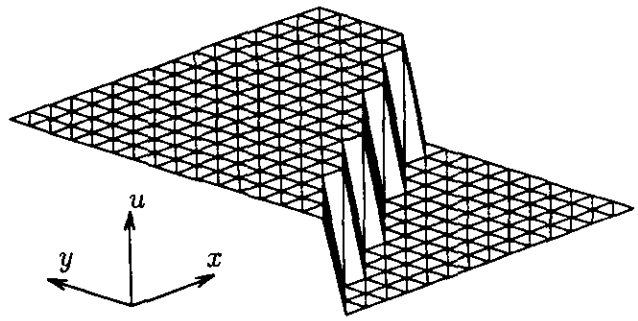

FIG. 2. Carpet plot, solution for $\alpha = 35°$, $N = 15$, $u_L = 2$, $u_R = 1$.

2.1. $\{S_h^1\}$ *Schemes*

THEOREM. *There are no schemes in* $\{S_h^1\}$ *that give the weak solution to* (1) *as* $h \to 0$.

Proof. The weak solution of (1) propagates the discontinuity at an angle α; therefore, we need to prove that any u_h^1 function that propagates the discontinuity at an angle α does not minimize (5).

Referring to Fig. 3, consider a segment of a u_h^1 function with $u_L^* = u_L$. In order for the function to propagate the discontinuity at the correct angle, this situation must arise somewhere along the discontinuity for $0° < \alpha < 45°$. Note that $v > 1$ for this range of α. Now let the value u_L^* be perturbed from u_L to $u_L - \delta$. Using (4), the six non-zero residuals of this segment contribute ΔLw_p to the overall Lw_p norm, given by

$$\begin{aligned}\Delta Lw_p = {} & \delta^p(w_1(v-1)^p + w_6) + w_2(\Delta u - v\delta)^p \\ & + (\Delta u - \delta)^p (w_4(v-1)^p + w_3) \\ & + w_5((v-1)\Delta u - v\delta)^p,\end{aligned} \tag{8}$$

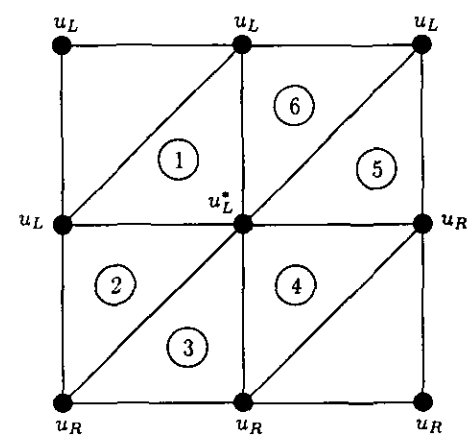

FIG. 3. 2×2 cell segment from a solution in $\{u_h^1\}$. Triangles with non-zero residuals are numbered.

where $\Delta u = u_L - u_R$, and $w_{1\to 6}$ are the weight factors corresponding to the triangles numbered in Fig. 3. This gives

$$\left.\frac{\partial Lw_p}{\partial \delta}\right|_{\delta=0} = -p(\Delta u)^{p-1} \times \{w_3 + w_2 v + w_4(v-1)^p + w_5 v(v-1)^{p-1}\}, \tag{9}$$

which is strictly negative. Therefore, at least for $0° < \alpha < 45°$, a u_h^1 function that propagates the discontinuity at the correct α is not at a local Lw_p minimum, and, in particular, it is not at a local L_1 minimum. A similar argument can be used for $45° < \alpha < 90°$.

Remark. In this proof it is assumed that the weights w_i are fixed, independent of the solution. The argument also holds for solution-dependent weights, as long as we "freeze" the weights during the minimization of (5). This in fact is the approach of the minimization procedure to be described in Section 2.3.

2.2. *Minimizing L_p in $\{u_h^1\}$*

THEOREM. *Considering only functions in $\{u_h^1\}$, those which globally minimize (7) do not propagate the discontinuity at the correct angle.*

Proof. Consider minimizing (7) over functions in $\{u_h^1\}$. Such solutions can be made up of stacked rows of cells, each of the family shown in Fig. 4. The number of each type of these rows is defined as M_k, where k is the number of cells the discontinuity is displaced for that particular row. Define the exit boundaries as

$$\Gamma_3 = \{(x, y) \in \Gamma : x = 1\}, \tag{10a}$$

$$\Gamma_4 = \{(x, y) \in \Gamma : y = 1\}. \tag{10b}$$

There are two possible constraints on the M_k values. If the discontinuity intersects the Γ_3 boundary, then the following constraint must be satisfied:

$$N = \sum_{i=1}^{N} iM_i, \qquad 0° < \alpha < 45°. \tag{11}$$

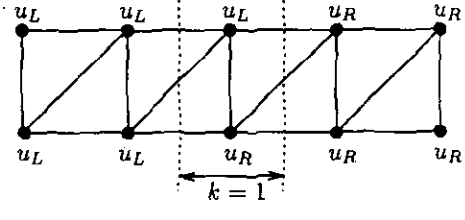

FIG. 4. Row of cells from a solution in $\{u_h^1\}$, $k = 1$, $N = 4$.

Otherwise, if the discontinuity intersects the Γ_4 boundary,

$$N = \sum_{i=0}^{N} M_i, \qquad 45° \leqslant \alpha < 90°. \tag{12}$$

For a function in $\{u_h^1\}$, using (4) gives

$$L_p = 2(u_L - u_R)^p \left\{ M_0 v^p + \sum_{k=1}^{N} M_k [|v-1|^p + k - 1] \right\}. \tag{13}$$

Already it is apparent that the norm is independent of the order in which the rows are stacked. The M_k are now chosen in such a way as to minimize L_p.

First consider those solutions under constraint (11). M_0 only increases L_p, so it must be zero for this case. M_1 may be eliminated using (11) to obtain

$$L_p = 2(u_L - u_R)^p \left\{ N(v-1)^p + f(v) \sum_{k=2}^{N} (k-1) M_k \right\}, \tag{14}$$

where $f(v) = 1 - (v-1)^p$. By inspecting Eq. (14), minimizing L_p is strictly dependent on the sign of $f(v)$:

- $f(v) > 0$: $M_1 = N$; $M_k = 0$ for $k \geqslant 2$. This solution corresponds to the discontinuity propagating at a 45° angle.
- $f(v) < 0$: $M_N = 1$; $M_k = 0$ for $k < N$. This solution corresponds to the discontinuity propagating along the x-axis.

As an example, for the L_1 norm, the cross-over between the above solutions occurs at $\cot(\alpha) = 2$, or $\alpha \approx 26.6°$.

For the constrain (12), M_0 may be eliminated from (13) to give

$$L_p = 2(u_L - u_R)^p \left\{ Nv^p + \sum_{k=1}^{M} M_k [(1-v)^p - v^p + k - 1] \right\}. \tag{15}$$

Given that $0 < v \leqslant 1$, to minimize (15), $M_k = 0$ for $k \geqslant 2$. Therefore,

$$L_p = 2(u_L - u_R)^p (Nv^p - g(v) M_1), \tag{16}$$

where $g(v) = v^p - (1-v)^p$. This result has similar properties to the result from the previous constraint, namely,

- $g(v) > 0$: $M_1 = N$; $M_0 = 0$. Again, this solution corresponds to the discontinuity propagating at a 45° angle.
- $g(v) < 0$: $M_1 = 0$; $M_0 = N$ for $k < N$. This solution corresponds to the discontinuity propagating along the y-axis.

These arguments show that for a general α, considering only functions in $\{u_h^1\}$, the function that propagates the discontinuity at the correct angle is not at a global minimum of an L_p norm.

2.3. *Schemes That Allow* $w_i \to \infty$

Jiang [4] suggested minimizing (5) with $p = 2$ (weighted least squares) and the weighting

$$w_i = \left(\sum_{n=1}^{\text{edges}} |u^+ - u^-|_{i,n} \right)^{-m}, \tag{17}$$

where m is a positive integer and $(u^+, u^-)_{n,i}$ are the endpoint values of the nth edge of the ith cell. For triangles this expression is equivalent to

$$w_i = (u_{\max} - u_{\min})_i^{-m}, \tag{18}$$

where $u_{\max}$, $u_{\min}$ are the respective maximum and minimum values for the ith triangle. The numerical procedure is as follows:

1. Initialize weight factors w_i to 1.
2. Minimize Eq. (5) using a weighted L_2 procedure.
3. Check convergence of solution with that of the previous weight factors used. If not converged, update the weight factors with Eq. (18) and return to step 2.

For $m = 6$, we are able to reproduce the results reported in [4], as shown in Fig. 2. Within computer round-off error, this solution is a u_h^1 function that propagates the discontinuity at approximately the correct α. Since these results show that this scheme will tolerate some large residuals, it is tempting to think of it as somehow approximating the L_1 norm.

Given the analysis of Section 2.1, there appears to be a discrepancy. We may be apt to suppose that since the weights w_i are solution-dependent, the previous analysis is invalid for this procedure. Note, however, that the weights are "frozen" during the minimization process in step 2.

The reason the previous analysis is invalid for this procedure is that the weightings given by (18) become unbounded in regions of constants u, violating constraint (6). By allowing $w_i \to \infty$, it is clear that no true norm is being measured; this scheme is not in $\{S_h^1\}$. This could be encouraging, because the analysis of Section 2.1 shows that, in this context at least, norm minimization is not desirable.

The question that remains is how is Jiang's procedure able to produce solutions in $\{u_h^1\}$. In the remainder of this section, we will put forth one possible explanation, which is largely based on numerical experiments.

Again consider the segment of a u_h^1 function in Fig. 3. Note that as the iteration process converges to the u_h^1 function, of the six numbered triangles, in theory the residuals R_1 and R_6 approach zero, while their respective weights $w_{1,6} \to \infty$. The other residuals and weights $(2 \to 5)$ remain finite and non-zero. We will now show that under certain circumstances, the products $w_1 |R_1|$ and $w_6 |R_6|$ are non-zero, significantly altering Eq. (9).

Consider the case where the scheme is nearly converged to a u_h^1 function, so that

$$|R_1| = \varepsilon_1, \tag{19}$$

$$|R_6| = \varepsilon_6, \tag{20}$$

where $0 < \varepsilon_{1,6} \ll 1$. Note that in application, $\varepsilon_{1,6}$ could be the result of computer round-off error. Proceeding as in Section 2.1, by perturbing u_L^* from u_L to $u_L - \delta$ gives

$$|R_1| = \delta(\nu - 1) + \varepsilon_1, \tag{21}$$

$$|R_6| = \delta + \varepsilon_6. \tag{22}$$

To first order in $\varepsilon_{1,6}$, this gives

$$\left.\frac{\partial L w_2'}{\partial \delta}\right|_{\delta=0} = \left.\frac{\partial L w_2}{\partial \delta}\right|_{\delta=0} + 2(w_1 \varepsilon_1 (\nu - 1) + w_2 \varepsilon_2), \tag{23}$$

where the first term is given by Eq. (9) for $p = 2$. In order for this expression to be positive, indicating a local minimum with respect to u_L^*, it is required that

$$w_1 \varepsilon_1 (\nu - 1) + w_2 \varepsilon_2 > -\frac{1}{2} \left.\frac{\partial L w_2}{\partial \delta}\right|_{\delta=0}. \tag{24}$$

Comparing Eqs. (4) and (18), we can assume that $w_{1,6} \sim |R_{1,6}|^{-m} = (\varepsilon_{1,6})^{-m}$, which gives

$$w_i \varepsilon_i \sim (\varepsilon_i)^{1-m}, \qquad i = 1, 6. \tag{25}$$

This relation indicates that (24) may be satisfied for large m. Indeed, for the $\alpha = 35°$ case, we have found good results for $m \geqslant 2$. For $1 < m < 2$, the solution found either propagates at the wrong angle or is not in $\{u_h^1\}$.

Note that to prevent computer overflow, in application, we must set

$$w_i := \min(w_{\max}, w_i), \tag{26}$$

where normally $w_{\max}$ is chosen as a function of round-off error:

$$w_{\max} > 1/\sigma, \tag{27}$$

$$\sigma = \min_{\sigma \geqslant 0} \quad \text{such that} \quad 1 + \sigma \neq 1. \tag{28}$$

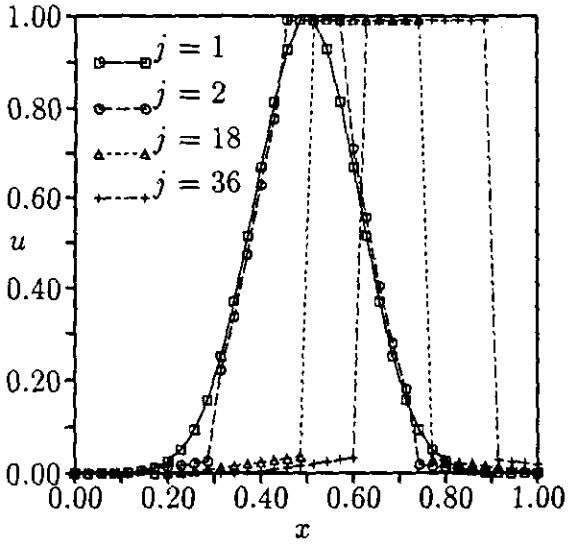

FIG. 5. Solution along several mesh lines ($j = [1, 36]$ corresponds to $y = [0, 1]$). Gaussian inflow, $\alpha = 75°$, $N = 35$.

We have observed that diffusion of the discontinuity increases by decreasing $w_{\max}$; however, as long as $w_{\max}$ is chosen as above, the solution is in $\{u_h^1\}$ to within round-off error.

The unbounded weightings have the effect of ignoring residuals that lie along the discontinuity while forcing residuals away from the discontinuity to be identically satisfied. Apparently, the initial L_2 solution ($w_i = 1$) locates the discontinuity in approximately the correct location, and the weighting procedure (18) refines this solution to give a solution in $\{u_h^1\}$. However, we stress that this solution is *not* at an L_1 minimum, as has been stated in [3].

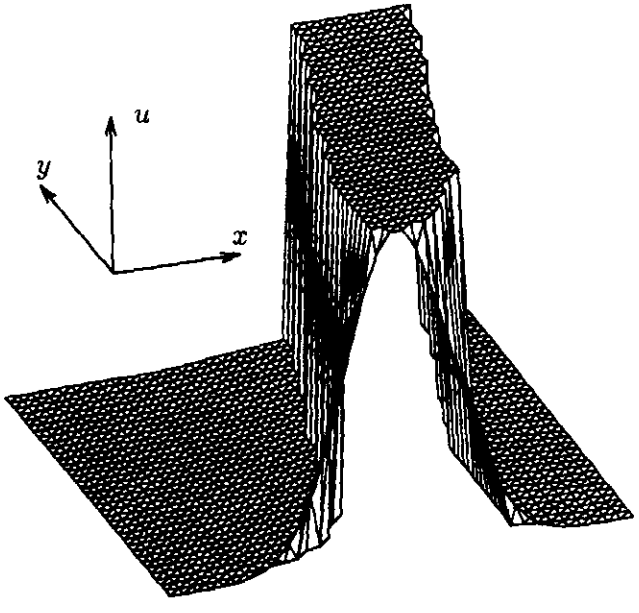

FIG. 6. Carpet plot corresponding to Fig. 5.

Although this procedure may appear promising as a way of handling discontinuities, we found that it has difficulties with smooth data. Consider a change in the boundary conditions of (1) to

$$u = 0 \quad \text{on} \quad \Gamma_1 = \{(x, y) \in \Gamma : x = 0\}, \tag{29a}$$

$$u = \exp[-40(x - \tfrac{1}{2})^2] \quad \text{on} \quad \Gamma_2 = \{(x, y) \in \Gamma : x > 0 \text{ and } y = 0\}. \tag{29b}$$

Figures 5 and 6 show results for $\alpha = 75°$. Note how the solution procedure misinterprets high-gradient regions as discontinuities. Clearly this solution procedure is unacceptable for smooth data and is reminiscent of results from over compressive flux-limiting schemes.

3. CONCLUSIONS

The minimization of residual norms has been considered for the solution of conservation laws. By studying 2D linear advection, the following may be concluded:

- Solutions in $\{u_h^1\}$ cannot minimize any weighted residual norm, while at the same time propagating the discontinuity at the correct angle.
- By allowing unbounded weight factors, a procedure may be used to obtain solutions in $\{u_h^1\}$ that apparently propagate a discontinuity at the correct angle. However, this method is unacceptable for smooth data, with results that are similar to those of over compressive flux-limiting schemes.

Given the quality of the solutions presented by [3, 4], these conclusions are somewhat distressing. However, there are several comments that can be made with regards to developing future schemes that are more applicable to a wide range of problems.

Essentially the scheme discussed in this note, and those in [1, 2], may be summarized as schemes that somehow discard key residuals. This concept certainly deserves more attention. As with the flux-limiting approach, somehow we must balance the ability to track discontinuities, with the ability to accurately represent smooth data. Furthermore, when discarding residuals, we must ensure that conservation is satisfied in some sense. Also under consideration is allowing discontinuities to be resolved over more than one cell. Each of these issues should be addressed in future work.

REFERENCES

1. J. E. Lavery, *J. Comput. Phys.* **79**, 436 (1988).
2. J. E. Lavery, *SIAM J. Numer. Anal.* **26** (5), 1081 (1989).
3. B. N. Jiang, ICOMP Report 91-03, 1991.
4. B. N. Jiang, *J. Comput. Phys.* **105**, 108 (1993).