NASA Technical Memorandum 106226

New Developments in the Method of Space-Time Conservation Element and Solution Element—Applications to the Euler and Navier-Stokes Equations

Sin-Chung Chang
Lewis Research Center
Cleveland, Ohio

Prepared for the
Second U.S. National Congress on Computational Mechanics
sponsored by the U.S. Association for Computational Mechanics,
August 16–18, 1993

NEW DEVELOPMENTS IN THE METHOD OF SPACE-TIME CONSERVATION ELEMENT AND SOLUTION ELEMENT—APPLICATIONS TO THE EULER AND NAVIER-STOKES EQUATIONS

Sin-Chung Chang
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135

Abstract

A new numerical framework for solving conservation laws is being developed. This new approach differs substantially in both concept and methodology from the well-established methods–i.e., finite difference, finite volume, finite element, and spectral methods. It is conceptually simple and designed to overcome several key limitations of the above traditional methods.

A two-level scheme for solving the convection-diffusion equation

$$\partial u/\partial t + a\,\partial u/\partial x - \mu\,\partial^2 u/\partial x^2 = 0$$

is constructed and used to illuminate major differences between the current method and those mentioned above. This *explicit* scheme, referred to as the a-μ scheme, has the unusual property that its stability is limited only by the *CFL* condition, i.e., it is independent of μ. Also it will be shown that the amplification factors of the a-μ scheme are identical to those of the Leapfrog scheme if $\mu = 0$, and to those of the DuFort-Frankel scheme if $a = 0$. These coincidences are unexpected because the a-μ scheme and the above classical schemes are derived from completely different perspectives, and the a-μ scheme *does not* reduce to the above classical schemes in the limiting cases.

The a-μ scheme is extended to solve the 1–D time-dependent Navier-Stokes equations of a perfect gas. Stability of this *explicit* solver also is limited only by the *CFL* condition. In spite of the fact that it does not use (i) *any* techniques related to the high-resolution upwind methods, and (ii) *any* ad hoc parameter, the current *Navier-Stokes* solver is capable of generating highly accurate shock tube solutions. Particularly, shock discontinuites can be resolved within one mesh interval.

The inviscid ($\mu = 0$) a-μ scheme is neutrally stable, i.e., free from numerical diffusion. Such a scheme generally can not be extended to solve the Euler equations. Thus, the inviscid version is modified. Stability of this modified scheme, referred to as the a-ϵ scheme, is limited by the *CFL* condition and $0 \leq \epsilon \leq 1$ where ϵ is a special parameter that controls numerical diffusion. Moreover, if $\epsilon = 0$, the amplification factors of the a-ϵ scheme are identical to those of the Leapfrog scheme, which has no numerical diffusion. On the other hand, if $\epsilon = 1$, these amplification factors unexpectedly become identical to each other and to the amplification factor of the highly diffusive Lax scheme. Note that, because the Lax scheme is very diffusive and it uses a mesh that is staggered in time, a two-level scheme using such a mesh is often associated with a highly diffusive scheme. The a-ϵ scheme, which also uses a mesh staggered in time, demonstrates that it can also be a scheme with no numerical diffusion.

The a-ϵ scheme is extended to become an Euler solver. The extension has stability conditions similar to those of the a-ϵ scheme. It also has the unusual property that numerical diffusion at all mesh points can be controlled by a set of local parameters. Moreover, it will be shown that the Euler extension is capable of generating accurate shock tube solutions with the *CFL* number ranging from 0.88 to 0.022.

1. Introduction

The method of space-time conservation element and solution element [1-3] is a new numerical framework for solving conservation laws. This new approach differs substantially in both concept and methodology from the well-established methods–i.e., finite difference, finite volume, finite element, and spectral methods [4-8]. It is conceived and designed to overcome several key limitations of the above traditional methods. Thus, we shall begin this paper with a discussion of several considerations that motivate the current development:

(a) A set of physical conservation laws is a collection of statements of *flux conservation in space-time*. Mathematically, these laws are represented by a set of integral equations. The differential form of these laws is obtained from the integral form with the assumption that *the physical solution is smooth*. For a physical solution in a region of rapid change (e.g., a boundary layer), this smoothness assumption is difficult to realize by a numerical approximation that can use only a limited number of discrete variables. This difficulty becomes even worse in the presence of discontinuities (e.g., shocks). Thus, a method designed to obtain numerical solutions to the differential form without enforcing flux conservation is at a fundamental disadvantage in modeling physical phenomena with high-gradient regions. Particularly, it may not be used to solve flow problems involving shocks. Contrarily, a numerical solution obtained from a method that also enforces flux-conservation locally (i.e., down to a computational cell) and globally (i.e., over the entire computational domain) will always retain the basic physical reality of flux conservation even in a region involving discontinuities. For this reason, the enforcement of *both local and global flux conservation in space and time* is a tenet in the current development. As will be shown, the concept of *conservation element* is introduced to serve this purpose.

Among the traditional methods, finite difference, finite element, and spectral methods are designed to solve the differrential form of the conservation laws. Note that the set of integral equations usually solved in a finite-element scheme is equivalent to the differential form of the conservation laws assuming certain smoothness conditions. However, these integral equations generally are different from the integral equations representing the conservation laws. Even if they are cast into a conservative form, the resulting flux-conservation conditions generally do not represent the physical conservation laws.

The finite volume method is the only traditional method designed to enforce flux conservation. A finite-volume scheme may enforce flux conservation in space only, or in both space and time. As a preliminary to this enforcement, a flux must be assigned at any interface separating two neighboring conservation cells. In a typical finite-volume scheme, it is evaluated by extrapolating or interpolating the mesh values at the neighboring cells. This evaluation generally requires an ad hoc choice of a special flux model among many models available [9-11]. Generally numerical results obtained are dependent on which model one chooses.

Contrarily, by design, making the above ad hoc choice is not needed in the current

numerical framework. As will be shown, by using the concept of *solution element*, flux evaluation at an interface becomes an integral part of the solution procedure and requires no interpolation or extrapolation.

(b) The numerical variables used in a spectral method, i.e., the expansion coefficients, are global parameters pertaining to the entire computational domain. As a result, a spectral method generally (i) lacks local flexibility and thus may be applied only to problems with simple geometry, and (ii) is hindered by the fact that it must deal with a full matrix that is diffucult to invert.

By design, only local parameters will be used in the current method. Moreover, the set of discrete variables in any one of the numerical equations to be solved is either associated with a single solution element or a few immediately neighboring solution elements. Thus, one needs only to deal with a very sparse matrix. As will be shown, the maximum number of solution elements involved in a numerical equation of the current framework is independent of the order of accuracy of a particular scheme. Contrarily, the order of accuracy of a classical finite-difference scheme generally can be increased only by using variables of more mesh points in each of its equations. Usually, a side effect of this practice is an increase in numerical diffusion, a subject to be discussed shortly. Note that, in the absence of body force, *direct* physical interactions occur only among the *immediate* neighbors. The current design is also consistent with this physical reality.

(c) Space and time traditionally are treated separately in the time marching schemes. Generally one obtains a system of ordinary differential equations with time being the independent variable after a spatial discretization. As an example, elements in the finite element method usually are used for spatial discretization. These elements are domains in space only.

Because flux conservation is fundamentally a property in *space-time*, space and time are unified and treated on the same footing in the current method. For example, conservation elements and solution elements used in the time-dependent version of the current method are domains in space-time. The significance of this unified approach cannot be overemphasized. As will be shown, it makes it easier for a numerical analogue to share the same space-time symmetry of the physical laws.

(d) In a finite-difference scheme, derivatives at mesh points are expressed in terms of mesh values of dependent variables by using finite-difference approximations. Accuracy of these approximations, especially those of higher-order accuracy, generally is excellent as long as dependent variables vary slowly across a mesh interval. However, it may not be adequate if these variables vary too rapidly. Thus, in a high-gradient region, e.g., a boundary layer, accuracy may demand the use of an extremely fine mesh. In turn, a prohibitively high computing cost may result.

The current method avoids the above pitfall by expressing the numerical solution within a solution element as an expansion in terms of certain base functions. As in a spectral method, the expansion coefficients are considered as *the independent numerical variables to be solved simultaneously*. For simplicity, Taylor's expansions will be

used in the current paper. For this special case, the expansion coefficients are interpreted as the numerical analogues of the derivatives. Note that (i) van Leer [12] also has attempted to improve accuracy by introducing two independent numerical variables for each independent physical variable, and (ii) the current solution procedure has no resemblance with those used in compact difference schemes.

(e) With a few exceptions, numerical diffusion generally appears in a numerical solution of a time-marching problem. In other words, the numerical solution dissipates faster than the corresponding physical solution. For a nearly inviscid problem, e.g., flow with a high Reynolds number, this could be very serious because numerical dissipation may overwhelm physical dissipation and cause a complete distortion of solutions. One may argue that numerical diffusion can be reduced by increasing the order of accuracy of the scheme used. However, because the order of accuracy of a scheme is generally determined with the aid of Taylor's expansion, and the latter is valid only for a smooth solution, it has meaning only for a smooth solution. Thus the use of a scheme of higher-order accuracy may not reduce numerical diffusion associated with high-frequency Fourier components of a numerical solution. This is the reason that the Leapfrog scheme, which is free from numerical diffusion, can outperform schemes with higher-order accuracy in solving some wave equations [13].

In a study of finite-difference analogues of a simple convection equation [14], it was shown that a numerical analogue will be free from numerical diffusion if it does not violate certain space-time invariant properties of the convection equation. In other words, numerical diffusion may be considered as a result of *symmetry-breaking* by the numerical scheme. Because of its intrinsic nature of space-time unity, the current framework is an excellent vehicle for constructing a numerical analogue that shares the same space-time invariant properties with the physical equation.

It is recognized that a certain amount of numerical diffusion may be needed to prevent large dispersive errors [15] that are often caused by the presence of high-frequency disturbances (such as round-off errors). Therefore, in the current paper we shall construct a model scheme for a simple convection equation in which its numerical diffusion is controlled by a single adjustable parameter. The numerical diffusion is shut off when this parameter is set to zero. Furthermore, an Euler solver will be constructed such that *its numerical diffusion at all mesh points can be controlled by a set of local parameters*.

(f) High-resolution upwind methods for solving the Euler equations [8], which we consider to be a branch of the finite volume method, are heavily dependent on characteristics-based techniques. For the 1-D time-dependent case, the characteristics are curves in space-time, and the coefficient matrix associated with the Euler equations [16] also can be diagonalized easily. As a result, these techniques are easy to apply. However, for multi-dimensional cases, the characteristics are 2-D or 3-D surfaces in space-time [17]. Moreover, the coefficient matrices cannot be diagonalized simultaneously by the same matrix [16]. Because of the above complexites, application of these techniques to multi-dimensional problems is much more diffucult. Furthermore, high-resolution

methods generally require the use of ad hoc parameters, e.g., flux-limiters and/or slope-limiters, and other ad hoc techniques. These ad hoc techniques generally are also difficult to extend to a space of higher dimension.

Because the current framework is developed to solve multi-dimensional problems, simplicity and generality weigh heavily in its design. Thus, we do not use characteristics-based techniques, and also try to avoid using ad hoc techniques. Moreover, the concept of characteristics generally is not applicable to the Navier-Stokes equations, which are non-hyperbolic in nature. Therefore, the above decision also makes it easier for the current framework to solve the Navier-Stokes equations.

This completes the discussion of the motivation for the current development. In summary, the development is guided by the following requirements: (i) To enforce both local and global flux consevation in space and time with flux evaluation at an interface being an integral part of the solution procedure and requiring no interpolation or extrapolation; (ii) To use local discrete variables such that the set of variables in any one of the numerical equations to be solved is associated with a set of immediately neighboring cells; (iii) Space and time are unified and treated on the same footing; (iv) Mesh values of dependent variables and their derivatives are considered as independent variables to be solved simultaneously; (v) To minimize numerical diffusion, a numerical analogue should be constructed, as much as possible, to be compatible with the space-time invariant properties of the corresponding physical equations; and (vi) To exclude the use of the characteristics-based techniques, and to avoid the use of ad hoc techniques as much as possible. It is the purpose of this paper to show that the above requirements can be met with a simple unified numerical framework.

For any reader who is interested in getting an advance idea on how simple the present method can be, he is referred to the computer program listed at the end of the present paper. It is a shock-tube-problem solver constructed using the present method. The simplicity of the solver is easily appreciated by a comparison of the listed program and a typical program associated with high-resolution upwind methods. *Not only is the listed program much smaller in size (it is self-contained and the main loop contains only* 33 *lines), but it contains no Fortran statements such as "if", "amax", and "amin" which are used so often in the programs implementing high-resolution methods.* The absence of the above Fortran statements in the listed program results from the effort in avoiding the use of the ad hoc techniques in the development of the present method. In spite of its simplicity, it will be shown in Sec. 8 that the present solver is capable of generating highly accurate shock tube solutions.

2. The a-μ Scheme

In this section, we consider a dimensionless form of the 1-D convection–diffusion equation, i.e.,

$$\frac{\partial u}{\partial t} + a\,\frac{\partial u}{\partial x} - \mu\,\frac{\partial^2 u}{\partial x^2} = 0 \tag{2.1}$$

where the convection speed a, and the viscosity coefficient μ (≥ 0) are constants. Let $x_1 = x$, and $x_2 = t$ be considered as the coordinates of a two-dimensional Euclidean space E_2. By using Gauss' divergence theorem in the space-time E_2, it can be shown that Eq. (2.1) is the differential form of the integral conservation law

$$\oint_{S(V)} \vec{h} \cdot d\vec{s} = 0. \tag{2.2}$$

As depicted in Fig. 1, here (i) $S(V)$ is the boundary of an arbitrary space-time region V in E_2, (ii) $\vec{h} = (au - \mu\partial u/\partial x,\ u)$ is a current density vector in E_2, and (iii) $d\vec{s} = d\sigma\,\vec{n}$ with $d\sigma$ and $\vec{n}$, respectively, being the area and the outward unit normal of a surface element on $S(V)$. Note that (i) $\vec{h} \cdot d\vec{s}$ is the *space-time* flux of $\vec{h}$ leaving the region V through the surface element $d\vec{s}$, and (ii) all mathematical operations can be carried out as though E_2 were an ordinary two-dimensional Euclidean space.

At this juncture, note that the conservation law given in Eq. (2.2) is formulated in a form in which space and time are unified and treated on the same footing. *This unity of space and time is also a tenet in the following numerical development. It is a key characteristic that distinguishes the current method from most of the traditional methods.*

Let Ω denote the set of mesh points (j,n) in E_2 (dots in Fig. 2(a)) where $n = 0, \pm 1/2, \pm 1, \pm 3/2, \pm 2, \pm 5/2, \ldots$, and, for each n, $j = n \pm 1/2, n \pm 3/2, n \pm 5/2, \ldots$. There is a solution element (SE) associated with each $(j,n) \in \Omega$. Let the solution element SE(j,n) be the *interior* of the *space-time* region bounded by a dashed curve depicted in Fig. 2(b). It includes a horizontal line segment, a vertical line segment, and their immediate neighborhood. For the following discussions, the exact size of this neighborhood does not matter.

For any $(x,t) \in$ SE(j,n), $u(x,t)$, and $\vec{h}(x,t)$, respectively, are approximated by $u^*(x,t\,;j,n)$ and $\vec{h}^*(x,t\,;j,n)$ which we shall define shortly. Let

$$u^*(x,t\,;j,n) = u_j^n + (u_x)_j^n (x - x_j) + (u_t)_j^n (t - t^n) \tag{2.3}$$

where (i) u_j^n, $(u_x)_j^n$, and $(u_t)_j^n$ are constants in SE(j,n), and (ii) (x_j, t^n) are the coordinates of the mesh point (j,n). Note that

$$u^*(x_j,t^n;j,n) = u_j^n, \qquad \frac{\partial u^*(x,t\,;j,n)}{\partial x} = (u_x)_j^n, \qquad \frac{\partial u^*(x,t\,;j,n)}{\partial t} = (u_t)_j^n. \tag{2.4}$$

Moreover, if we identify u_j^n, $(u_x)_j^n$, and $(u_t)_j^n$, respectively, with the values of u, $\partial u/\partial x$, and $\partial u/\partial t$ at (x_j, t^n), the expression on the right side of Eq. (2.3) becomes the first-order

Taylor's expansion of $u(x,t)$ at (x_j,t^n). As a result of these considerations, u_j^n, $(u_x)_j^n$, and $(u_t)_j^n$ will be considered as the numerical analogues of the values of u, $\partial u/\partial x$, and $\partial u/\partial t$ at (x_j,t^n), respectively.

We shall require that $u = u^*(x,t;j,n)$ satisfy Eq. (2.1) within SE(j,n). As a result of Eq. (2.4), this implies that

$$(u_t)_j^n = -a(u_x)_j^n. \tag{2.5}$$

Because Eq. (2.3) is a first-order Taylor's expansion, the diffusion term in Eq. (2.1) has no counterpart in Eq. (2.5). As a result, the diffusion term has no impact on how $u^*(x,t;j,n)$ varies with time *within SE*(j,n). However, as will be shown shortly, through its role in the numerical analogue of Eq. (2.2), it does influence time-dependence of numerical solutions. Note that, for a higher-order scheme, how $u^*(x,t;j,n)$ varies with time within SE(j,n) will be influenced by the presence of the diffusion term. Combining Eqs. (2.3) and (2.5), one has

$$u^*(x,t;j,n) = u_j^n + (u_x)_j^n\left[(x-x_j) - a(t-t^n)\right], \quad (x,t) \in \text{SE}(j,n). \tag{2.6}$$

Because $\vec{h} = (au - \mu\partial u/\partial x, u)$, we define

$$\vec{h}^*(x,t;j,n) = \left(au^*(x,t;j,n) - \mu\partial u^*(x,t;j,n)/\partial x,\ u^*(x,t;j,n)\right). \tag{2.7}$$

Let E_2 be divided into nonoverlapping rectangular regions (see Fig. 2(a)) referred to as conservation elements (CE's). As depicted in Figs. 2(c) and 2(d), a CE with its top-right (top-left) vertex being the mesh point $(j,n) \in \Omega$ is denoted by CE$_-(j,n)$ (CE$_+(j,n)$). Obviously the boundary of CE$_-(j,n)$ (CE$_+(j,n)$), excluding two isolated points B and C (C and D), is formed by the subsets of SE(j,n) and SE$(j-1/2,n-1/2)$ (SE$(j+1/2,n-1/2)$). The current approximation of Eq. (2.2) is

$$F_\pm(j,n) \stackrel{\text{def}}{=} \oint_{S(CE_\pm(j,n))} \vec{h}^* \cdot d\vec{s} = 0 \tag{2.8}$$

for all $(j,n) \in \Omega$. In other words, the total flux leaving the boundary of any conservation element is zero. Note that the flux at any interface separating two neighboring CE's is calculated using the information from a single SE. As an example, the interface AC depicted in Figs. 2(c) and 2(d) is a subset of SE(j,n). Thus the flux at this interface is calculated using the information associated with SE(j,n). Also note that an SE is the *interior* of a space-time region. Thus the vertices B, C, and D, strictly speaking, do not belong to any SE. As a result, $\vec{h}^*$ is not defined at these points. However, contributions to the above integral from these isolated points are zero no matter what values of $\vec{h}^*$ are assigned to them. For this reason, one may simply exclude them from the above surface integration.

Because the surface integration across any interface separating two neighboring CE's is evaluated using the information from a single SE, obviously the local conservation condition

Eq. (2.8) will lead to a global conservation relation, i.e., *the total flux leaving the boundary of any space-time region that is the union of any combination of CE's will also vanish.*

Because each $S(CE_{\pm}(j,n))$ is a simple closed curve in E_2 (see Fig. 1), the surface integration in Eq. (2.8) can be converted into a line integration. Let

$$\vec{g}^* \stackrel{\text{def}}{=} (-u^*, au^* - \mu \partial u^*/\partial x), \quad \text{and} \quad d\vec{r} \stackrel{\text{def}}{=} (dx, dt) \tag{2.9}$$

Thus, $d\vec{r}$ is normal to $d\vec{s}$ and points in the tangential direction of the line segment joining the two points (x,t) and $(x+dx, t+dt)$. Because $d\vec{s} = \pm(dt, -dx)$ [1, p.14], we have

$$\vec{h}^* \cdot d\vec{s} = \pm \vec{g}^* \cdot d\vec{r} \tag{2.10}$$

where the upper (lower) sign should be chosen if the 90° rotation from $d\vec{s}$ to $d\vec{r}$ is in the counterclockwise (clockwise) direction. By combining Eqs. (2.8) and (2.10), one concludes that

$$F_{\pm}(j,n) = \oint_{S(CE_{\pm}(j,n))}^{c.c.} \vec{g}^* \cdot d\vec{r} \tag{2.11}$$

Note that the notation *c.c.* indicates that the line integration should be carried out in the counterclockwise direction. Substituting Eq. (2.6) into Eq. (2.11), and using the fact that the boundary of a CE is formed by the subsets of two SE's, one has

$$\begin{aligned} \frac{4}{(\Delta x)^2} F_{\pm}(j,n) = \pm(1/2) \left[(1-\nu^2+\xi)(u_x)_j^n + (1-\nu^2-\xi)(u_x)_{j\pm 1/2}^{n-1/2} \right] \\ + \frac{2(1 \mp \nu)}{\Delta x} \left(u_j^n - u_{j\pm 1/2}^{n-1/2} \right) \end{aligned} \tag{2.12}$$

where

$$\nu \stackrel{\text{def}}{=} \frac{a\Delta t}{\Delta x}, \qquad \text{and} \qquad \xi \stackrel{\text{def}}{=} \frac{4\mu \Delta t}{(\Delta x)^2}. \tag{2.13}$$

Note that (i) the parameter ν is the Courant number, and (ii) a more efficient method of flux evaluation will be presented later in this section.

With the aid of Eqs. (2.8) and (2.12), u_j^n and $(u_x)_j^n$ can be solved in terms of $u_{j\pm 1/2}^{n-1/2}$ and $(u_x)_{j\pm 1/2}^{n-1/2}$ if $1-\nu^2+\xi \neq 0$, i.e., for all SE(j,n),

$$\vec{q}(j,n) = Q_+ \, \vec{q}(j-1/2, n-1/2) + Q_- \, \vec{q}(j+1/2, n-1/2) \quad (1-\nu^2+\xi \neq 0). \tag{2.14}$$

Here

$$\vec{q}(j,n) \stackrel{\text{def}}{=} \begin{pmatrix} u_j^n \\ \\ (\Delta x/4)(u_x)_j^n \end{pmatrix} \tag{2.15}$$

for all $(j,n) \in \Omega$, and

$$Q_+ \stackrel{\text{def}}{=} (1/2)\begin{pmatrix} 1+\nu & 1-\nu^2-\xi \\ \dfrac{-(1-\nu^2)}{1-\nu^2+\xi} & \dfrac{-(1-\nu)(1-\nu^2-\xi)}{1-\nu^2+\xi} \end{pmatrix}, \tag{2.16}$$

and

$$Q_- \stackrel{\text{def}}{=} (1/2)\begin{pmatrix} 1-\nu & -(1-\nu^2-\xi) \\ \dfrac{1-\nu^2}{1-\nu^2+\xi} & \dfrac{-(1+\nu)(1-\nu^2-\xi)}{1-\nu^2+\xi} \end{pmatrix}. \tag{2.17}$$

Because numerical variables at a higher time level can be evaluated in terms of those at a lower time level by using Eq. (2.14), it defines a marching scheme. Furthermore, because this scheme models Eq. (2.1) which is characterized by two parameters a and μ, hereafter it will be referred to as the a-μ scheme.

As a preliminary for future developments, we apply Eq. (2.14) successively and obtain

$$\vec{q}(j,\, n+1) = (Q_+)^2\vec{q}(j-1,\, n) + (Q_+Q_- + Q_-Q_+)\vec{q}(j,n) + (Q_-)^2\vec{q}(j+1,\, n) \qquad (1-\nu^2+\xi \neq 0). \tag{2.18}$$

A result of Eq. (2.18) is

$$\vec{q}(j,\, n+1) \to \vec{q}(j,n) \quad \text{as} \quad \Delta t \to 0, \tag{2.19}$$

if a, μ, and Δx are held constant. The proof follows from the fact that

$$(Q_+)^2 \to 0, \quad (Q_+Q_- + Q_-Q_+) \to 1 \quad \text{and} \quad (Q_-)^2 \to 0, \quad \text{as} \quad \Delta t \to 0, \tag{2.20}$$

if a, μ, and Δx are held constant.

Alternatively, Eq. (2.19) can be proved using the fact that the total flux of $\vec{h}^*$ leaving the boundary of any space-time region that is the union of any combination of CE's vanishes. Consider the union of $\text{CE}_+(j,n+1)$ and $\text{CE}_-(j+1/2,n+1/2)$ (see Fig. 2). This union is a rectangle with the vertices $(j+1/2,n+1)$, $(j,n+1)$, (j,n), and $(j+1/2,n)$. The flux leaving this rectangle through its two vertical edges approaches zero as $\Delta t \to 0$. Because the total flux leaving its boundary vanishes, one concludes that the total flux leaving its two horizontal edges also approaches zero as $\Delta t \to 0$. In other words, the flux entering the rectangle through the lower horizontal edge approaches that leaving through the upper horizontal edge as $\Delta t \to 0$. Because these two fluxes are evaluated using $\vec{q}(j,n)$ and $\vec{q}(j,\, n+1)$, respectively, the above limiting condition implies a limiting relation between $\vec{q}(j,n)$ and $\vec{q}(j,\, n+1)$. Similarly, by considering the union of $\text{CE}_-(j,n+1)$ and $\text{CE}_+(j-1/2,n+1/2)$, one obtains another limiting relation for $\vec{q}(j,n)$ and $\vec{q}(j,\, n+1)$. Eq. (2.19) is a result of the above two limiting relations.

The a-μ scheme has several nontraditional features. They are summarized in the following remarks:

(a) Space and time are unified and treated on the same footing in the construction of the a-μ scheme.

(b) The expansion coefficients u_j^n and $(u_x)_j^n$ in Eq. (2.6) are treated as independent variables, i.e., $(u_x)_j^n$ is not expressed in terms of u_j^n's by using a finite-difference approximation.

(c) As a result of Eq. (2.12), each of the conservation conditions $F_\pm(j,n) = 0$ involves only numerical variables associated with two neighboring SE's. This fact remains true for a scheme of higher-order accuracy in which Eq. (2.3) is replaced by a Taylor's expansion of higher-order. The contrast with the finite difference method and its physical significance were dicussed in Sec. 1.

(d) The a-μ scheme has the simplest stencil, i.e., a triangle with a vertex at the upper time level and the other two vertices at the lower time level. Eq. (2.14), which relates numerical variables at these vertices, was derived using the flux conservation conditions $F_\pm(j,n) = 0$. Because the flux at an interface separating two neighboring CE's is evaluated using information of a single SE, no interpolation or extrapolation is required. Moreover, accuracy of flux evaluation is enhanced by requiring that $u = u^*(x,t;j,n)$ satisfy Eq. (2.1) within SE(j,n). This makes the use of characteristics-based techniques less necessary.

(e) The a-μ scheme uses a mesh that is staggered in time. As will be explained in Appendix A, for a two-level scheme using such a mesh, e.g., the Lax scheme [4, p.97], generally the numerical variable at $(j,n+1)$ does not approach that at (j,n) as $\Delta t \to 0$, if a, μ, and Δx are held constant. This is a key reason why the Lax scheme is very diffusive when the Courant number ν is small. According to Eq. (2.19), the a-μ scheme is an exception to the above general rule.

(f) Eq. (2.1) can be solved numerically using the Leapfrog/DuFort-Frankel scheme [4, p.161]. This scheme is reduced to the Leapfrog scheme [4, p.100] if diffusion is absent (i.e., $\mu = 0$), and to the DuFort-Frankel scheme [4, p.114] if convection is absent (i.e., $a = 0$). It is well known that a solution of any of the above schemes is formed by two decoupled solutions with each being associated with a mesh that is also staggered in time. Traditionally the von Neumann stability analysis for the above schemes is performed without taking into account this decoupled nature [4]. In Appendix A, it is performed for each decoupled solution using the mesh depicted in Fig. 2(a). It is shown that the amplification factors of the Leapfrog/DuFort-Frankel scheme are

$$A_\pm = \left[\frac{\xi \cos(\theta/2) - i\nu \sin(\theta/2) \pm \sqrt{[\xi \cos(\theta/2) - i\nu \sin(\theta/2)]^2 + 1 - \xi^2}}{1+\xi} \right]^2 \qquad (2.21)$$

Here the amplification factors are defined to be those between the time levels n and $n+1$, i.e., they are the amplification factors of the solution after two marching steps.

The reason behind this definition is that the mesh points at the time levels n and $n+1$ are not staggered. Hereafter the same definition will be used for other schemes. Let $1-\nu^2 \neq 0$. Then the amplification factors $G_{\pm}^{(1)}$ of the current a-μ scheme (see Eq. (6.9)) are identical to those given by Eq. (2.21) except that the parameter ξ should be replaced by $\hat{\xi} \stackrel{\text{def}}{=} \xi/(1-\nu^2)$. Because (i) $\hat{\xi} = \xi = 0$ if $\mu = 0$, and (ii) $\nu = 0$ and thus $\hat{\xi} = \xi$, if $a = 0$, one concludes that $G_{\pm}^{(1)}$ are completely identical to those of the Leapfrog scheme if $\mu = 0$, and to those of the DuFort-Frankel scheme if $a = 0$. These coincidences are unexpected because the a-μ scheme and the above classical schemes are derived from completely different perspectives. Moreover, the a-μ scheme is a two-level scheme with two variables u_j^n and $(u_x)_j^n$ associated with the mesh point (j, n), while the above classical schemes are three-level schemes with a single variable u_j^n associated with the same point.

Because the amplification factors of the inviscid a-μ scheme (i.e., the a-μ scheme with $\mu = 0$) are identical to those of the Leapfrog scheme, the former, as in the case of the latter, is neutrally stable (i.e., free of numerical diffusion) if $\nu^2 < 1$. Note that the case with $\mu = 0$ and $\nu^2 = 1$ is ruled out by the assumption $1 - \nu^2 + \xi \neq 0$ of Eq. (2.14). Similarly, the pure-diffusion a-μ scheme (i.e., the a-μ scheme with $a = 0$), as in the case of the DuFort-Frankel scheme, is unconditionally stable. Furthermore, it is proved in Sec. 6 that the stability of the general a-μ scheme, as in the case of the Leapfrog/DuFort-Frankel scheme, *is independent of μ, and restricted only by the CFL condition, i.e.,* $\nu^2 \leq 1$. The a-μ scheme is the only *two-level explicit* scheme known to the author to possesss the above properties. Also it will be shown later that the same stability condition is retained by a natural 1–D time-dependent Navier-Stokes extension of the a-μ scheme.

Because stability of the a-μ scheme is restricted only by the *CFL* condition, the stability bound for Δt is proportional to Δx. In contrast, the stability condition of a typical classical explicit scheme generally is more restrictive than the *CFL* condition. For a small mesh Reynolds number, the stability bound for Δt is approximately proportional to $(\Delta x)^2$ for the MacCormack scheme [4, p.102].

Because a neutrally stable numerical analogue of the pure convection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{2.22}$$

usually becomes unstable when it is applied to a nonlinear inviscid generalization of Eq. (2.22), the inviscid a-μ scheme will be modified in Sec. 3 such that it can be extended to model the Euler equations. In this new version, numerical diffusion is introduced in a way that allows its magnitude to be adjusted by a special parameter.

(g) The conservation relations for $\mathrm{CE}_+(j-1/2,\, n+1/2)$ and $\mathrm{CE}_-(j+1/2,\, n+1/2)$ (see Figs. 2(e) and 2(f)) are

$$F_+(j-1/2,\, n+1/2) = 0, \quad \text{and} \quad F_-(j+1/2,\, n+1/2) = 0, \tag{2.23}$$

respectively. Combining Eqs. (2.12) and (2.23), and assuming $1-\nu^2-\xi \neq 0$, one has

$$\vec{q}(j,n) = \hat{Q}_+ \vec{q}(j+1/2,\, n+1/2) + \hat{Q}_- \vec{q}(j-1/2,\, n+1/2) \quad (1-\nu^2-\xi \neq 0). \tag{2.24}$$

Here

$$\hat{Q}_+ \stackrel{\text{def}}{=} (1/2) \begin{pmatrix} 1+\nu & -(1-\nu^2+\xi) \\ \dfrac{1-\nu^2}{1-\nu^2-\xi} & \dfrac{-(1-\nu)(1-\nu^2+\xi)}{1-\nu^2-\xi} \end{pmatrix}, \tag{2.25}$$

and

$$\hat{Q}_- \stackrel{\text{def}}{=} (1/2) \begin{pmatrix} 1-\nu & 1-\nu^2+\xi \\ \dfrac{-(1-\nu^2)}{1-\nu^2-\xi} & \dfrac{-(1+\nu)(1-\nu^2+\xi)}{1-\nu^2-\xi} \end{pmatrix}. \tag{2.26}$$

Eq. (2.24) defines a backward marching scheme, i.e., the numerical variables at the time level n are determined in terms of those at the time level $(n+1/2)$. Recall that both the forward marching scheme Eq. (2.14) and the backward marching scheme Eq. (2.24) are derived using the same set of conservation relations. As a matter of fact, Eqs. (2.14) and (2.24) are equvalent if $(1-\nu^2)^2 \neq (\xi)^2$ is assumed. For the above reason, the a-μ scheme may be referred to as a *two-way marching* scheme. For the case $\mu > 0$, it will be proved in Sec. 6 that the a-μ scheme cannot be stable for both the forward and the backward marching directions except for the singular case $\nu^2 = 1$ which is also on the threshold of instability. Thus, for all practical purposes the viscous a-μ scheme is irreversible in time. On the other hand, it is neutrally stable for both the forward and backward marching directions, and thus is reversible in time, if $\mu = 0$, and $\nu^2 < 1$. Again, the a-μ scheme is the only two-level explicit two-way marching scheme known to the author.

(h) Several invariant properties of Eq. (2.1) with respect to space and time are discussed in [14]. In the same paper, these properties are also defined for the numerical analogues of Eq. (2.1). It is also shown that the neutral stability of several finite-difference analogues of Eq. (2.22) can be established by using their invariant properties with respect to space-time inversion. Because solutions of Eq. (2.22) do not dissipate with time, it is not surprising that solutions of a numerical analogue also will not dissipate with time, i.e., the scheme is neutrally stable, if it shares with Eq. (2.22) some space-time invariant properties. It will be shown in a future paper that the a-μ scheme shares with Eq. (2.1) the same space-time invariant properties. Also note that these invariant properties are closely linked with the other properties discussed in (a), (e), (f), and (g).

This completes the discussion on nontraditional features of the a-μ scheme. In the following, it will be shown that this scheme can also be constructed from a completely different perspective. As a part of this construction, SE's and CE's of different types will be used and discussed.

In the new construction, the locations of mesh points (dots in Fig. 3(a)) are identical to those used in the original construction. However, $\mathrm{SE}(j,n)$ is defined to be the *interior* of a rhombus centered at (j,n) (see Fig. 3(b)). $\mathrm{CE}(j,n)$ is the union of $\mathrm{SE}(j,n)$ and its boundary. Readers are warned *not* to confuse the sides of the rhombus with the characteristics of Eq. (2.22). Any one of these sides is simply a line segment joining two points of intersection (not marked by dots) of horizontal and vertical mesh lines. For any $(x,t) \in \mathrm{SE}(j,n)$, $u(x,t)$ and $\vec{h}(x,t)$, respectively, again are approximated by $u^*(x,t;j,n)$ and $\vec{h}^*(x,t;j,n)$ which are defined by Eqs. (2.3) and (2.7), respectively. However, Eq. (2.5) will be derived from a consideration of flux conservation.

Let Eq. (2.2) be approximated by

$$\oint_{S(V^*)} \vec{h}^* \cdot d\vec{s} = 0 \tag{2.27}$$

where V^* is *the union of any combination of CE's*. Because an SE is the interior of a CE, $\vec{h}^*$ is not defined on $S(V^*)$, the boundary of V^*. As a result, the above surface integration is to be carried out over a surface that is in the interior of V^* and immediately adjacent to $S(V^*)$. A necessary condition of Eq. (2.27) is that, for all $(j,n) \in \Omega$,

$$\oint_{S(CE(j,n))} \vec{h}^* \cdot d\vec{s} = 0, \tag{2.28}$$

i.e., the total flux leaving any conservation element is zero.

Note that the center of a current SE no longer sits on an interface separating two CE's. It coincides with the center of a CE. Thus $\vec{h}^*$ at one side of an interface is evaluated using information from one SE, while that at the other side is evaluated using information from another SE. As an example, $\vec{h}^*$ at BC and B'C' depicted in Fig. 3(d), respectively, are evaluated using information from $\mathrm{SE}(j,n)$ and $\mathrm{SE}(j-1/2,n-1/2)$. Another necessary condition for Eq. (2.27) *is the equality between the fluxes entering and leaving any interface*. This can be seen by applying Eq. (2.27) separately to two neighboring CE's, and then to their union. Obviously the local flux conservation relations at all interfacs, and within all CE's (i.e., Eq. (2.28)) are equivalent to the global conservation relation Eq. (2.27). The equations representing the above conservation conditions are the numerical equations to be solved. Note that, in the current construction, a flux is not preassigned at an interface using an interpolation or extrapolation of information from both sides of this interface. The current method of interface flux evaluation obviously is different from that used in the finite volume method which was disscussed in Sec. 1.

By using Eqs. (2.3) and (2.7), one concludes that, for any $(x,t) \in \mathrm{SE}(j,n)$, the divergence of $\vec{h}^*$ in E_2 is

$$\begin{aligned} \nabla \cdot \vec{h}^* &\stackrel{\text{def}}{=} \frac{\partial [au^*(x,t;j,n) - \mu \partial u^*(x,t;j,n)/\partial x]}{\partial x} + \frac{\partial u^*(x,t;j,n)}{\partial t} \\ &= a(u_x)_j^n + (u_t)_j^n. \end{aligned} \tag{2.29}$$

Because $(u_x)_j^n$ and $(u_t)_j^n$ are constants within an SE, Eq. (2.29) implies that $\nabla \cdot \vec{h}^*$ is also a constant. Thus Eq. (2.28) coupled with Gauss' divergence theorem implies that, within any SE,

$$\nabla \cdot \vec{h}^* = 0. \tag{2.30}$$

Eq. (2.5) is a direct result of Eqs. (2.29) and (2.30).

Note that Eq. (2.30) follows from Eq. (2.28) because $u^*(x,t\,;j,n)$ defined in Eq. (2.3) is a first-order Taylor's expansion. For a higher-order expansion, the condition that Eq. (2.30) being valid uniformly within an SE is stronger than Eq. (2.28). For the general case, the stronger condition should be imposed. Because Eq. (2.30) is the numerical analogue of Eq. (2.1), the imposition of the stronger condition ensures that, *within an SE, the numerical solution uniformly satisfies the differential form of the conservation law Eq. (2.2).*

With the aid of Gauss' divergence theorem, Eq. (2.30) implies that the surface integration of $\vec{h}^*$ over any closed surface located within any SE vanishes. As a result,

$$\oint_{S(\triangle ABC)} \vec{h}^* \cdot d\vec{s} = 0, \quad \text{and} \quad \oint_{S(\triangle A'B'C')} \vec{h}^* \cdot d\vec{s} = 0, \tag{2.31}$$

where the triangles $\triangle ABC$ and $\triangle A'B'C'$ are those depicted in Fig. 3(d). Because the net flux of $\vec{h}^*$ entering an interface from both sides vanishes, the sum of the flux leaving $\mathrm{CE}(j,n)$ through BC and that leaving $\mathrm{CE}(j-1/2,n-1/2)$ through $B'C'$ vanishes. Thus, Eq. (2.31) implies that $F_-(j,n) = 0$ where $F_-(j,n)$ is defined in Eq. (2.11). Similarly, it can be shown that $F_+(j,n) = 0$.

Assuming Eqs. (2.3) and (2.7), it has been shown that both Eqs. (2.5) and (2.8) can be derived using Eq. (2.27). Conversely, Eq. (2.27) also follows from Eqs. (2.5) and (2.8). Obviously both the forward marching scheme Eq. (2.14) and the backward marching scheme Eq. (2.22) can also be obtained by assuming Eqs. (2.3), (2.7), and (2.27).

Note that the equivalence between Eq. (2.27) and the pair of equations Eqs. (2.5) and (2.8) hinges on the fact that $\nabla \cdot \vec{h}^* = 0$ within an SE of either type I or type II. As will be shown immediately, this condition can be used to simplify evaluation of the flux across a simple curve that *lies entirely within an SE of either type.*

According to the top expression given in Eq. (2.29), $\nabla \cdot \vec{h}^* = 0$ implies that there exists a function $\psi(x,t\,;j,n)$ such that

$$\frac{\partial \psi(x,t\,;j,n)}{\partial t} = a u^*(x,t\,;j,n) - \mu \frac{\partial u^*(x,t\,;j,n)}{\partial x}, \tag{2.32}$$

and

$$-\frac{\partial \psi(x,t\,;j,n)}{\partial x} = u^*(x,t\,;j,n) \tag{2.33}$$

for any $(x,t) \in \mathrm{SE}(j,n)$. Substituting Eq. (2.6) into Eqs. (2.32) and (2.33), one concludes that, up to an arbitrary constant,

$$\begin{aligned} \psi(x,t\,;j,n) = -\frac{(u_x)_j^n}{2}\Big\{ \big[(x-x_j) - a(t-t^n)\big]^2 + 2\mu(t-t^n) \Big\} \\ - u_j^n \big[(x-x_j) - a(t-t^n)\big]. \end{aligned} \tag{2.34}$$

Moreover, with the aid of Eq. (2.9), Eqs. (2.32) and (2.33) imply that

$$\vec{g}^* \cdot d\vec{r} = d\psi. \tag{2.35}$$

Let $(x,t) \in \mathrm{SE}(j,n)$ and $(x',t') \in \mathrm{SE}(j,n)$. Let Γ be a simple curve joining (x,t) and (x',t'), and *lying entirely within SE*(j,n) (see Fig. 4). Then Eqs. (2.10) and (2.35) imply that

$$\int_\Gamma \vec{h}^* \cdot d\vec{s} = \psi(x',t';j,n) - \psi(x,t;j,n). \tag{2.36}$$

Here we assume that $d\vec{s}$ points to the right of Γ if one moves forward from (x,t) to (x',t') (see Fig. 4). Eq. (2.36) states that the flux of $\vec{h}^*$ across the curve Γ is given by the difference in the values of ψ at its two end-points. For this reason, $\psi(x,t;j,n)$ will be referred to as the potential function associated with $\mathrm{SE}(j,n)$. Obviously, Eq. (2.12) can be obtained using Eq. (2.36).

In [1], the a-μ scheme is subjected to a thorough theoretical and numerical analysis on stability, dissipation, dispersion, consistency, truncation error, and accuracy. It is shown that it has many advantages over the MacCormack and the Leapfrog/DuFort-Frankel schemes. Particularly, by using a discrete Fourier analysis, it is shown that the a-μ scheme is more accurate than the Leapfrog/DuFort-Frankel scheme by one order in both initial-value specification and the marching scheme itself.

In conclusion, a model scheme has been constructed from two different perspectives using SE's and CE's of different types. Using either perspective, one can say that a *numerical solution generated using the current framework satisfies (i) the differential form of the conservation law uniformly within an SE, and (ii) the integral form over any region that is the union of any combination of CE's.* In the author's opinion, the second perspective that uses the SE's and CE's of type II depicted in Fig. 3 is more fundamental and thus was used in the initial development of the current method [1]. However, we believe that the first perspective is easier to use in constructing explicit schemes. Because most schemes to be constructed in the current paper are explicit, the first perspective will be adopted unless specified otherwise.

3. The a-ϵ Scheme

The inviscid a-μ scheme is neutrally stable and reversible in time. It is well known that a neutrally stable numerical analogue of Eq. (2.22) generally becomes unstable when it is extended to model the Euler equations. It is also obvious that a scheme that is reversible in time cannot model a physical problem that is irreversible in time, e.g., an inviscid flow problem involving shocks. In this section, we assume $\mu = 0$ and attempt to modify the inviscid a-μ scheme such that it can be extended to model the Euler equations.

The current path of development is almost identical to that given in Sec. 2. We continue to assume Eqs. (2.3)–(2.7), and use SE's of type I depicted in Fig. 2. In addition to $\mu = 0$, the *only* other modification is the replacement of the assumption $F_{\pm}(j,n) = 0$ by

$$F_{\pm}(j,n) = \pm\frac{\epsilon(1-\nu^2)(\Delta x)^2}{4}(du_x)_j^n, \tag{3.1}$$

where ϵ is a parameter independent of numerical variables, and

$$(du_x)_j^n \stackrel{\text{def}}{=} (1/2)\left[(u_x)_{j+1/2}^{n-1/2} + (u_x)_{j-1/2}^{n-1/2}\right] - \left(u_{j+1/2}^{n-1/2} - u_{j-1/2}^{n-1/2}\right)\Big/\Delta x. \tag{3.2}$$

In other words, we add two terms of the same magnitude but with opposite signs, respectively, to the right sides of the original conservation conditions $F_+(j,n) = 0$ and $F_-(j,n) = 0$. The beauty of this modification will be fully explained later in this section. For now it suffices to say that this modification injects a higher-order *finite-difference* error into the inviscid a-μ scheme. It breaks the space-time symmetry of the latter. In turn, numerical diffusion is introduced as a result of this symmetry breaking. Because the magnitude of the terms added in this modification is controlled by ϵ, numerical diffusion is controlled by ϵ in the modified scheme just as physical diffusion is controlled by μ in the a-μ scheme. Note that, as a result of Eq. (3.1) and the assumption $\mu = 0$, the modified scheme is characterized by two parameters a and ϵ. Thus, hereafter it will be referred to as the a-ϵ scheme. Also note that, because there is no upwind bias in the a-ϵ scheme, *upwind bias is not the source of numerical diffusion.* Additional remarks on Eqs. (3.1) and (3.2) are:

(a) By definition, $F_+(j,n)$ and $F_-(j,n)$ represent total fluxes leaving $\text{CE}_+(j,n)$ and $\text{CE}_-(j,n)$, respectively (see Figs. 2(c) and 2(d)). Because $F_{\pm}(j,n) \neq 0$ if $\epsilon \neq 0$, $\text{CE}_+(j,n)$ and $\text{CE}_-(j,n)$ generally are no longer conservation elements in the a-ϵ scheme.

(b) Let $\text{CE}(j,n)$ be the union of $\text{CE}_+(j,n)$ and $\text{CE}_-(j,n)$ (see Fig. 5(b)). Note that *this definition of $CE(j,n)$ differs from that given in Sec. 2 and depicted in Fig. 3(c).* Let

$$F(j,n) \stackrel{\text{def}}{=} \oint_{S(CE(j,n))} \vec{h}^* \cdot d\vec{s}. \tag{3.3}$$

Because the net flux entering the interface separating $\text{CE}_+(j,n)$ and $\text{CE}_-(j,n)$ is zero. $F(j,n)$ is the sum of $F_+(j,n)$ and $F_-(j,n)$. With the aid of Eq. (3.1), we have

$$F(j,n) = F_+(j,n) + F_-(j,n) = 0, \tag{3.4}$$

i.e., the total flux leaving CE(j,n) vanishes. As a result, CE(j,n) *is a conservation element in the a-ϵ scheme* . Note that Eq. (3.4) leads to a global conservation relation in the form of Eq. (2.27) where V^* is the union of any combination of these new CE's.

(c) Because $\xi = 0$ if $\mu = 0$, Eq. (3.4) coupled with Eq. (2.12) implies that

$$u_j^n = (1/2)\left[(1+\nu)u_{j-1/2}^{n-1/2} + (1-\nu)u_{j+1/2}^{n-1/2}\right] + \frac{\Delta x(1-\nu^2)}{8}\left[(u_x)_{j-1/2}^{n-1/2} - (u_x)_{j+1/2}^{n-1/2}\right] \tag{3.5}$$

Thus, u_j^n is independent of ϵ.

(d) Because $(u_x)_{j\pm1/2}^{n-1/2}$ is a numerical analogue of $\partial u/\partial x$ at $(j \pm 1/2, n-1/2)$, the simple average

$$(1/2)\left[(u_x)_{j+1/2}^{n-1/2} + (u_x)_{j-1/2}^{n-1/2}\right]$$

is a numerical analogue of $\partial u/\partial x$ at $(j, n-1/2)$, the midpoint of a line segment joining $(j+1/2, n-1/2)$ and $(j-1/2, n-1/2)$ (see Fig. 2(a)). Note that $(j, n-1/2) \notin \Omega$ if $(j,n) \in \Omega$. Also note that

$$\left(u_{j+1/2}^{n-1/2} - u_{j-1/2}^{n-1/2}\right) \Big/ \Delta x$$

is a central-difference analogue of $\partial u/\partial x$ at $(j, n-1/2)$. Thus, $(du_x)_j^n$ *represents the difference of two numerical analogues of* $\partial u/\partial x$ *at the same mesh point* $(j, n-1/2)$. By using Taylor's expansion at $(j, n-1/2)$, it can be shown that $(du_x)_j^n = O\left[(\Delta x)^2\right]$, if $(u_x)_{j\pm1/2}^{n-1/2}$ are identified with $\partial u(x_{j\pm1/2}, t^{n-1/2})/\partial x$, respectively. Hereafter a quantity is denoted by $O\left[(\Delta x)^\ell\right]$ if there exists a constant $C > 0$ such that the absolute value of this quantity $\leq C\,|\Delta x|^\ell$ for all sufficiently small $|\Delta x|$. Note that we have constructed an expression of $O\left[(\Delta x)^2\right]$ without explicitly introducing the factor $(\Delta x)^2$. This natural construction leads to the simple stability conditions to be given in Eq. (3.14). It is possible only because there are two discrete variables u_j^n and $(u_x)_j^n$ associated with the mesh point (j,n).

(e) Eq. (3.1) could have been written as $F_\pm(j,n) = \pm\epsilon'(du_x)_j^n$ with $\epsilon' = \epsilon(1-\nu^2)(\Delta x)^2/4$. However, this *simplified* expression would lead to much more complicated equations later.

This completes the discussion of Eqs. (3.1) and (3.2). Now, let $1-\nu^2 \neq 0$. Then Eqs. (2.12), (3.1), and (3.2) can be used to obtain the current counterparts of Eqs. (2.14) and (2.18). They are

$$\vec{q}(j,n) = M_+\,\vec{q}(j-1/2,\, n-1/2) + M_-\,\vec{q}(j+1/2,\, n-1/2) \quad (1-\nu^2 \neq 0) \tag{3.6}$$

and

$$\vec{q}(j,\, n+1) = (M_+)^2\vec{q}(j-1,\, n) + (M_+M_- + M_-M_+)\vec{q}(j,n) + (M_-)^2\vec{q}(j+1,\, n) \quad (1-\nu^2 \neq 0), \tag{3.7}$$

respectively. Here

$$M_+ \stackrel{\text{def}}{=} (1/2)\begin{pmatrix} 1+\nu & 1-\nu^2 \\ \epsilon-1 & 2\epsilon-1+\nu \end{pmatrix} \tag{3.8}$$

and

$$M_- \stackrel{\text{def}}{=} (1/2)\begin{pmatrix} 1-\nu & -(1-\nu^2) \\ 1-\epsilon & 2\epsilon-1-\nu \end{pmatrix}. \tag{3.9}$$

Obviously, $M_\pm = Q_\pm$ if $\epsilon = 0$ and $\xi = 0$. Furthermore, the limiting condition given in Eq. (2.19) is still valid *if we assume that* $\epsilon = \epsilon(\Delta t)$ *and* $\lim_{\Delta t \to 0} \epsilon(\Delta t) = 0$. However, unlike the a-μ scheme, the a-ϵ scheme *is not a two-way marching scheme if* $\epsilon \neq 0$.

Eq. (3.6) represents a pair of equations. The first is Eq. (3.5). With the aid of Eqs. (2.5) and (2.13), the second equation can be expressed as

$$(u_x)_j^n = \left(u'^n_{j+1/2} - u'^n_{j-1/2}\right) \Big/ \Delta x + (2\epsilon - 1)(du_x)_j^n. \tag{3.10}$$

Here

$$u'^n_{j\pm1/2} \stackrel{\text{def}}{=} u_{j\pm1/2}^{n-1/2} + (\Delta t/2)(u_t)_{j\pm1/2}^{n-1/2}, \tag{3.11}$$

i.e., $u'^n_{j\pm1/2}$ is a first-order Taylor's approximation of u at $(j \pm 1/2, n)$. Thus, the expression on the right side of Eq. (3.10) is the sum of a central-difference approximation of $\partial u/\partial x$ at (j,n) and the extra term $(2\epsilon - 1)(du_x)_j^n$. Because $(du_x)_j^n = O\left[(\Delta x)^2\right]$, *the presence of this extra term will not lower the order of accuracy of the entire sum as an approximation of* $\partial u/\partial x$ *at* (j,n). Also note that *this extra term vanishes when* $\epsilon = 1/2$ *while the term associated with* $(du_x)_j^n$ *in Eq. (3.1) vanishes when* $\epsilon = 0$.

Next we shall study the influence of ϵ on the stability and numerical diffusion of the a-ϵ scheme. Let $G_+^{(2)}$ and $G_-^{(2)}$ be the principal and spurious amplification factors of the a-ϵ scheme, respectively. Then, it will be shown in Sec. 6 that

$$G_\pm^{(2)} = [\lambda_\pm(\epsilon, \nu, \theta)]^2, \tag{3.12}$$

with

$$\lambda_\pm(\epsilon,\nu,\theta) \stackrel{\text{def}}{=} \epsilon\cos(\theta/2) - i\nu\sin(\theta/2) \pm \sqrt{(1-\epsilon)\left[(1-\epsilon)\cos^2(\theta/2) + (1-\nu^2)\sin^2(\theta/2)\right]}. \tag{3.13}$$

Here θ, $-\pi < \theta \leq \pi$ [1, p.30], is the phase angle variation per Δx. Also it will be proved that

$$0 \leq \epsilon \leq 1, \qquad \text{and} \qquad \nu^2 < 1 \tag{3.14}$$

are necessary conditions for the stability of the a-ϵ scheme. Thus, Eq. (3.14) will be assumed in the remainder of this section.

It was pointed out in Sec. 2 that the amplification factors of the Leapfrog scheme are identical to those of the inviscid a-μ scheme. Because the latter scheme is a special case of the a-ϵ scheme with $\epsilon = 0$, $G_{\pm}^{(2)}$ become the amplification factors of the Leapfrog scheme when $\epsilon = 0$. This fact can be reverified by comparing Eqs. (2.21), (3.12) and (3.13) with $\xi = 0$ and $\epsilon = 0$.

Also, we have

$$\lambda_{\pm}(1, \nu, \theta) = \cos(\theta/2) - i\nu \sin(\theta/2). \tag{3.15}$$

Thus, $G_{+}^{(2)} = G_{-}^{(2)}$ when $\epsilon = 1$. Moreover, it is shown in Appendix A that *the coalesced amplification factor is identical to that of the Lax scheme.* Note that, like the Leapfrog scheme, a solution of the Lax scheme is also composed of two decoupled solutions with each being associated with a mesh that is staggered in time. However, because the Lax scheme is a two-level scheme, it does not have a spurious amplification factor.

Thus, at one extreme, i.e., when $\epsilon = 0$, $G_{\pm}^{(2)}$ become the amplification factors of the Leapfrog scheme, which *is free of numerical diffusion.* At another extreme, i.e., when $\epsilon = 1$, $G_{+}^{(2)}$ and $G_{-}^{(2)}$ coalesce into one and it becomes the amplification factor of the Lax scheme, which is notorious for its large diffusive errors. From the above observations, one may infer the following conclusion that will be established shortly, i.e., *the a-ϵ scheme becomes more diffusive as the value of ϵ increases.* Note that, *because the Lax scheme is very diffusive and uses a mesh that is staggered in time, a two-level scheme using such a mesh is usually associated with a highly diffusive scheme* [18]. The a-ϵ scheme demonstrates that it can also be a scheme with no diffusive error!

As a result of Eq. (3.14), the expression under the radical sign in Eq. (3.13) is non-negative. Thus, it can be shown that

$$1 - |G_{\pm}^{(2)}| = \chi_{\pm}(\epsilon, \nu, \theta) \stackrel{\text{def}}{=} \epsilon \Bigg\{ (1 - \nu^2)\sin^2(\theta/2) + 2\cos(\theta/2) \times \Big[(1 - \epsilon)\cos(\theta/2) \mp \sqrt{(1 - \epsilon)\left[(1 - \epsilon)\cos^2(\theta/2) + (1 - \nu^2)\sin^2(\theta/2)\right]} \Big] \Bigg\} \tag{3.16}$$

Because solutions to the physical equation Eq. (2.22) do not dissipate with time, a numerical analogue to Eq. (2.22) is said to be free of numerical diffusion if its solutions also do not dissipate with time, i.e., its amplification factors are of unit magnitude. As a result, numerical diffusion of the a-ϵ scheme may be measured by $1 - |G_{\pm}^{(2)}|$, i.e., $\chi_{\pm}(\epsilon, \nu, \theta)$. Obviously the a-ϵ scheme is free of numerical diffusion if $\epsilon = 0$. Also, by using Eqs. (3.14) and (3.16), it is shown in Sec. 6 that, for all θ with $-\pi < \theta \leq \pi$, and all ϵ and ν satisfying Eq. (3.14), we have

$$0 \leq \chi_{+}(\epsilon, \nu, \theta) + 4\epsilon(1 - \epsilon)\cos^2(\theta/2) \leq \chi_{-}(\epsilon, \nu, \theta) \leq \min\{1, 4\epsilon\}, \tag{3.17}$$

and

$$0 \le \chi_+(\epsilon, \nu, \theta) \le \epsilon(1 - \nu^2)\sin^2(\theta/2). \tag{3.18}$$

The significance of Eqs. (3.17) and (3.18) is discussed in the following remarks:

(a) Because $0 \le \chi_\pm(\epsilon, \nu, \theta)$, Eq. (3.16) implies that $|G_\pm^{(2)}| \le 1$. It is proved in Sec. 6 that this result and other considerations lead to the conclusion that Eq. (3.14) is also sufficient for stability.

(b) For a numerical analogue of Eq. (2.22) that has both principal and spurious amplification factors, a numerical solution with periodic boundary conditions is the sum of a principal solution and a spurious solution [1, p.37]. Only the principal solution contributes to the accuracy of the scheme. Given a smooth initial condition, the spurious solution at $t = 0$ generally is very small compared with the principal solution. Also, the behaviors of the principal and the spurious solutions as functions of time are determined by the principal and spurious amplification factors, respectively. Because $0 \le \epsilon(1 - \epsilon)$ if $0 \le \epsilon \le 1$, Eq. (3.17) implies that $\chi_+(\epsilon, \nu, \theta) \le \chi_-(\epsilon, \nu, \theta)$. Thus, the spurious solution will dissipate not slower than the principal solution. Let ϵ be not too close to 0 or 1. Then Eq. (3.17) also implies that the Fourier components of the spurious solution with smaller $|\theta|$, i.e., longer wavelength, will dissipate much faster than those of the principal solution. In other words, the spurious solution will rapidly disappear from the long-wavelength components of a numerical solution. Note that $\chi_-(1/2, \nu, 0) = 1$. Thus, *the long-wavelength components of the spurious solution are annihilated almost completely in a single time step if $\epsilon = 1/2$, i.e., if the last term in Eq. (3.10) is dropped.*

(c) The upper bound of $\chi_+(\epsilon, \nu, \theta)$ given in Eq. (3.18) is proportional to $\sin^2(\theta/2)$. As a result, the long-wavelength Fourier components in the principal solution are nearly free of numerical diffusion. On the other hand, short-wavelength components may decay rapidly.

(d) For a fixed ϵ, Eq. (3.18) implies that the principal solution is more diffusive for a smaller $|\nu|$. To compensate, one may choose a small ϵ for a small $|\nu|$. One may even choose ϵ to be a monotonic function of $|\nu|$, subjected to the condition $0 \le \epsilon \le 1$.

(e) Eqs. (3.17) and (3.18) imply that, for all ν with $\nu^2 < 1$ and all θ with $-\pi < \theta \le \pi$, we have

$$0 \le \chi_+(\epsilon, \nu, \theta) \le \epsilon, \qquad \text{and} \qquad 0 \le \chi_-(\epsilon, \nu, \theta) \le \min\{1, 4\epsilon\}, \tag{3.19}$$

which, according to Eq. (3.16), is equivalent to

$$1 - \epsilon \le |G_+^{(2)}| \le 1, \qquad \text{and} \qquad 1 - \min\{1, 4\epsilon\} \le |G_-^{(2)}| \le 1. \tag{3.20}$$

As a result, by choosing ϵ small enough, both $|G_+^{(2)}|$ and $|G_-^{(2)}|$ can be confined within an arbitrarily narrow range. As noted previously, the spurious part of a numerical

solution generally is insignificantly small assuming a smooth initial condition. It does not contribute to accuracy and usually dissipates faster than the principal part. Thus, our primary concerns is how the principal part dissipates. From Eq.(3.20), one concludes that, for any ϵ with $0 < \epsilon < 1$, $|G_+^{(2)}|$ will be bounded *uniformly* from below by a *positive* number $1-\epsilon$ for all ν with $\nu^2 < 1$ and all θ with $-\pi < \theta \leq \pi$. By choosing an ϵ of proper magnitude, *one can suppress artificial numerical oscillations without causing large diffusive errors for any combination of* ν *and* θ. This fact contrasts sharply with what one expects from typical classical schemes which are usually very diffusive with respect to certain ν and θ, while not at all with respect to other ν and θ. As an example, we consider the Lax-Wendroff scheme [4, p.101]. Its amplification factor is of unit magnitude, for all θ at $\nu = 0$, or $\nu = 1$. On the other hand, the amplification factor $= 0$ if $\nu^2 = 1/2$ and $\theta = \pi$.

In nonlinear flow solutions, e.g., shock-tube solutions to be discussed in Sec. 8, analogues of ν are dependent on local velocity components. Thus, they may vary from one location to another. Also, at some neighborhood, the Fourier spectrum of the local solution may have peaks spread over a wide range of θ. Thus, for a numerical analogue of Eq. (2.22), a large variation in numerical diffusivity with respect to θ and ν generally means that numerical solutions obtained using its nonlinear extensions will suffer annihilations of sharply different degrees at different locations and different θ. Such selective annihilations may cause large distortions of numerical solutions [19].

This completes the discussion of Eqs. (3.17) and (3.18). In conclusion, the a-ϵ scheme has been constructed to solve Eq. (2.22). It has the unique property that numerical diffusion can be controlled by a parameter ϵ. *Because neither characteristics-based techniques nor knowledge about the upwind direction is used in the construction of the* a-ϵ *scheme, as will be shown in the next section, it can be easily extended to model the Euler equations.*

4. The Euler Solver

We consider a dimensionless form of the 1-D unsteady Euler equations of a perfect gas. Let ρ, v, p, and γ be the mass density, velocity, static pressure, and constant specific heat ratio, respectively. Let

$$u_1 = \rho, \qquad u_2 = \rho v, \qquad u_3 = p/(\gamma - 1) + (1/2)\rho v^2, \tag{4.1}$$

$$f_1 = u_2, \tag{4.2}$$

$$f_2 = (\gamma - 1)u_3 + (1/2)(3 - \gamma)(u_2)^2/u_1, \tag{4.3}$$

and

$$f_3 = \gamma u_2 u_3/u_1 - (1/2)(\gamma - 1)(u_2)^3/(u_1)^2. \tag{4.4}$$

Then the Euler equations can be expressed as

$$\frac{\partial u_m}{\partial t} + \frac{\partial f_m}{\partial x} = 0, \qquad m = 1, 2, 3. \tag{4.5}$$

The integral form of Eq. (4.5) in space-time E_2 is

$$\oint_{S(V)} \vec{h}_m \cdot d\vec{s} = 0, \qquad m = 1, 2, 3, \tag{4.6}$$

where $\vec{h}_m = (f_m, u_m)$, $m = 1, 2, 3$, are the space-time mass, momentum, and energy current density vectors, respectively.

As a preliminary, let

$$f_{m,k} \stackrel{\text{def}}{=} \partial f_m/\partial u_k, \qquad m, k = 1, 2, 3. \tag{4.7}$$

Let F be the matrix formed by $f_{m,k}$, $m, k = 1, 2, 3$. Then Eqs. (4.2)–(4.4) imply that

$$F = \begin{pmatrix} 0 & 1 & 0 \\ -\dfrac{(3-\gamma)}{2}\left(\dfrac{u_2}{u_1}\right)^2 & (3-\gamma)\dfrac{u_2}{u_1} & \gamma - 1 \\ (\gamma-1)\left(\dfrac{u_2}{u_1}\right)^3 - \gamma\dfrac{u_2 u_3}{(u_1)^2} & \gamma\dfrac{u_3}{u_1} - \dfrac{3(\gamma-1)}{2}\left(\dfrac{u_2}{u_1}\right)^2 & \gamma\dfrac{u_2}{u_1} \end{pmatrix} \tag{4.8}$$

Let c be the sonic speed. Then

$$c = \sqrt{\gamma(\gamma - 1)\left[\frac{u_3}{u_1} - \frac{1}{2}\left(\frac{u_2}{u_1}\right)^2\right]}. \tag{4.9}$$

Let G be the 3×3 matrix defined by

$$G \stackrel{\text{def}}{=} \begin{pmatrix} 1 & \frac{u_1}{\sqrt{2}c} & \frac{u_1}{\sqrt{2}c} \\ \frac{u_2}{u_1} & \frac{u_2}{\sqrt{2}c} - \frac{u_1}{\sqrt{2}} & \frac{u_2}{\sqrt{2}c} + \frac{u_1}{\sqrt{2}} \\ \frac{1}{2}\left(\frac{u_2}{u_1}\right)^2 & \frac{(u_2)^2}{2\sqrt{2}cu_1} - \frac{u_2}{\sqrt{2}} + \frac{u_1 c}{\sqrt{2}(\gamma-1)} & \frac{(u_2)^2}{2\sqrt{2}cu_1} + \frac{u_2}{\sqrt{2}} + \frac{u_1 c}{\sqrt{2}(\gamma-1)} \end{pmatrix} \tag{4.10}$$

Then the inverse of G is given by

$$G^{-1} = \begin{pmatrix} 1 - \frac{(\gamma-1)}{2c^2}\left(\frac{u_2}{u_1}\right)^2 & \frac{(\gamma-1)u_2}{c^2 u_1} & -\frac{(\gamma-1)}{c^2} \\ \frac{(\gamma-1)(u_2)^2}{2\sqrt{2}c(u_1)^3} + \frac{u_2}{\sqrt{2}(u_1)^2} & -\frac{1}{\sqrt{2}u_1} - \frac{(\gamma-1)u_2}{\sqrt{2}c(u_1)^2} & \frac{\gamma-1}{\sqrt{2}cu_1} \\ \frac{(\gamma-1)(u_2)^2}{2\sqrt{2}c(u_1)^3} - \frac{u_2}{\sqrt{2}(u_1)^2} & \frac{1}{\sqrt{2}u_1} - \frac{(\gamma-1)u_2}{\sqrt{2}c(u_1)^2} & \frac{\gamma-1}{\sqrt{2}cu_1} \end{pmatrix} \tag{4.11}$$

For any numbers $a_1, a_2, \ldots, a_n$, let $diag(a_1, a_2, \ldots, a_n)$ denote the diagonal matrix with a_1, $a_2, \ldots, a_n$ being the diagonal elements on the first, second, ..., and n-th rows, respectively. Then, by using Eqs. (4.8)–(4.11) and $v = u_2/u_1$, one has

$$G^{-1} F G = diag(v,\ v - c,\ v + c). \tag{4.12}$$

Consider SE's of type I depicted in Fig. 2. For any $(x,t) \in \mathrm{SE}(j,n)$, $u_m(x,t)$, $f_m(x,t)$, and $\vec{h}_m(x,t)$ are approximated by $u_m^*(x,t\,;j,n)$, $f_m^*(x,t\,;j,n)$, and $\vec{h}_m^*(x,t\,;j,n)$, respectively. They will be defined shortly. Let

$$u_m^*(x,t\,;j,n) \stackrel{\text{def}}{=} (u_m)_j^n + (u_{mx})_j^n (x - x_j) + (u_{mt})_j^n (t - t^n), \qquad m = 1,2,3, \tag{4.13}$$

where $(u_m)_j^n$, $(u_{mx})_j^n$, and $(u_{mt})_j^n$ are constants in $\mathrm{SE}(j,n)$. Obviously, they can be considered as the numerical analogues of the values of u_m, $\partial u_m/\partial x$, and $\partial u_m/\partial t$ at (x_j, t^n), respectively.

Let $(f_m)_j^n$ and $(f_{m,k})_j^n$ denote the values of f_m and $f_{m,k}$, respectively, when u_m, $m = 1,2,3$, respectively, assume the values of $(u_m)_j^n$, $m = 1,2,3$. Let

$$(f_{mx})_j^n \stackrel{\text{def}}{=} \sum_{k=1}^{3} (f_{m,k})_j^n (u_{kx})_j^n, \qquad m = 1,2,3, \tag{4.14}$$

and

$$(f_{mt})_j^n \stackrel{\text{def}}{=} \sum_{k=1}^{3} (f_{m,k})_j^n (u_{kt})_j^n, \qquad m = 1,2,3. \tag{4.15}$$

Because

$$\frac{\partial f_m}{\partial x} = \sum_{k=1}^{3} f_{m,k} \frac{\partial u_k}{\partial x}, \tag{4.16}$$

and

$$\frac{\partial f_m}{\partial t} = \sum_{k=1}^{3} f_{m,k} \frac{\partial u_k}{\partial t}, \tag{4.17}$$

$(f_{mx})_j^n$ and $(f_{mt})_j^n$ can be considered as the numerical analogues of the values of $\partial f_m/\partial x$ and $\partial f_m/\partial t$ at (x_j, t^n), respectively. As a result, we assume that

$$f_m^*(x,t\,;j,n) = (f_m)_j^n + (f_{mx})_j^n (x - x_j) + (f_{mt})_j^n (t - t^n), \qquad m = 1,2,3. \tag{4.18}$$

Because $\vec{h}_m = (f_m,\, u_m)$, we also assume that

$$\vec{h}_m^*(x,t\,;j,n) = (f_m^*(x,t\,;j,n),\, u_m^*(x,t\,;j,n)), \qquad m = 1,2,3. \tag{4.19}$$

Note that, by their definitions, (i) $(f_m)_j^n$ and $(f_{m,k})_j^n$, $m = 1,2,3$, are functions of $(u_m)_j^n$, $m = 1,2,3$, (ii) $(f_{mx})_j^n$, $m = 1,2,3$, are functions of $(u_m)_j^n$ and $(u_{mx})_j^n$, $m = 1,2,3$, and (iii) $(f_{mt})_j^n$ are functions of $(u_m)_j^n$ and $(u_{mt})_j^n$, $m = 1,2,3$.

Moreover we assume that, for any $(x,t) \in \text{SE}(j,n)$, $u_m = u_m^*(x,t\,;j,n)$ and $f_m = f_m^*(x,t\,;j,n)$ satisfy Eq. (4.5), i.e.,

$$\frac{\partial u_m^*(x,t\,;j,n)}{\partial t} + \frac{\partial f_m^*(x,t\,;j,n)}{\partial x} = 0. \tag{4.20}$$

According to Eqs. (4.13) and (4.18), Eq. (4.20) is equivalent to

$$(u_{mt})_j^n = -(f_{mx})_j^n. \tag{4.21}$$

Because $(f_{mx})_j^n$ are functions of $(u_m)_j^n$ and $(u_{mx})_j^n$, Eq. (4.21) implies that $(u_{mt})_j^n$ are also functions of $(u_m)_j^n$ and $(u_{mx})_j^n$. From this result and the facts stated following Eq. (4.19), one concludes that *the only independent discrete variables needed to be solved in the current marching scheme are* $(u_m)_j^n$ *and* $(u_{mx})_j^n$.

From Eq. (4.20), one concludes that the generalization of the potential function $\psi(x,t\,;j,n)$ introduced in Sec. 2 to the current solver are $\psi_m(x,t\,;j,n)$, $m = 1,2,3$, which satisfy

$$\frac{\partial \psi_m(x,t\,;j,n)}{\partial t} = f_m^*(x,t\,;j,n), \tag{4.22}$$

and

$$-\frac{\partial \psi_m(x,t\,;j,n)}{\partial x} = u_m^*(x,t\,;j,n). \tag{4.23}$$

Substituting Eqs. (4.13) and (4.18) into Eqs. (4.22) and (4.23), and using Eq. (4.21), one concludes that, up to an arbitrary constant,

$$\begin{aligned}\psi_m(x,t\,;j,n) = (f_m)_j^n(t-t^n) - (u_m)_j^n(x-x_j) + (1/2)(f_{mt})_j^n(t-t^n)^2 \\ - (1/2)(u_{mx})_j^n(x-x_j)^2 + (f_{mx})_j^n(x-x_j)(t-t^n).\end{aligned} \tag{4.24}$$

By using an argument similar to that leading to Eq. (2.36), one concludes that

$$\int_\Gamma \vec{h}_m^* \cdot d\vec{s} = \psi_m(x',t'\,;j,n) - \psi_m(x,t\,;j,n). \tag{4.25}$$

Here Γ is a simple curve joining (x,t) and (x',t'), and *lying entirely within* $SE(j,n)$. We also assume that $d\vec{s}$ points to the right of Γ if one moves forward from (x,t) to (x',t').

As in the a-ϵ scheme, we assume that the flux of $\vec{h}_m^*$ is conserved over $\mathrm{CE}(j,n)$, i.e.,

$$\oint_{S(CE(j,n))} \vec{h}_m^* \cdot d\vec{s} = 0. \tag{4.26}$$

Combining Eqs. (4.25) and (4.26), one has

$$\begin{aligned}&\psi_m(x_j - \Delta x/2, t^n;\, j,n) - \psi_m(x_j + \Delta x/2, t^n;\, j,n) \\ &+ \psi_m(x_{j-1/2} + \Delta x/2, t^{n-1/2};\, j-1/2, n-1/2) \\ &- \psi_m(x_{j-1/2}, t^{n-1/2} + \Delta t/2;\, j-1/2, n-1/2) \\ &+ \psi_m(x_{j+1/2}, t^{n-1/2} + \Delta t/2;\, j+1/2, n-1/2) \\ &- \psi_m(x_{j+1/2} - \Delta x/2, t^{n-1/2};\, j+1/2, n-1/2) = 0.\end{aligned} \tag{4.27}$$

Substitution of Eq. (4.24) into Eq. (4.27) yields

$$(u_m)_j^n = \frac{1}{2}\left[(u_m)_{j-1/2}^{n-1/2} + (u_m)_{j+1/2}^{n-1/2} + (s_m)_{j-1/2}^{n-1/2} - (s_m)_{j+1/2}^{n-1/2}\right], \tag{4.28}$$

where, for all $(j,n) \in \Omega$,

$$(s_m)_j^n \stackrel{\text{def}}{=} \frac{\Delta x}{4}(u_{mx})_j^n + \frac{\Delta t}{\Delta x}(f_m)_j^n + \frac{(\Delta t)^2}{4\Delta x}(f_{mt})_j^n, \qquad m = 1,2,3. \tag{4.29}$$

Eq. (4.28) forms the first half of the current marching scheme. The second half which solves $(u_{mx})_j^n$ will come from a generalization of Eq. (3.10).

For all $(j,n) \in \Omega$, let

$$(du_{mx})_j^n \stackrel{\text{def}}{=} \frac{1}{2}\left[(u_{mx})_{j+1/2}^{n-1/2} + (u_{mx})_{j-1/2}^{n-1/2}\right] - \left[(u_m)_{j+1/2}^{n-1/2} - (u_m)_{j-1/2}^{n-1/2}\right] \Big/ \Delta x, \tag{4.30}$$

and

$$(u'_m)_{j\pm1/2}^n \stackrel{\text{def}}{=} (u_m)_{j\pm1/2}^{n-1/2} + (\Delta t/2)(u_{mt})_{j\pm1/2}^{n-1/2}, \tag{4.31}$$

for $m = 1,2,3$. Because Eqs. (4.30) and (4.31) are the generalizations of Eqs. (3.2) and (3.11), respectively, a natural generalization of Eq. (3.10) is

$$(u_{mx})_j^n = \left[(u'_m)_{j+1/2}^n - (u'_m)_{j-1/2}^n\right] \Big/ \Delta x + (2\epsilon - 1)(du_{mx})_j^n, \qquad m = 1,2,3, \tag{4.32}$$

where ϵ is a parameter independent of numerical variables. Note that the last term in Eq. (4.32) vanishes if $\epsilon = 1/2$. The marching scheme presented in [2] is formed by Eqs. (4.28) and (4.32) with $\epsilon = 1/2$.

To construct a larger class of generalizations to Eq. (3.10), for all $(j,n) \in \Omega$, let

$$(\hat{u}_m)_j^n \stackrel{\text{def}}{=} \frac{1}{2}\left[(u_m)_{j+1/2}^{n-1/2} + (u_m)_{j-1/2}^{n-1/2}\right], \qquad m = 1,2,3. \tag{4.33}$$

Let $(\hat{\epsilon}_m)_j^n$, $m = 1,2,3$, be parameters that can be functions of $(\hat{u}_m)_j^n$, $m = 1,2,3$. There can be many choices of these functions. Let $(\hat{g}_{mk})_j^n$ be the value of the (m,k)–element of the matrix G when u_m, $m = 1,2,3$, respectively, assume the values of $(\hat{u}_m)_j^n$, $m = 1,2,3$. Similarly, let $(\hat{g}_{mk}^{-1})_j^n$ be the value of the (m,k)–element of the matrix G^{-1} when u_m, $m = 1,2,3$, respectively, assume the values of $(\hat{u}_m)_j^n$, $m = 1,2,3$. Let

$$(\hat{e}_{mk})_j^n \stackrel{\text{def}}{=} \sum_{l=1}^{3} (\hat{g}_{ml})_j^n (\hat{\epsilon}_l)_j^n (\hat{g}_{lk}^{-1})_j^n, \qquad m,k = 1,2,3. \tag{4.34}$$

Then Eq. (3.10) can be generalized as

$$(u_{mx})_j^n = \left[(u'_m)_{j+1/2}^n - (u'_m)_{j-1/2}^n\right] \Big/ \Delta x + \sum_{k=1}^{3} \left[2(\hat{e}_{mk})_j^n - \delta_{mk}\right] (du_{kx})_j^n, \tag{4.35}$$

where $m = 1,2,3$, and δ_{mk} is the kronecker-delta symbol.

Consider the special case in which, for all $(j,n) \in \Omega$, $(\hat{\epsilon}_1)_j^n = (\hat{\epsilon}_2)_j^n = (\hat{\epsilon}_3)_j^n$. Let $(\hat{\epsilon}_m)_j^n = (\hat{\epsilon})_j^n$, $m = 1,2,3$. Then $(\hat{e}_{mk})_j^n = (\hat{\epsilon})_j^n \delta_{mk}$, and thus Eq. (4.35) is reduced to

$$(u_{mx})_j^n = \left[(u'_m)_{j+1/2}^n - (u'_m)_{j-1/2}^n\right] \Big/ \Delta x + \left[2(\hat{\epsilon})_j^n - 1\right] (du_{mx})_j^n, \qquad m = 1,2,3, \tag{4.36}$$

Note that Eq. (4.36) reduces to Eq. (4.32) if $(\hat{\epsilon})_j^n = \epsilon$ for all $(j,n) \in \Omega$.

The current Euler solver is a straightforward extension of the a-ϵ scheme. Eqs. (4.28) and (4.35), which form the marching scheme, will be converted into a matrix form similar to Eq. (3.6). Eqs. (4.43) and (4.44), to be obtained during this conversion, will also be used in a stabilty analysis presented in Sec. 6. To proceed, note that f_m, $m = 1,2,3$, are homogeneous functions of degree 1 [20, p.11] in the variables u_m, $m = 1,2,3$. Thus

$$(f_m)_j^n = \sum_{k=1}^{3} (f_{m,k})_j^n (u_k)_j^n. \tag{4.37}$$

Also, by using Eqs. (4.15), (4.21) and (4.14), one has

$$(f_{mt})_j^n = -\sum_{k=1}^{3}\sum_{l=1}^{3} (f_{m,l})_j^n (f_{l,k})_j^n (u_{kx})_j^n. \tag{4.38}$$

Substituting Eqs. (4.37) and (4.38) into Eq. (4.29), and using the definitions

$$(u_{mx}^+)_j^n \stackrel{\text{def}}{=} \frac{\Delta x}{4}(u_{mx})_j^n, \qquad m = 1,2,3, \tag{4.39}$$

and

$$(f_{m,k}^+)_j^n \stackrel{\text{def}}{=} \frac{\Delta t}{\Delta x}(f_{m,k})_j^n, \qquad m,k = 1,2,3, \tag{4.40}$$

one has

$$(s_m)_j^n = \sum_{k=1}^{3} (f_{m,k}^+)_j^n (u_k)_j^n + \sum_{k=1}^{3} \left[\delta_{mk} - \sum_{l=1}^{3} (f_{m,l}^+)_j^n (f_{l,k}^+)_j^n\right] (u_{kx}^+)_j^n, \qquad m = 1,2,3. \tag{4.41}$$

Substituting Eqs. (4.21) and (4.14) into Eq. (4.31) and using Eqs. (4.39) and (4.40), one also has

$$(u_m')_{j\pm 1/2}^n = (u_m)_{j\pm 1/2}^{n-1/2} - 2\sum_{k=1}^{3} (f_{m,k}^+)_{j\pm 1/2}^{n-1/2} (u_{kx}^+)_{j\pm 1/2}^{n-1/2}. \tag{4.42}$$

With the aid of Eqs. (4.39)–(4.42) and (4.30), Eqs. (4.28) and (4.35) can be expressed as

$$\begin{aligned}
(u_m)_j^n = \frac{1}{2}\sum_{k=1}^{3}\Bigg\{ & \left[\delta_{mk} + (f_{m,k}^+)_{j-1/2}^{n-1/2}\right](u_k)_{j-1/2}^{n-1/2} \\
& + \left[\delta_{mk} - \sum_{l=1}^{3} (f_{m,l}^+)_{j-1/2}^{n-1/2} (f_{l,k}^+)_{j-1/2}^{n-1/2}\right](u_{kx}^+)_{j-1/2}^{n-1/2} \\
& + \left[\delta_{mk} - (f_{m,k}^+)_{j+1/2}^{n-1/2}\right](u_k)_{j+1/2}^{n-1/2} \\
& - \left[\delta_{mk} - \sum_{l=1}^{3} (f_{m,l}^+)_{j+1/2}^{n-1/2} (f_{l,k}^+)_{j+1/2}^{n-1/2}\right](u_{kx}^+)_{j+1/2}^{n-1/2}\Bigg\}, \qquad m = 1,2,3,
\end{aligned} \tag{4.43}$$

and

$$(u^+_{mx})^n_j = \frac{1}{2}\sum_{k=1}^{3}\Bigg\{\left[(\hat{e}_{mk})^n_j - \delta_{mk}\right]\left[(u_k)^{n-1/2}_{j-1/2} - (u_k)^{n-1/2}_{j+1/2}\right]$$

$$+\left[2(\hat{e}_{mk})^n_j - \delta_{mk} + (f^+_{m,k})^{n-1/2}_{j-1/2}\right](u^+_{kx})^{n-1/2}_{j-1/2} \tag{4.44}$$

$$+\left[2(\hat{e}_{mk})^n_j - \delta_{mk} - (f^+_{m,k})^{n-1/2}_{j+1/2}\right](u^+_{kx})^{n-1/2}_{j+1/2}\Bigg\}, \qquad m = 1,2,3,$$

respectively.

For all $(j,n) \in \Omega$, let $\vec{u}^n_j$ and $(\vec{u}^+_x)^n_j$, respectively, denote the column matrices formed by $(u_m)^n_j$ and $(u^+_{mx})^n_j$, $m = 1,2,3$. Let $(\hat{E})^n_j$ and $(F^+)^n_j$, respectively, denote the 3×3 matrices formed by $(\hat{e}_{mk})^n_j$ and $(f^+_{m,k})^n_j$, $m,k = 1,2,3$. Let I denote the 3×3 identity matrix, i.e., the matrix formed by δ_{mk}, $m,k = 1,2,3$. For all $(j,n) \in \Omega$, let

$$\vec{\mathbf{q}}(j,n) \stackrel{\text{def}}{=} \begin{pmatrix} \vec{u}^n_j \\ (\vec{u}^+_x)^n_j \end{pmatrix}, \tag{4.45}$$

$$\mathbf{M}_+(j,n) \stackrel{\text{def}}{=} (1/2)\begin{pmatrix} I + (F^+)^{n-1/2}_{j-1/2} & I - \left[(F^+)^{n-1/2}_{j-1/2}\right]^2 \\ (\hat{E})^n_j - I & 2(\hat{E})^n_j - I + (F^+)^{n-1/2}_{j-1/2} \end{pmatrix}, \tag{4.46}$$

and

$$\mathbf{M}_-(j,n) \stackrel{\text{def}}{=} (1/2)\begin{pmatrix} I - (F^+)^{n-1/2}_{j+1/2} & \left[(F^+)^{n-1/2}_{j+1/2}\right]^2 - I \\ I - (\hat{E})^n_j & 2(\hat{E})^n_j - I - (F^+)^{n-1/2}_{j+1/2} \end{pmatrix}. \tag{4.47}$$

By their definitions, $\vec{\mathbf{q}}(j,n)$ is a 6×1 column matrix while $\mathbf{M}_+(j,n)$ and $\mathbf{M}_-(j,n)$ are 6×6 matrices. With the aid of Eqs. (4.45)–(4.47), Eqs. (4.43) and (4.44) can be cast into the matrix form

$$\vec{\mathbf{q}}(j,n) = \mathbf{M}_+(j,n)\vec{\mathbf{q}}(j-1/2, n-1/2) + \mathbf{M}_-(j,n)\vec{\mathbf{q}}(j+1/2, n-1/2). \tag{4.48}$$

Note that the set of equations given in Eqs. (4.45)–(4.48) and the set given in Eqs. (2.15), (3.8), (3.9) and (3.6) are very similar in their forms.

Recall that both v and c are functions of u_m, $m = 1,2,3$. For all SE(j,n), let $\hat{v}_j^n$ and $\hat{c}_j^n$, respectively, denote the values of v and c when u_m, $m = 1,2,3$, respectively, assume the values of $(\hat{u}_m)_j^n$, $m = 1,2,3$. It will be shown in Sec. 6 that the marching scheme defined by Eq. (4.48) can be linearized and decoupled into three pairs of equations with each pair being in the form of Eq. (3.6). This decoupling and other considerations lead to the conclusion that Eq. (4.48) is stable if, for all $(j,n) \in \Omega$,

$$(\hat{\nu}_{max})_j^n < 1, \qquad \text{and} \qquad 0 \le (\hat{\epsilon}_m)_j^n \le 1, \quad m = 1,2,3, \tag{4.49}$$

where

$$(\hat{\nu}_{max})_j^n \stackrel{\text{def}}{=} \left(|\hat{v}_j^n| + |\hat{c}_j^n|\right) \frac{\Delta t}{\Delta x}. \tag{4.50}$$

We conclude this section by introducing some possible modifications to the above solver. Note that $(u'_m)_{j\pm1/2}^n$, by its definition, represents a finite-difference approximation of u_m at $(j \pm 1/2, n)$. As a result,

$$(u_{mx}^c)_j^n \stackrel{\text{def}}{=} \left[(u'_m)_{j+1/2}^n - (u'_m)_{j-1/2}^n\right] \Big/ \Delta x, \qquad m = 1,2,3, \tag{4.51}$$

respectively, are the central-difference approximations for $\partial u_m/\partial x$, $m = 1,2,3$, at (j,n). Note that $(u_{mx}^c)_j^n$ is the first term on the right side of each of Eqs. (4.32), (4.35) and (4.36). The above central-difference approximation is valid as long as no discontinuity of u_m (or its derivatives) occurs between $(j-1/2,n)$ and $(j+1/2,n)$ (see Fig. 5). In the following discussion, we develop alternates which are valid even in the presence of discontinuity.

Let

$$(u_{mx\pm})_j^n \stackrel{\text{def}}{=} \pm \frac{(u'_m)_{j\pm1/2}^n - (u_m)_j^n}{\Delta x/2}, \qquad m = 1,2,3, \tag{4.52}$$

where $(u_m)_j^n$ can be obtained from Eq. (4.28). Because $(u'_m)_{j-1/2}^n$, $(u_m)_j^n$ and $(u'_m)_{j+1/2}^n$, are the numerical analogues of u_m at $(j-1/2,n)$, (j,n) and $(j+1/2,n)$, respectively, $(u_{mx-})_j^n$ and $(u_{mx+})_j^n$ are two numerical analogues of the value of $\partial u_m/\partial x$ at (j,n) with one being evaluated from the left and another from the right. Note that

$$(u_{mx}^c)_j^n = \frac{1}{2}\left[(u_{mx-})_j^n + (u_{mx+})_j^n\right]. \tag{4.53}$$

In case a discontinuity occurs between (j,n) and $(j+1/2,n)$ but not between (j,n) and $(j-1/2,n)$, one would expect that $|(u_{mx+})_j^n| \gg |(u_{mx-})_j^n|$. Moreover, because (j,n) and $(j-1/2,n)$ are on the same side of the discontinuity while (j,n) and $(j+1/2,n)$ are on the opposite sides, $(u_{mx})_j^n$ should be a weighted average of $(u_{mx+})_j^n$ and $(u_{mx-})_j^n$ biased toward the one with the smaller magnitude.

As a result of the above considerations, $(u_{mx}^c)_j^n$ can be replaced by

$$(u_{mx}^{w_o})_j^n \stackrel{\text{def}}{=} W_o\left((u_{mx-})_j^n, (u_{mx+})_j^n; \alpha\right), \qquad m = 1,2,3. \tag{4.54}$$

Here α is an adjustable constant and the function W_o is defined by (i) $W_o(0,0,\alpha)=0$ and (ii)

$$W_o(x_-,x_+;\alpha)=\frac{|x_+|^\alpha x_- + |x_-|^\alpha x_+}{|x_+|^\alpha + |x_-|^\alpha}, \qquad (|x_+|+|x_-|>0) \tag{4.55}$$

where x_+ and x_- are any two real variables. Note that $W_o(x_-,x_+;\alpha)=(x_-+x_+)/2$, i.e., $(u_{mx}^{w_o})_j^n = (u_{mx}^c)_j^n$, if $\alpha=0$ or $|x_-|=|x_+|$. Also the expression on the right side of Eq. (4.55) represents a weighted average of x_- and x_+ with the weight factors $|x_+|^\alpha/(|x_+|^\alpha+|x_-|^\alpha)$ and $|x_-|^\alpha/(|x_+|^\alpha+|x_-|^\alpha)$. *For $\alpha>0$, this average is biased toward the one among x_+ and x_- with the smaller magnitude. For the same value of $|x_+|$ and $|x_-|$, the bias increases as α increases.* Thus, we should always choose $\alpha \geq 0$.

Note that the special weighted averages $W_o(x_-,x_+;1)$ and $W_o(x_-,x_+;2)$ are used in the slope-limiters proposed by van Leer [21] and van Albada [22], respectively.

The above modification, i.e., $(u_{mx}^c)_j^n$ replaced by $(u_{mx}^{w_o})_j^n$, is first given in [2]. It is shown in [2] and also Sec. 8 of the current paper that it is an efficient tool to suppress overshoots and/or numerical oscillations near a discontinuity. Moreover, *because $(u_{mx\pm})_j^n$ are constructed using only the data associated with the mesh points $(j-1/2,n-1/2)$ and $(j+1/2,n-1/2)$, the effect of this modification is highly local, i.e., it generally will not cause the smearing of shock discontinuities.*

However, there may be a price to pay for the above modification. Because a fractional power is costly to evaluate, so is $W_o(x_-,x_+;\alpha)$ if α is not an integer. Moreover, because the bias of this weighted average increases with α, a situation may arise such that the use of an α with $|\alpha|<1$ may be desirable. To obtain a computationally efficient weighted average of arbitrary small bias, let

$$W(x_-,x_+;\alpha,\beta) \stackrel{\text{def}}{=} (1-\beta)W_o(x_-,x_+;0)+\beta W_o(x_-,x_+;\alpha), \tag{4.56}$$

where $\beta \geq 0$ is an adjustable weight factor, and α generally is an integer. Because $W_o(x_-,x_+;0)$ is the simple average of x_- and x_+, Eq. (4.56) defines a *linear weighted average of this simple average and the nonlinear weighted average defined in Eq. (4.55).* Obviously, $W(x_-,x_+;\alpha,\beta)=(1/2)(x_-+x_+)$ if $x_-=x_+$. Furthermore, because

$$W_o(x_-,x_+;-\alpha)=\frac{|x_+|^\alpha x_+ + |x_-|^\alpha x_-}{|x_+|^\alpha + |x_-|^\alpha}, \qquad (|x_+|+|x_-|>0), \tag{4.57}$$

alternatively, $W(x_-,x_+;\alpha,\beta)$ can also be expressed as

$$W(x_-,x_+;\alpha,\beta)=\left(\frac{1+\beta}{2}\right)W_o(x_-,x_+,;\alpha)+\left(\frac{1-\beta}{2}\right)W_o(x_-,x_+;-\alpha). \tag{4.58}$$

The application of the more general modification, i.e., $(u_{mx}^c)_j^n$ is replaced by

$$(u_{mx}^w)_j^n \stackrel{\text{def}}{=} W\left((u_{mx-})_j^n,(u_{mx+})_j^n;\alpha,\beta\right), \qquad m=1,2,3, \tag{4.59}$$

will be demonstrated in Sec. 8.

Finally, note that $W(x_-,x_+;\alpha,\beta)$ can be further generalized by a linear weighted average of several $W_o(x_-,x_+;\alpha)$ with different values of α.

5. The Navier–Stokes Solver

We consider a dimensionless form of the 1–D unsteady Navier–Stokes equations of a perfect gas [4, pp.191-193]. (Note: the expressions on the right sides of the last three equations in Eq. (5-47) of [4] have incorrect signs in the earlier versions. The conduction heat-flux vector should be proportional to the negative of the gradient of temperature.) These equations are extensions of the Euler equations defined in Sec. 4. Thus, unless specified otherwise, the symbols, definitions, and equations given there will be used in this section.

Let Re_L and Pr denote the Reynolds number and Prandtl number. respectively. They are assumed to be nonnegative constants. Let

$$\tilde{f}_1 \stackrel{\text{def}}{=} 0, \tag{5.1}$$

$$\tilde{f}_2 \stackrel{\text{def}}{=} \frac{4}{3Re_L}\frac{u_2}{u_1}, \tag{5.2}$$

and

$$\tilde{f}_3 \stackrel{\text{def}}{=} \frac{2}{3Re_L}\left(\frac{u_2}{u_1}\right)^2 + \frac{\gamma}{Re_L\, Pr}\left[\frac{u_3}{u_1} - \frac{(u_2)^2}{2(u_1)^2}\right]. \tag{5.3}$$

Then, the Navier–Stokes euations can be expressed as

$$\frac{\partial u_m}{\partial t} + \frac{\partial f_m}{\partial x} - \frac{\partial^2 \tilde{f}_m}{\partial x^2} = 0, \qquad m = 1,2,3. \tag{5.4}$$

The integral form of Eq. (5.4) in space-time E_2 is Eq. (4.6) with

$$\vec{h}_m \stackrel{\text{def}}{=} (f_m - \partial \tilde{f}_m/\partial x,\, u_m), \qquad m = 1,2,3. \tag{5.5}$$

As a preliminary, let

$$\tilde{f}_{m,k} \stackrel{\text{def}}{=} \partial \tilde{f}_m/\partial u_k, \qquad m,k = 1,2,3, \tag{5.6}$$

and

$$\tau_1 \stackrel{\text{def}}{=} \frac{4}{3Re_L}, \qquad \tau_2 \stackrel{\text{def}}{=} \frac{\gamma}{Re_L\, Pr}, \qquad \text{and} \qquad \tau_3 \stackrel{\text{def}}{=} \tau_2 - \tau_1. \tag{5.7}$$

Let $\tilde{F}$ denote the 3×3 matrix formed by $\tilde{f}_{m,k}$, $m,k = 1,2,3$. Then Eqs. (5.1)–(5.3) imply that

$$\tilde{F} = \begin{pmatrix} 0 & 0 & 0 \\ -\dfrac{\tau_1 u_2}{(u_1)^2} & \dfrac{\tau_1}{u_1} & 0 \\ \tau_3\dfrac{(u_2)^2}{(u_1)^3} - \tau_2\dfrac{u_3}{(u_1)^2} & -\dfrac{\tau_3 u_2}{(u_1)^2} & \dfrac{\tau_2}{u_1} \end{pmatrix}. \tag{5.8}$$

Again we consider SE's of type I depicted in Fig. 2. For any $(x,t) \in \mathrm{SE}(j,n)$, $u_m(x,t)$, $f_m(x,t)$, $\tilde{f}_m(x,t)$, and $\vec{h}_m(x,t)$, respectively, are approximated by $u^*_m(x,t\,;j,n)$, $f^*_m(x,t\,;j,n)$, $\tilde{f}^*_m(x,t\,;j,n)$, and $\vec{h}^*_m(x,t\,;j,n)$. $u^*_m(x,t\,;j,n)$ and $f^*_m(x,t\,;j,n)$, respectively, are defined in Eqs. (4.13) and (4.18). $\tilde{f}^*_m(x,t\,;j,n)$ and $\vec{h}^*_m(x,t\,;j,n)$ will be defined immediately.

Both $\tilde{f}_m$ and $\tilde{f}_{m,k}$ are functions of u_m, $m = 1,2,3$. Let $(\tilde{f}_m)^n_j$ and $(\tilde{f}_{m,k})^n_j$, respectively, denote the values of $\tilde{f}_m$ and $\tilde{f}_{m,k}$ when u_m, $m = 1,2,3$, respectively, assume the values of $(u_m)^n_j$, $m = 1,2,3$. Let

$$(\tilde{f}_{mx})^n_j \stackrel{\text{def}}{=} \sum_{k=1}^{3} (\tilde{f}_{m,k})^n_j (u_{kx})^n_j, \qquad m = 1,2,3, \tag{5.9}$$

and

$$(\tilde{f}_{mt})^n_j \stackrel{\text{def}}{=} \sum_{k=1}^{3} (\tilde{f}_{m,k})^n_j (u_{kt})^n_j, \qquad m = 1,2,3. \tag{5.10}$$

Using an argument similar to that leading to Eq. (4.18), we assume that

$$\tilde{f}^*_m(x,t\,;j,n) = (\tilde{f}_m)^n_j + (\tilde{f}_{mx})^n_j (x - x_j) + (\tilde{f}_{mt})^n_j (t - t^n), \qquad m = 1,2,3. \tag{5.11}$$

As a result of Eq. (5.5), we also assume that

$$\vec{h}^*_m(x,t\,;j,n) = \left(f^*_m(x,t\,;j,n) - \frac{\partial \tilde{f}^*_m(x,t\,;j,n)}{\partial x},\, u^*_m(x,t\,;j,n) \right), \quad m = 1,2,3. \tag{5.12}$$

Also, we assume that, for any $(x,t) \in \mathrm{SE}(j,n)$, $u_m = u^*_m(x,t\,;j,n)$, $f_m = f^*_m(x,t\,;j,n)$, and $\tilde{f}_m = \tilde{f}^*_m(x,t\,;j,n)$ satisfy Eq. (5.4), i.e.,

$$\frac{\partial u^*_m(x,t\,;j,n)}{\partial t} + \frac{\partial}{\partial x} \left[f^*_m(x,t\,;j,n) - \frac{\partial \tilde{f}^*_m(x,t\,;j,n)}{\partial x} \right] = 0. \tag{5.13}$$

The above condition again leads to Eq. (4.21). Thus, the diffusion term in Eq. (5.4) has no impact on how $u^*_m(x,t\,;j,n)$ varies with time *within* $\mathrm{SE}(j,n)$. This same fact was observed in Sec. 2. The reason behind it and its significance were also discussed there. As a result of Eq.(4.21), and other definitions given earlier in this section, one can conclude that the only independent discrete variables needed to be solved in the current solver, as in the Euler solver described in Sec. 4, are also $(u_m)^n_j$ and $(u_{mx})^n_j$.

A comparison between Eqs. (4.20) and (5.13) reveals that, for the current solver, Eqs. (4.22) and (4.23) should be replaced by

$$\frac{\partial \psi_m(x,t\,;j,n)}{\partial t} = f^*_m(x,t\,;j,n) - \frac{\partial \tilde{f}^*_m(x,t\,;j,n)}{\partial x}, \tag{5.14}$$

and

$$-\frac{\partial \psi_m(x,t;j,n)}{\partial x} = u_m^*(x,t;j,n), \tag{5.15}$$

respectively. Note that Eqs. (5.15) and (4.23) are identical. According to Eq. (5.11), the second term on the right side of Eq. (5.14) is simply the constant $-(\tilde{f}_{mx})_j^n$. Thus, for the current solver, Eq. (4.24) should be replaced by

$$\begin{aligned}\psi_m(x,t;j,n) = (\acute{f}_m)_j^n(t-t^n) - (u_m)_j^n(x-x_j) + (1/2)(f_{mt})_j^n(t-t^n)^2 \\ - (1/2)(u_{mx})_j^n(x-x_j)^2 + (f_{mx})_j^n(x-x_j)(t-t^n),\end{aligned} \tag{5.16}$$

where

$$(\acute{f}_m)_j^n \stackrel{\text{def}}{=} (f_m)_j^n - (\tilde{f}_{mx})_j^n. \tag{5.17}$$

The only difference between Eqs. (4.24) and (5.16) is that $(f_m)_j^n$ in Eq. (4.24) is replaced by $(\acute{f}_m)_j^n$ in Eq. (5.16). Obviously, Eq. (4.25) is still valid for the current solver. Because $\psi_m(x,t;j,n)$ is independent of $(\tilde{f}_m)_j^n$ and $(\tilde{f}_{mt})_j^n$, Eq. (4.25) implies that the last two parameters are irrelevant in flux evaluation. Moreover, because the current solver will be constructed using only flux-balance conditions, these parameters are also irrelevant in the following construction.

For all $(j,n) \in \Omega$, we assume that

$$\oint_{S(CE_\pm(j,n))} \vec{h}_m^* \cdot d\vec{s} = 0. \tag{5.18}$$

With the aid of Eqs.(5.16) and (4.25), Eq. (5.18) implies that, for all $(j,n) \in \Omega$,

$$\begin{aligned}&(u_m)_j^n - (u_m)_{j\pm1/2}^{n-1/2} \pm \frac{\Delta x}{4}\left[(u_{mx})_{j\pm1/2}^{n-1/2} + (u_{mx})_j^n\right] \\ &\pm \frac{\Delta t}{\Delta x}\left[(\acute{f}_m)_{j\pm1/2}^{n-1/2} - (\acute{f}_m)_j^n\right] \pm \frac{(\Delta t)^2}{4\Delta x}\left[(f_{mt})_{j\pm1/2}^{n-1/2} + (f_{mt})_j^n\right] = 0.\end{aligned} \tag{5.19}$$

Adding the two equations given in Eq. (5.19) results in

$$(u_m)_j^n = \frac{1}{2}\left[(u_m)_{j-1/2}^{n-1/2} + (u_m)_{j+1/2}^{n-1/2} + (\acute{s}_m)_{j-1/2}^{n-1/2} - (\acute{s}_m)_{j+1/2}^{n-1/2}\right], \tag{5.20}$$

where, for all $(j,n) \in \Omega$,

$$(\acute{s}_m)_j^n \stackrel{\text{def}}{=} \frac{\Delta x}{4}(u_{mx})_j^n + \frac{\Delta t}{\Delta x}(\acute{f}_m)_j^n + \frac{(\Delta t)^2}{4\Delta x}(f_{mt})_j^n, \qquad m = 1,2,3. \tag{5.21}$$

Eqs. (5.20) and (5.21) are the current counterparts of Eqs. (4.28) and (4.29), respectively. By using Eq. (5.20), $(u_m)_j^n$ can be solved explicitly in terms of discrete variables at the next lower time level.

By substraction of the two equations given in Eq. (5.19), and using Eq. (5.17), one has

$$\frac{\Delta x}{4}(u_{mx})_j^n + \frac{(\Delta t)^2}{4\Delta x}(f_{mt})_j^n + \frac{\Delta t}{\Delta x}(\tilde{f}_{mx})_j^n = (b_m)_j^n, \qquad m = 1,2,3, \tag{5.22}$$

where, for all $(j,n) \in \Omega$, and $m = 1,2,3$,

$$(b_m)_j^n \stackrel{\text{def}}{=} \frac{\Delta t}{\Delta x}(f_m)_j^n + \frac{1}{2}\left[(u_m)_{j+1/2}^{n-1/2} - (u_m)_{j-1/2}^{n-1/2} - (\acute{s}_m)_{j+1/2}^{n-1/2} - (\acute{s}_m)_{j-1/2}^{n-1/2}\right]. \tag{5.23}$$

Note that $(f_m)_j^n$, $m = 1,2,3$, are functions of $(u_m)_j^n$, $m = 1,2,3$, and the latter can be evaluated by using Eq. (5.21). Thus, $(b_m)_j^n$, $m = 1,2,3$, can also be evaluated in terms of the variables at the $(n-1/2)$-th time level.

Moreover, for all $(j,n) \in \Omega$, let

$$(\tilde{f}_{m,k}^+)_j^n \stackrel{\text{def}}{=} \frac{4\Delta t}{(\Delta x)^2}(\tilde{f}_{m,k})_j^n, \qquad m,k = 1,2,3, \tag{5.24}$$

and

$$(a_{mk})_j^n \stackrel{\text{def}}{=} \delta_{mk} + (\tilde{f}_{m,k}^+)_j^n - \sum_{l=1}^{3}(f_{m,l}^+)_j^n (f_{l,k}^+)_j^n, \qquad m,k = 1,2,3. \tag{5.25}$$

Then, with the aid of Eqs. (4.38)–(4.40) and (5.9), Eq. (5.22) can be reexpressed as

$$\sum_{k=1}^{3}(a_{mk})_j^n (u_{kx}^+)_j^n = (b_m)_j^n, \qquad m = 1,2,3. \tag{5.26}$$

Because $(f_{m,k}^+)_j^n$ and $(\tilde{f}_{m,k}^+)_j^n$, $m,k = 1,2,3$, are all functions of $(u_m)_j^n$, $m = 1,2,3$, so are $(a_{mk})_j^n$, $m,k = 1,2,3$. Thus, $(a_{mk})_j^n$ can also be evaluated in terms of the variables at the $(n-1/2)$-th time level. It follows that, for each $(j,n) \in \Omega$, Eq. (5.26) represents a system of three linear equations for three unknowns $(u_{mx}^+)_j^n$, $m = 1,2,3$. These unknowns (and thus $(u_{mx})_j^n$, $m = 1,2,3$, through Eq. (4.39)) can be solved easily by a matrix inversion. Eqs. (5.20) and (5.26) form the current marching scheme.

For all $(j,n) \in \Omega$, let $(\tilde{F}^+)_j^n$ denote the matrix formed by $(\tilde{f}_{m,k}^+)_j^n$, $m,k = 1,2,3$. Also, let

$$(C_\pm)_j^n \stackrel{\text{def}}{=} I \pm (F^+)_j^n, \tag{5.27}$$

and

$$(D_\pm)_j^n \stackrel{\text{def}}{=} I - \left[(F^+)_j^n\right]^2 \pm (\tilde{F}^+)_j^n, \tag{5.28}$$

where $(F^+)_j^n$ is defined in Sec. 4. Note that $(D_+)_j^n$ is the matrix formed by $(a_{mk})_j^n$, $m,k = 1,2,3$. Existence of its inverse $\left[(D_+)_j^n\right]^{-1}$ will be assumed. As a resut, one can define

$$\mathbf{Q}_+(j,n) \stackrel{\text{def}}{=} \frac{1}{2}\begin{pmatrix} (C_+)_{j-1/2}^{n-1/2} & (D_-)_{j-1/2}^{n-1/2} \\ -\left[(D_+)_j^n\right]^{-1}(C_-)_j^n(C_+)_{j-1/2}^{n-1/2} & -\left[(D_+)_j^n\right]^{-1}(C_-)_j^n(D_-)_{j-1/2}^{n-1/2} \end{pmatrix}, \tag{5.29}$$

and

$$\mathbf{Q}_-(j,n) \stackrel{\text{def}}{=} \frac{1}{2}\begin{pmatrix} (C_-)_{j+1/2}^{n-1/2} & -(D_-)_{j+1/2}^{n-1/2} \\ \left[(D_+)_j^n\right]^{-1}(C_+)_j^n(C_-)_{j+1/2}^{n-1/2} & -\left[(D_+)_j^n\right]^{-1}(C_+)_j^n(D_-)_{j+1/2}^{n-1/2} \end{pmatrix}, \tag{5.30}$$

Using Eqs. (5.27)–(5.30), and mathematical manipulations similar to those leading to Eq. (4.48), Eq. (5.20) and (5.26) can be cast into the matrix form

$$\vec{\mathbf{q}}(j,n) = \mathbf{Q}_+(j,n)\vec{\mathbf{q}}(j-1/2, n-1/2) + \mathbf{Q}_-(j,n)\vec{\mathbf{q}}(j+1/2, n-1/2), \tag{5.31}$$

where $\vec{\mathbf{q}}(j,n)$ is defined in Eq. (4.45). Note that $\vec{\mathbf{q}}(j,n)$ is converted into $\vec{q}(j,n)$ (see Eq. (2.15)) if $\vec{u}_j^n$ and $(\vec{u}_x^+)_j^n$, are replaced by u_j^n and $(\Delta x/4)(u_x)_j^n$, respectively. Also $\mathbf{Q}_+(j,n)$ and $\mathbf{Q}_-(j,n)$ are converted into Q_+ and Q_-, respectively, if (i) $(F^+)_j^n$ and $(F^+)_{j\pm1/2}^{n-1/2}$ are all replaced by ν, and (ii) $(\tilde{F}^+)_j^n$ and $(\tilde{F}^+)_{j\pm1/2}^{n-1/2}$ are all replaced by ξ. Thus, Eq. (5.31) is converted into Eq. (2.14) after the above substitutions.

6. Stability Analysis

The stability of the a-μ scheme will be studied using the von Neumann analysis. For all $(j,n) \in \Omega$, let

$$\vec{q}(j,n) = \vec{q}^*(n,\theta)e^{ij\theta} \qquad (i \stackrel{\text{def}}{=} \sqrt{-1}, \quad -\pi < \theta \leq \pi) \tag{6.1}$$

where $\vec{q}^*(n,\theta)$ is a 2×1 column matrix. Substituting Eq. (6.1) into Eq. (2.18), one obtains

$$\vec{q}^*(n+1,\theta) = [Q(\nu,\xi,\theta)]^2 \vec{q}^*(n,\theta) \tag{6.2}$$

where

$$Q(\nu,\xi,\theta) \stackrel{\text{def}}{=} e^{-i\theta/2}Q_+ + e^{i\theta/2}Q_-. \tag{6.3}$$

According to Eq. (6.2), the amplification matrix is the square of the matrix $Q(\nu,\xi,\theta)$. Substituting Eqs. (2.16) and (2.17) into Eq. (6.3), one has

$$Q(\nu,\xi,\theta) = \begin{pmatrix} \cos(\theta/2) - i\nu\sin(\theta/2) & -i(1-\nu^2-\xi)\sin(\theta/2) \\ \dfrac{i(1-\nu^2)\sin(\theta/2)}{1-\nu^2+\xi} & -\dfrac{1-\nu^2-\xi}{1-\nu^2+\xi}[\cos(\theta/2) + i\nu\sin(\theta/2)] \end{pmatrix}. \tag{6.4}$$

Let

$$\eta(\nu,\xi,\theta) \stackrel{\text{def}}{=} \xi\cos(\theta/2) - i\nu(1-\nu^2)\sin(\theta/2). \tag{6.5}$$

Then the eigenvalues of $Q(\nu,\xi,\theta)$ are

$$\sigma_\pm(\nu,\xi,\theta) \stackrel{\text{def}}{=} \frac{\eta(\nu,\xi,\theta) \pm \sqrt{[\eta(\nu,\xi,\theta)]^2 + (1-\nu^2)^2 - \xi^2}}{1-\nu^2+\xi}. \tag{6.6}$$

Thus the amplification factors $G_+^{(1)}$ and $G_-^{(1)}$ of the a-μ scheme are given by

$$G_\pm^{(1)} = [\sigma_\pm(\nu,\xi,\theta)]^2. \tag{6.7}$$

Note that

$$G_+^{(1)} \to 1 \quad \text{and} \quad G_-^{(1)} \to \left(\frac{1-\nu^2-\xi}{1-\nu^2+\xi}\right)^2 \quad \text{as} \quad \theta \to 0 \tag{6.8}$$

if $1-\nu^2 \geq 0$. Because the amplification factor of a plane-wave solution to Eq. (2.1) approaches 1 as $\theta \to 0$, $G_+^{(1)}$ and $G_-^{(1)}$ are referred to as the principal and the spurious amplification factors, respectively. Moreover, Eqs. (6.5)–(6.7) imply that

$$G_\pm^{(1)} = \left[\frac{\hat{\xi}\cos(\theta/2) - i\nu\sin(\theta/2) \pm \sqrt{[\hat{\xi}\cos(\theta/2) - i\nu\sin(\theta/2)]^2 + 1 - \hat{\xi}^2}}{1+\hat{\xi}}\right]^2 \tag{6.9}$$

if $1-\nu^2 \neq 0$, and $\hat{\xi} \stackrel{\text{def}}{=} \xi/(1-\nu^2)$. Similarity between Eqs. (6.9) and (2.21) was noted in Sec. 2.

In [1], the stability of the a-μ scheme is studied using a rigorous discrete Fourier analysis. The von Neumann stability analysis can be considered as a limiting case of the discreate Fourier analysis. By using Eqs. (4.33) and (4.34) in [1], one can infer that the a-μ scheme is stable if and only if, for all θ with $-\pi < \theta \leq \pi$,

$$\max\{|G_+^{(1)}|, |G_-^{(1)}|\} \leq 1 \quad \text{if} \quad Q(\nu, \xi, \theta) \text{ is nondefective} \tag{6.10}$$

and

$$|G_+^{(1)}| < 1 \quad \text{if} \quad Q(\nu, \xi, \theta) \text{ is defective.} \tag{6.11}$$

Note that $G_+^{(1)} = G_-^{(1)}$ if $Q(\nu, \xi, \theta)$ is defective [23, p.353]. Assuming $\xi \geq 0$ and $1-\nu^2+\xi \neq 0$ (the latter is a basic assumption of Eq. (2.14)), it is proved in [1] that the current scheme is stable if and only if $\nu^2 \leq 1$.

Let $(1-\nu^2)^2 \neq \xi^2$ such that both Eqs. (2.14) and (2.24) are valid. Combining Eqs. (6.5)–(6.7), one has

$$G_+^{(1)} G_-^{(1)} = \left(\frac{1-\nu^2-\xi}{1-\nu^2+\xi}\right)^2. \tag{6.12}$$

Because the amplification factors of the backward-marching scheme are $(G_+^{(1)})^{-1}$ and $(G_-^{(1)})^{-1}$, stability of both Eqs. (2.14) and (2.24) requires that $|G_+^{(1)}| = |G_-^{(1)}| = 1$. According to Eq. (6.12), the last condition cannot be met if $\mu > 0$ and $\nu^2 \neq 1$. This result was used in a discussion given in Sec. 2.

Next we study the stability of the a-ϵ scheme. By substituting Eq. (6.1) into Eq. (3.7), one has

$$\bar{q}^*(n+1, \theta) = [M(\epsilon, \nu, \theta)]^2 \bar{q}^*(n, \theta) \tag{6.13}$$

where

$$M(\epsilon, \nu, \theta) \stackrel{\text{def}}{=} e^{-i\theta/2} M_+ + e^{i\theta/2} M_-. \tag{6.14}$$

According to Eq. (6.13), the amplification matrix of the a-ϵ scheme is the square of the matrix $M(\epsilon, \nu, \theta)$. Substituting Eqs. (3.8) and (3.9) into Eq. (6.14), one has

$$M(\epsilon, \nu, \theta) = \begin{pmatrix} \cos(\theta/2) - i\nu \sin(\theta/2) & -i(1-\nu^2)\sin(\theta/2) \\ i(1-\epsilon)\sin(\theta/2) & (2\epsilon-1)\cos(\theta/2) - i\nu\sin(\theta/2) \end{pmatrix}. \tag{6.15}$$

The eigenvalues $\lambda_\pm(\epsilon, \nu, \theta)$ of $M(\epsilon, \nu, \theta)$ were given in Eq. (3.13). The principal amplification factor $G_+^{(2)}$ and the spurious amplification factor $G_-^{(2)}$ of the a-ϵ scheme were given in Eq. (3.12). Note that

$$G_+^{(2)} \to 1 \quad \text{and} \quad G_-^{(2)} \to 2\epsilon - 1 \quad \text{as} \quad \theta \to 0 \tag{6.16}$$

if Eq.(3.14) is assumed. Moreover, from Eqs. (6.10) and (6.11), one infers that the a-ϵ scheme is stable if and only if, for all θ with $-\pi < \theta \leq \pi$,

$$\max\{|G_+^{(2)}|, |G_-^{(2)}|\} \leq 1 \quad \text{if} \quad M(\epsilon, \nu, \theta) \text{ is nondefective} \tag{6.17}$$

and

$$|G_+^{(2)}| < 1 \quad \text{if} \quad M(\epsilon, \nu, \theta) \text{ is defective.} \tag{6.18}$$

Eq. (3.13) implies that

$$|\lambda_+(\epsilon, \nu, 0)||\lambda_-(\epsilon, \nu, 0)| = |2\epsilon - 1|. \tag{6.19}$$

By using Eqs. (3.12) and (6.17)–(6.19), one concludes that stability requires that $|2\epsilon - 1| \leq 1$, i.e., $0 \leq \epsilon \leq 1$. This is the first part of Eq. (3.14). Eq. (3.13) also implies that

$$\lambda_\pm(\epsilon, \nu, \pi) = -i\nu \pm \sqrt{(1-\epsilon)(1-\nu^2)}. \tag{6.20}$$

Thus,

$$\max\{|\lambda_+(\epsilon, \nu, \pi)|, |\lambda_-(\epsilon, \nu, \pi)|\} > 1 \quad \text{if } \nu^2 > 1 \text{ and } \epsilon \leq 1. \tag{6.21}$$

The first part of Eq. (3.14) coupled with Eqs. (6.17), (6.18), and (6.21) implies that $\nu^2 \leq 1$ is necessary for stability. Because the case $\nu^2 = 1$ is ruled out by the basic assumption $1 - \nu^2 \neq 0$ of Eq. (3.6), the second part of Eq. (3.14) is now proved.

To prove Eqs. (3.17) and (3.18), note that Eq. (3.16) implies that

$$\chi_\pm(\epsilon, \nu, \theta) = \epsilon(\chi' \mp \chi'') \tag{6.22}$$

where

$$\chi' \stackrel{\text{def}}{=} (1-\nu^2)\sin^2(\theta/2) + 2(1-\epsilon)\cos^2(\theta/2), \tag{6.23}$$

and

$$\chi'' \stackrel{\text{def}}{=} 2\cos(\theta/2)\sqrt{(1-\epsilon)\left[(1-\epsilon)\cos^2(\theta/2) + (1-\nu^2)\sin^2(\theta/2)\right]}. \tag{6.24}$$

With the aid of Eq. (3.14) and $-\pi < \theta \leq \pi$, Eqs. (6.23) and (6.24) imply that

$$\chi' = \chi'' = 0 \qquad \text{if } \epsilon = 1 \text{ and } \theta = 0, \tag{6.25}$$

$$\chi' \begin{cases} = 0, & \text{if } \epsilon = 1 \text{ and } \theta = 0; \\ > 0, & \text{if } \epsilon \neq 1 \text{ or } \theta \neq 0, \end{cases} \tag{6.26}$$

$$\chi'' \geq 2(1-\epsilon)\cos^2(\theta/2) \geq 0, \tag{6.27}$$

$$\chi' - \chi'' \leq (1-\nu^2)\sin^2(\theta/2), \tag{6.28}$$

and

$$(\chi' - \chi'')(\chi' + \chi'') = (\chi')^2 - (\chi'')^2 = (1-\nu^2)^2 \sin^4(\theta/2). \tag{6.29}$$

For the case $\epsilon = 1$ and $\theta = 0$, Eqs. (3.17) and (3.18) follow immediately from Eqs. (6.22) and (6.25). Thus, in the following proof of Eqs. (3.17) and (3.18), we assume that

$$\epsilon \neq 1 \qquad \text{or} \qquad \theta \neq 0. \tag{6.30}$$

Combining Eqs. (6.26), (6.27), and (6.30), one concludes that

$$\chi' + \chi'' > 0. \tag{6.31}$$

Eqs. (6.29) and (6.31) imply that

$$\chi' - \chi'' \geq 0. \tag{6.32}$$

Eq. (3.18) now follows from Eqs. (3.14), (6.22), (6.28), and (6.32). The validity of the first inequality sign in Eq. (3.17) follows from Eq. (3.18) and the fact that $\epsilon(1-\epsilon) \geq 0$ if $0 \leq \epsilon \leq 1$. The validity of the second inequality sign follows from the fact that

$$\chi_-(\epsilon, \nu, \theta) - \chi_+(\epsilon, \nu, \theta) = 2\epsilon\chi'' \geq 4\epsilon(1-\epsilon)\cos^2(\theta/2). \tag{6.33}$$

Eq. (6.33) is a simple result of Eqs. (6.22) and (6.27). To establish the validity of the last inequality sign in Eq. (3.17), note that

$$\begin{aligned} \chi_-(\epsilon, \nu, \theta) &= \epsilon(\chi' + \chi'') = \epsilon[2\chi' - (\chi' - \chi'')] \leq 2\epsilon\chi' \\ &= 2\epsilon\left[(1-\nu^2)\sin^2(\theta/2) + 2(1-\epsilon)\cos^2(\theta/2)\right], \\ &\leq \max\{2\epsilon(1-\nu^2), 4\epsilon(1-\epsilon)\} \leq 4\epsilon \end{aligned} \tag{6.34}$$

where Eqs. (6.22), (6.32), (6.23), and (3.14) have been used. Moreover, because $|G_-^{(2)}| \geq 0$, Eq. (3.16) implies that

$$\chi_-(\epsilon, \nu, \theta) \leq 1. \tag{6.35}$$

The validity of the last inequality sign in Eq. (3.17) now follows from Eqs. (6.34) and (6.35). Q.E.D.

Next we shall prove that Eq. (3.14) is also sufficient for stability. Note that, as a result of Eqs. (3.17) and (3.18), $0 \leq \chi_\pm(\epsilon, \nu, \theta)$, and thus $|G_\pm^{(2)}| \leq 1$, for all ϵ, ν, and θ satisfying Eq. (3.14) and $-\pi < \theta \leq \pi$. As a result, Eq. (6.17) is always satisfied. To complete the proof, we need only to show that Eq.(6.18) is also satsfied. To proceed, note that $G_+^{(2)} = G_-^{(2)}$ if $M(\epsilon, \nu, \theta)$ is defective. From Eqs. (3.12)–(3.14), one also concludes that $\epsilon = 1$ is necessary if $G_+^{(2)} = G_-^{(2)}$. Moreover, Eq. (6.15) implies that $M(1, \nu, 0)$ is the identity matrix. Thus, one concludes that $\epsilon = 1$ and $\theta \neq 0$ are necessary if $M(\epsilon, \nu, \theta)$ is defective. Because (i)

$$G_\pm^{(2)} = [\cos(\theta/2) - i\nu\sin(\theta/2)]^2 \qquad \text{if } \epsilon = 1, \tag{6.36}$$

and (ii)

$$\left|[\cos(\theta/2) - i\nu\sin(\theta/2)]^2\right| < 1 \qquad \text{if } \nu^2 < 1 \text{ and } \theta \neq 0, \tag{6.37}$$

one arrives at the conclusion that Eq. (6.18) is also satisfied. Q.E.D.

Next we shall study the stability of the Euler solver defined by Eqs. (4.43) and (4.44). Because the von Neumann stability analysis, strictly speaking, is applicable only to a system of constant-coeeficient linear equations, we begin with a linearization of the above nonlinear equations.

Recall that the independent variables for Eqs. (4.43) and (4.44) are $(u_m)_j^n$ and $(u_{mx}^+)_j^n$, with $m = 1,2,3$ and $(j,n) \in \Omega$. The other variables in these equations are functions of them. Let $(u_m)_j^n$ and $(u_{mx}^+)_j^n$ also denote a solution to these equations. Let $(u_m)_j^n + \delta(u_m)_j^n$ and $(u_{mx}^+)_j^n + \delta(u_{mx}^+)_j^n$ denote another solution with $\delta(u_m)_j^n$ and $\delta(u_{mx}^+)_j^n$ being *small* perturbations to $(u_m)_j^n$ and $(u_{mx}^+)_j^n$, respectively. One may consider the second solution (hereafter to be referred to as the perturbed solution) to be the result of the first solution (hereafter to be referred to as the background solution) being perturbed initially by round-off errors at some time level. The purpose of the stability study is to determine whether the induced perturbation will amplify or die off as it propagates down the subsequent time levels.

In the current linearization, we assume that

$$\left|(u_{mx}^+)_j^n\right| \ll \left|(u_m)_j^n\right|, \qquad \text{for} \qquad m = 1,2,3, \quad \text{and} \quad (j,n) \in \Omega. \tag{6.38}$$

According to Eqs. (4.13) and (4.39), the above assumption is equivalent to

$$|u_m^*(x_j + \Delta x/4, t^n; j, n) - u_m^*(x_j, t^n; j, n)| \ll |u_m^*(x_j, t^n; j, n)|, \tag{6.39}$$

i.e., the change in the value of u_m^* in SE(j,n) over a spatial distance of $\Delta x/4$ is negligible compared with the value of u_m^* at (x_j, t^n). Similarly, we assume that

$$\left|(u_m)_{j+1/2}^{n-1/2} - (u_m)_{j-1/2}^{n-1/2}\right| \ll \left|(u_m)_{j\pm1/2}^{n-1/2}\right|. \tag{6.40}$$

With the aid of Eq. (4.33), Eq. (6.40) implies that

$$\left|(u_m)_{j\pm1/2}^{n-1/2} - (\hat{u}_m)_j^n\right| \ll \left|(\hat{u}_m)_j^n\right|. \tag{6.41}$$

Because the magnitude of the round-off error of a small quantity is not necessarily smaller than that of a large quantity, in contrast to Eq. (6.38), we assume that

$$\left|\delta(u_{mx}^+)_j^n\right| \approx \left|\delta(u_m)_j^n\right|, \qquad \text{for} \qquad m = 1,2,3, \quad \text{and} \quad (j,n) \in \Omega. \tag{6.42}$$

Here the symbol "$\approx$" implies that the quantities on both sides have the same order of magnitude. Also because round-off errors could vary erratically from one mesh point to another, in contrast to Eq. (6.40), we assume that

$$\left|\delta(u_m)_{j+1/2}^{n-1/2} - \delta(u_m)_{j-1/2}^{n-1/2}\right| \approx \left|\delta(u_m)_{j\pm1/2}^{n-1/2}\right| \approx \left|\delta(\hat{u}_m)_j^n\right|. \tag{6.43}$$

Let

$$(f^+_{m,k})^{n-1/2}_{j\pm1/2}(u_k)^{n-1/2}_{j\pm1/2} + \delta\left\{(f^+_{m,k})^{n-1/2}_{j\pm1/2}(u_k)^{n-1/2}_{j\pm1/2}\right\}$$

denote the perturbed-solution counterpart of $(f^+_{m,k})^{n-1/2}_{j\pm1/2}(u_k)^{n-1/2}_{j\pm1/2}$. Then

$$\sum_{k=1}^{3}\delta\left\{(f^+_{m,k})^{n-1/2}_{j\pm1/2}(u_k)^{n-1/2}_{j\pm1/2}\right\} = \sum_{k=1}^{3}(f^+_{m,k})^{n-1/2}_{j\pm1/2}\delta(u_k)^{n-1/2}_{j\pm1/2}. \tag{6.44}$$

The proof follows directly from the fact that $f_{m,k}$, $m,k=1,2,3$, are homogeneous functions of degree 0 [20, p11] in the variables u_m, $m=1,2,3$, and thus

$$\sum_{l=1}^{3}\frac{\partial^2 f_m}{\partial u_k \partial u_l}u_l = 0, \qquad m,k=1,2,3. \tag{6.45}.$$

Q.E.D. With the aid of Eqs. (6.38) and (6.40)–(6.44), linearization of Eqs.(4.43) and (4.44) can now proceed by using the fact that both background and perturbed solutions satisfy these two equations.

Recall that, as defined in Eqs. (4.7) and (4.8), $f_{m,k}$, $m,k=1,2,3$, are functions of u_m, $m=1,2,3$. In Sec. 4, we also define $(f_{m,k})^n_j$ to be the value of $f_{m,k}$ when u_m, $m=1,2,3$, respectively, assume the values of $(u_m)^n_j$, $m=1,2,3$. Moreover, $(\hat{u}_m)^n_j$, which is different from $(u_m)^n_j$, is also defined in Eq. (4.33). To simplify the linearized versions of Eqs. (4.49) and (4.44), we define $(\hat{f}_{m,k})^n_j$ to be the value of $f_{m,k}$ when u_m, $m=1,2,3$, respectively, assume the values of $(\hat{u}_m)^n_j$, $m=1,2,3$. Furthermore, let

$$(\hat{f}^+_{m,k})^n_j \stackrel{\text{def}}{=} \frac{\Delta t}{\Delta x}(\hat{f}_{m,k})^n_j, \qquad m,k=1,2,3. \tag{6.46}$$

Then the linearized versions of Eqs. (4.49) and (4.44) can be expressed as

$$\begin{aligned}
\delta(u_m)^n_j = \frac{1}{2}\sum_{k=1}^{3}\Bigg\{ & \left[\delta_{mk} + (\hat{f}^+_{m,k})^n_j\right]\delta(u_k)^{n-1/2}_{j-1/2} \\
& + \left[\delta_{mk} - \sum_{l=1}^{3}(\hat{f}^+_{m,l})^n_j(\hat{f}^+_{l,k})^n_j\right]\delta(u^+_{kx})^{n-1/2}_{j-1/2} \\
& + \left[\delta_{mk} - (\hat{f}^+_{m,k})^n_j\right]\delta(u_k)^{n-1/2}_{j+1/2} \\
& - \left[\delta_{mk} - \sum_{l=1}^{3}(\hat{f}^+_{m,l})^n_j(\hat{f}^+_{l,k})^n_j\right]\delta(u^+_{kx})^{n-1/2}_{j+1/2}\Bigg\}, \qquad m=1,2,3,
\end{aligned} \tag{6.47}$$

and

$$\delta(u^+_{mx})^n_j = \frac{1}{2}\sum_{k=1}^{3}\Bigg\{ \left[(\hat{e}_{mk})^n_j - \delta_{mk}\right]\left[\delta(u_k)^{n-1/2}_{j-1/2} - \delta(u_k)^{n-1/2}_{j+1/2}\right]$$
$$+ \left[2(\hat{e}_{mk})^n_j - \delta_{mk} + (\hat{f}^+_{m,k})^n_j\right]\delta(u^+_{kx})^{n-1/2}_{j-1/2} \qquad (6.48)$$
$$+ \left[2(\hat{e}_{mk})^n_j - \delta_{mk} - (\hat{f}^+_{m,k})^n_j\right]\delta(u^+_{kx})^{n-1/2}_{j+1/2}\Bigg\}, \qquad m = 1,2,3,$$

respectively. Note that, in arriving at the final forms given above, we have replaced both $(f^+_{m,k})^{n-1/2}_{j+1/2}$ and $(f^+_{m,k})^{n-1/2}_{j-1/2}$ with $(\hat{f}^+_{m,k})^n_j$. According to Eq. (6.41), these substitutions introduce only errors which are higher order than every terms presented in Eqs. (6.47) and (6.48).

A comparison among Eqs. (4.43), (4.44), (6.47), and (6.48) reveals that the first two equations will turn into the last two equations if, for all $m,k = 1,2,3$, and $(j,n) \in \Omega$, $(u_m)^n_j$, $(u^+_{mx})^n_j$ and $(f^+_{m,k})^{n-1/2}_{j\pm1/2}$ are replaced by $\delta(u_m)^n_j$, $\delta(u^+_{mx})^n_j$ and $(\hat{f}^+_{m,k})^n_j$, respectively. Roughly speaking, one can say that Eqs. (6.47) and (6.48) can be obtained from Eqs. (4.43) and (4.44) by "freezing" the coefficients and replacing the "background" variables $(u_m)^n_j$ and $(u^+_{mx})^n_j$ with the "perturbation" variables $\delta(u_m)^n_j$ and $\delta(u^+_{mx})^n_j$, respectively, for all $m = 1,2,3$, and $(j,n) \in \Omega$.

The von Neumann stability analysis for Eqs. (6.47) and (6.48) can be performed easily after each of them is decoupled. To proceed, first we convert them into matrix forms. For all $(j,n) \in \Omega$, let $\delta\vec{u}^n_j$ and $\delta(\vec{u}^+_x)^n_j$, respectively, denote the column matrices formed by $\delta(u_m)^n_j$ and $\delta(u^+_{mx})^n_j$, $m = 1,2,3$. Let $(\hat{F}^+)^n_j$, $(\hat{G})^n_j$, and $(\hat{G}^{-1})^n_j$, respectively, denote the 3×3 matrices formed by $(\hat{f}^+_{m,k})^n_j$, $(\hat{g}_{mk})^n_j$, and $(\hat{g}^{-1}_{mk})^n_j$, $m,k = 1,2,3$. Then Eq. (4.34) is equivalent to

$$(\hat{E})^n_j = (\hat{G})^n_j \; diag\left((\hat{\epsilon}_1)^n_j, (\hat{\epsilon}_2)^n_j, (\hat{\epsilon}_3)^n_j\right)(\hat{G}^{-1})^n_j, \qquad (6.49)$$

where $(\hat{E})^n_j$ was defined in Sec. 4. Furthermore, Eqs. (6.47) and (6.48) can be reexpressed as

$$\delta\vec{u}^n_j = \frac{1}{2}\Bigg\{\left[I + (\hat{F}^+)^n_j\right]\delta\vec{u}^{n-1/2}_{j-1/2} + \left[I - \left((\hat{F}^+)^n_j\right)^2\right]\delta(\vec{u}^+_x)^{n-1/2}_{j-1/2}$$
$$+ \left[I - (\hat{F}^+)^n_j\right]\delta\vec{u}^{n-1/2}_{j+1/2} - \left[I - \left((\hat{F}^+)^n_j\right)^2\right]\delta(\vec{u}^+_x)^{n-1/2}_{j+1/2}\Bigg\}, \qquad (6.50)$$

and

$$\delta(\vec{u}_x^+)_j^n = \frac{1}{2}\Big\{\left[(\hat{E})_j^n - I\right]\left[\delta\vec{u}_{j-1/2}^{n-1/2} - \delta\vec{u}_{j+1/2}^{n-1/2}\right]$$
$$+ \left[2(\hat{E})_j^n - I + (\hat{F}^+)_j^n\right]\delta(\vec{u}_x^+)_{j-1/2}^{n-1/2} + \left[2(\hat{E})_j^n - I - (\hat{F}^+)_j^n\right]\delta(\vec{u}_x^+)_{j+1/2}^{n-1/2}\Big\}, \tag{6.51}$$

respectively.

A result of Eqs. (4.12) and (4.40) is

$$(\hat{G}^{-1})_j^n\,(\hat{F}^+)_j^n\,(\hat{G})_j^n = diag((\hat{\nu}_1)_j^n, (\hat{\nu}_2)_j^n, (\hat{\nu}_3)_j^n), \tag{6.52}$$

where

$$(\hat{\nu}_1)_j^n \stackrel{\text{def}}{=} \frac{\hat{v}_j^n \Delta t}{\Delta x}, \quad (\hat{\nu}_2)_j^n \stackrel{\text{def}}{=} \frac{(\hat{v}_j^n - \hat{c}_j^n)\Delta t}{\Delta x}, \quad \text{and} \quad (\hat{\nu}_3)_j^n \stackrel{\text{def}}{=} \frac{(\hat{v}_j^n + \hat{c}_j^n)\Delta t}{\Delta x}. \tag{6.53}$$

By using Eqs. (6.49) and (6.52) and matrix manupulations such as

$$(\hat{G}^{-1})_j^n(\hat{F}^+)_j^n\delta\vec{u}_{j\pm1/2}^{n-1/2} = (\hat{G}^{-1})_j^n(\hat{F}^+)_j^n(\hat{G})_j^n(\hat{G}^{-1})_j^n\delta\vec{u}_{j\pm1/2}^{n-1/2}, \tag{6.54}$$

and

$$(\hat{G}^{-1})_j^n\left((\hat{F}^+)_j^n\right)^2\delta(\vec{u}_x^+)_{j\pm1/2}^{n-1/2} = \left((\hat{G}^{-1})_j^n(\hat{F}^+)_j^n(\hat{G})_j^n\right)^2(\hat{G}^{-1})_j^n\delta(\vec{u}_x^+)_{j\pm1/2}^{n-1/2}, \tag{6.55}$$

it is easy to see that both Eqs. (6.50) and (6.51) can be decoupled if the expressions on both sides of them are multiplied from left by the matrix $(\hat{G}^{-1})_j^n$. For all $(j,n) \in \Omega$, let $\delta'(u_m)_j^n$ and $\delta'(u_{mx}^+)_j^n$, $m = 1,2,3$, respectively, denote the components of the column matrices $(\hat{G}^{-1})_j^n\,\delta\vec{u}_j^n$ and $(\hat{G}^{-1})_j^n\,\delta(\vec{u}_x^+)_j^n$. Then the decoupled equations can be written as

$$\delta'(u_m)_j^n = \frac{1}{2}\Big\{\left[1 + (\hat{\nu}_m)_j^n\right]\delta'(u_m)_{j-1/2}^{n-1/2} + \left[1 - \left((\hat{\nu}_m)_j^n\right)^2\right]\delta'(u_{mx}^+)_{j-1/2}^{n-1/2}$$
$$+ \left[1 - (\hat{\nu}_m)_j^n\right]\delta'(u_m)_{j+1/2}^{n-1/2} - \left[1 - \left((\hat{\nu}_m)_j^n\right)^2\right]\delta'(u_{mx}^+)_{j+1/2}^{n-1/2}\Big\}, \qquad m = 1,2,3, \tag{6.56}$$

and

$$\delta'(u_{mx}^+)_j^n = \frac{1}{2}\Big\{\left[(\hat{\epsilon}_m)_j^n - 1\right]\left[\delta'(u_m)_{j-1/2}^{n-1/2} - \delta'(u_m)_{j+1/2}^{n-1/2}\right]$$
$$+ \left[2(\hat{\epsilon}_m)_j^n - 1 + (\hat{\nu}_m)_j^n\right]\delta'(u_{mx}^+)_{j-1/2}^{n-1/2} \tag{6.57}$$
$$+ \left[2(\hat{\epsilon}_m)_j^n - 1 - (\hat{\nu}_m)_j^n\right]\delta'(u_{mx}^+)_{j+1/2}^{n-1/2}\Big\}. \qquad m = 1,2,3.$$

With the aid of Eqs. (2.15), (3.8) and (3.9), one can see that, for each m, Eqs. (6.56) and (6.57), respectively, will be converted into the two component equations contained in Eqs. (3.6) if, for all $(j,n) \in \Omega$, $\delta'(u_m)_j^n$, $\delta'(u_{mx}^+)_j^n$, $(\hat{\nu}_m)_j^n$, and $(\hat{\epsilon}_m)_j^n$, respectively, are replaced by u_j^n, $(\Delta x/4)(u_x)_j^n$, ν, and ϵ. Moreover, if we consider only the Fourier components of $\delta'(u_m)_j^n$ and $\delta'(u_{mx}^+)_j^n$ with sufficiently short wavelengths, i.e., $(\hat{\nu}_m)_j^n$ and $(\hat{\epsilon}_m)_j^n$ do not vary substantially over these wavelengths, the last two coefficients can be considered as constants in the von Neumann stability analysis. Note that round-off errors generally are dominated by short-wavelength Fourier components. As a result of the above considerations, at least approximately, one can obtain the stability conditions of the marching scheme defined by Eqs. (4.43) and (4.44) by a straightforward generalization of Eq. (3.14), i.e., the marching scheme is stable if, for all $(j,n) \in \Omega$,

$$0 \le (\hat{\epsilon}_m)_j^n \le 1, \qquad \text{and} \qquad \left[(\hat{\nu}_m)_j^n\right]^2 < 1, \qquad m = 1,2,3. \tag{6.58}$$

By using Eq. (6.53), it is easy to see that Eq. (4.49) is equivalent to Eq. (6.58).

7. Consistency and the Truncation Error

Consistency and the truncation error of the a-μ scheme were studied and given in Sec. 6 of [1]. In this section, a similar study for the a-ϵ scheme is presented.

As a preliminary, note that Eq. (3.7) can be expressed explicitly as

$$\begin{aligned} u_j^{n+1} = \frac{1}{4}\Big\{&(1+\nu)[\epsilon+(2-\epsilon)\nu]u_{j-1}^n + (1/2)(1-\nu^2)(\epsilon+\nu)\Delta x(u_x)_{j-1}^n \\ &+ 2(2-\epsilon)(1-\nu^2)u_j^n - \nu(1-\nu^2)\Delta x(u_x)_j^n \\ &+ (1-\nu)[\epsilon-(2-\epsilon)\nu]u_{j+1}^n - (1/2)(1-\nu^2)(\epsilon-\nu)\Delta x(u_x)_{j+1}^n\Big\}, \end{aligned} \tag{7.1}$$

and

$$\begin{aligned} &\Delta x(u_x)_j^{n+1} = \\ &2(\epsilon-1)(\epsilon+\nu)u_{j-1}^n + (1/4)\left[\epsilon(4\epsilon-3)+2(2\epsilon-1)\nu+(2-\epsilon)\nu^2\right]\Delta x(u_x)_{j-1}^n \\ &+ 4(1-\epsilon)\nu u_j^n + (1/2)\left[4\epsilon^2-5\epsilon+2+(\epsilon-2)\nu^2\right]\Delta x(u_x)_j^n \\ &+ 2(1-\epsilon)(\epsilon-\nu)u_{j+1}^n + (1/4)\left[\epsilon(4\epsilon-3)-2(2\epsilon-1)\nu+(2-\epsilon)\nu^2\right]\Delta x(u_x)_{j+1}^n. \end{aligned} \tag{7.2}$$

Eqs. (7.1) and (7.2) represent a system of two discrete equations for each $(j,n) \in \Omega$. It will be shown in this section that a solution to a pair of particular partial differential equations (PDE's) will satisfy the above discrete equations under certain limiting conditions. One of these PDE's is Eq. (2.22).

To proceed, let $\tilde{u}(x,t)$ and $\tilde{v}(x,t)$ be two smooth functions. let

$$\tilde{w}(x,t) \stackrel{\text{def}}{=} \tilde{v}(x,t) - \frac{\partial \tilde{u}(x,t)}{\partial x}. \tag{7.3}$$

For all $(j,n) \in \Omega$, let $\tilde{u}_j^n \stackrel{\text{def}}{=} \tilde{u}(x_j,t^n)$, $\tilde{v}_j^n \stackrel{\text{def}}{=} \tilde{v}(x_j,t^n)$, and $\tilde{w}_j^n \stackrel{\text{def}}{=} \tilde{w}(x_j,t^n)$. Let

$$\begin{aligned} [DE1]_j^n \stackrel{\text{def}}{=} \frac{\tilde{u}_j^{n+1}}{\Delta t} - \frac{1}{4\Delta t}\Big\{&(1+\nu)[\epsilon+(2-\epsilon)\nu]\tilde{u}_{j-1}^n + (1/2)(1-\nu^2)(\epsilon+\nu)\Delta x\tilde{v}_{j-1}^n \\ &+ 2(2-\epsilon)(1-\nu^2)\tilde{u}_j^n - \nu(1-\nu^2)\Delta x\tilde{v}_j^n \\ &+ (1-\nu)[\epsilon-(2-\epsilon)\nu]\tilde{u}_{j+1}^n - (1/2)(1-\nu^2)(\epsilon-\nu)\Delta x\tilde{v}_{j+1}^n\Big\}, \end{aligned} \tag{7.4}$$

and

$$
\begin{aligned}
&[DE2]_j^n \stackrel{\text{def}}{=} \tilde{v}_j^{n+1} \\
&- (2/\Delta x)(\epsilon-1)(\epsilon+\nu)\tilde{u}_{j-1}^n - (1/4)\left[\epsilon(4\epsilon-3)+2(2\epsilon-1)\nu+(2-\epsilon)\nu^2\right]\tilde{v}_{j-1}^n \\
&- (4/\Delta x)(1-\epsilon)\nu\tilde{u}_j^n - (1/2)\left[4\epsilon^2-5\epsilon+2+(\epsilon-2)\nu^2\right]\tilde{v}_j^n \\
&- (2/\Delta x)(1-\epsilon)(\epsilon-\nu)\tilde{u}_{j+1}^n - (1/4)\left[\epsilon(4\epsilon-3)-2(2\epsilon-1)\nu+(2-\epsilon)\nu^2\right]\tilde{v}_{j+1}^n.
\end{aligned}
\tag{7.5}
$$

Note that $[DE1]_j^n$ and $[DE2]_j^n$ are defined such that Eqs. (7.1) and (7.2) are equivalent to

$$[DE1]_j^n = 0, \tag{7.6}$$

and

$$[DE2]_j^n = 0, \tag{7.7}$$

respectively, if, for all $(j,n) \in \Omega$, $\tilde{u}_j^n$ and $\tilde{v}_j^n$ in Eqs. (7.6) and (7.7) are replaced by u_j^n and $(u_x)_j^n$, respectively.

Substituting Taylor series expansions of $\tilde{u}_j^{n+1}$, $\tilde{u}_{j\pm1}^n$, $\tilde{v}_j^{n+1}$ and $\tilde{v}_{j\pm1}^n$ about (x_j,t^n) into Eqs. (7.4) and (7.5), one concludes that, for *any* smooth functions $\tilde{u}(x,t)$ and $\tilde{v}(x,t)$,

$$[DE1]_j^n - [PDE]_j^n = [ER1]_j^n, \tag{7.8}$$

and

$$[DE2]_j^n - 4\epsilon(1-\epsilon)\tilde{w}_j^n = [ER2]_j^n. \tag{7.9}$$

Here, *assuming all derivatives are evaluated with* $x = x_j$ *and* $t = t^n$,

$$[PDE]_j^n \stackrel{\text{def}}{=} \frac{\partial\tilde{u}}{\partial t} + a\frac{\partial\tilde{u}}{\partial x}, \tag{7.10}$$

$$
\begin{aligned}
[ER1]_j^n \stackrel{\text{def}}{=}\ & \frac{\Delta t}{2}\left(\frac{\partial^2\tilde{u}}{\partial t^2} - a^2\frac{\partial^2\tilde{u}}{\partial x^2}\right) + \frac{\epsilon}{4}\left[\frac{(\Delta x)^2}{\Delta t} - a^2\Delta t\right]\frac{\partial\tilde{w}}{\partial x} \\
&+ \frac{(\Delta t)^2}{6}\left(\frac{\partial^3\tilde{u}}{\partial t^3} + a^3\frac{\partial^3\tilde{u}}{\partial x^3}\right) + \frac{a\left[(\Delta x)^2 - a^2(\Delta t)^2\right]}{24}\left(\frac{\partial^3\tilde{u}}{\partial x^3} - 3\frac{\partial^2\tilde{w}}{\partial x^2}\right) \\
&+ \frac{(\Delta t)^3}{24}\left(\frac{\partial^4\tilde{u}}{\partial t^4} - a^4\frac{\partial^4\tilde{u}}{\partial x^4}\right) + \frac{(\Delta x)^3}{24}\epsilon\left(\frac{\Delta x}{\Delta t} - a\right)\frac{\partial^3\tilde{w}}{\partial x^3} \\
&+ \frac{1}{48}\left[\epsilon\frac{(\Delta x)^2}{\Delta t} - 2a^2\Delta t\right]\left[(\Delta x)^2 - a^2(\Delta t)^2\right]\frac{\partial^4\tilde{u}}{\partial x^4} + O\left[(\Delta x)^4\right] + O\left[\Delta t(\Delta x)^3\right] \\
&+ O\left[(\Delta t)^2(\Delta x)^2\right] + O\left[(\Delta t)^4\right] + \epsilon\left\{O\left[\Delta t(\Delta x)^3\right] + \frac{\Delta x}{\Delta t}O\left[(\Delta x)^4\right]\right\},
\end{aligned}
\tag{7.11}
$$

and

$$
\begin{aligned}
[ER2]_j^n \stackrel{\text{def}}{=} & \Delta t\left[\frac{\partial \tilde{w}}{\partial t}+\frac{\partial}{\partial x}\left(\frac{\partial \tilde{u}}{\partial t}+a\frac{\partial \tilde{u}}{\partial x}\right)+(2\epsilon-1)a\frac{\partial \tilde{w}}{\partial x}\right] \\
& -\frac{1}{4}\left[\epsilon(4\epsilon-3)(\Delta x)^2+(2-\epsilon)a^2(\Delta t)^2\right]\frac{\partial^2 \tilde{w}}{\partial x^2} \\
& +\frac{(\Delta t)^2}{2}\left[\frac{\partial^2 \tilde{w}}{\partial t^2}+\frac{\partial}{\partial x}\left(\frac{\partial^2 \tilde{u}}{\partial t^2}-a^2\frac{\partial^2 \tilde{u}}{\partial x^2}\right)\right]-\frac{\epsilon\left[(4\epsilon-1)(\Delta x)^2-3a^2(\Delta t)^2\right]}{12}\frac{\partial^3 \tilde{u}}{\partial x^3} \\
& +\frac{(\Delta t)^3}{6}\left[\frac{\partial^3 \tilde{w}}{\partial t^3}+\frac{\partial}{\partial x}\left(\frac{\partial^3 \tilde{u}}{\partial t^3}+a^3\frac{\partial^3 \tilde{u}}{\partial x^3}\right)\right]+\frac{(2\epsilon-1)a\Delta t(\Delta x)^2}{6}\frac{\partial^3 \tilde{w}}{\partial x^3} \\
& +\frac{a\Delta t\left[\epsilon(\Delta x)^2-a^2(\Delta t)^2\right]}{6}\frac{\partial^4 \tilde{u}}{\partial x^4}+O\left[\Delta t(\Delta x)^3\right]+O\left[(\Delta t)^2(\Delta x)^2\right]+O\left[(\Delta t)^4\right] \\
& +\epsilon\Big\{O\left[(\Delta x)^4\right]+O\left[\Delta t(\Delta x)^3\right]+O\left[(\Delta t)^2(\Delta x)^2\right]\Big\}+\epsilon^2 O\left[(\Delta x)^4\right].
\end{aligned}
\tag{7.12}
$$

Because we assume that the parameter ϵ can vary with Δt and Δx, it is not treated as a constant in the derivation of Eqs. (7.8)–(7.12). In other words, *all the order-of-magnitude quantities given in Eqs. (7.11) and (7.12) are independent of* ϵ.

The significance of Eqs.(7.8) and (7.9) will be discussed under different assumptions about ϵ. Assuming that $\epsilon(1-\epsilon)\neq 0$, Eq. (7.9) can be rewritten as

$$[DE2']_j^n-\tilde{w}_j^n=[ER2']_j^n, \tag{7.13}$$

where

$$[DE2']_j^n \stackrel{\text{def}}{=} [DE2]_j^n/[4\epsilon(1-\epsilon)], \quad \text{and} \quad [ER2']_j^n \stackrel{\text{def}}{=} [ER2]_j^n/[4\epsilon(1-\epsilon)]. \tag{7.14}$$

The following comments are made for Eqs. (7.8) and (7.13):

(a) For all $(j,n)\in\Omega$,

$$[PDE]_j^n=0, \qquad \text{and} \qquad \tilde{w}_j^n=0, \tag{7.15}$$

if $u=\tilde{u}(x,t)$ and $v=\tilde{v}(x,t)$ satisfy both Eq. (2.22) and

$$v-\frac{\partial u}{\partial x}=0. \tag{7.16}$$

(b) Note that $u_j^n=\tilde{u}_j^n$ and $(u_x)_j^n=\tilde{v}_j^n$ satisfy Eqs. (7.1) and (7.2) if $\tilde{u}_j^n$ and $\tilde{v}_j^n$ satisfy Eqs. (7.6) and (7.7). As a result, $[DE1]_j^n$ and $[DE2']_j^n$ can be considered as the a-ϵ scheme's approximations for $[PDE]_j^n$ and $\tilde{w}_j^n$, respectively. Eqs. (7.8) and (7.13) then

state that $[ER1]_j^n$ and $[ER2']_j^n$ are the errors of these approximations, respectively. Let $[ER1]_j^n \to 0$ and $[ER2']_j^n \to 0$ as $\Delta t, \Delta x \to 0$. Then $[DE1]_j^n$ and $[DE2']_j^n$, respectively, approach $[PDE]_j^n$ and $\tilde{w}_j^n$ as $\Delta t, \Delta x \to 0$. Note that *the limits* $[PDE]_j^n$ *and* $\tilde{w}_j^n$ *are independent of* Δt *and* Δx.

(c) With the aid of the observations made in (a) and (b), one concludes that Eqs. (7.1) and (7.2) may be considered as the discrete approximations of Eqs. (2.22) and (7.16), respectively, with the understanding that u_j^n and $(u_x)_j^n$ are the discrete counterparts of u and v, respectively. *Note that Eq. (7.16), i.e.,* $v = \partial u/\partial x$, *is consistent with the fact that* $(u_x)_j^n$ *is the numerical analogue of the value of* $\partial u/\partial x$ *at the mesh point* (x_j, t^n).

(d) Let $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ be a solution to Eqs. (2.22) and (7.16). Then Eq. (7.3) implies that $\tilde{w}(x,t) = 0$. Moreover, it can be shown that

$$\frac{\partial^\ell \tilde{u}(x,t)}{\partial t^\ell} - (-1)^\ell a^\ell \frac{\partial^\ell \tilde{u}(x,t)}{\partial x^\ell} = 0, \qquad \ell = 1, 2, 3, \ldots . \tag{7.17}$$

As a result, Eqs. (7.11), (7.12) and (7.14) imply that

$$\begin{aligned}
[ER1]_j^n &= \frac{a\left[(\Delta x)^2 - a^2(\Delta t)^2\right]}{24}\frac{\partial^3 \tilde{u}}{\partial x^3} \\
&+ \frac{1}{48}\left[\epsilon\frac{(\Delta x)^2}{\Delta t} - 2a^2\Delta t\right]\left[(\Delta x)^2 - a^2(\Delta t)^2\right]\frac{\partial^4 \tilde{u}}{\partial x^4} + O\left[(\Delta x)^4\right] + O\left[\Delta t(\Delta x)^3\right] \\
&+ O\left[(\Delta t)^2(\Delta x)^2\right] + O\left[(\Delta t)^4\right] + \epsilon\left\{O\left[\Delta t(\Delta x)^3\right] + \frac{\Delta x}{\Delta t}O\left[(\Delta x)^4\right]\right\},
\end{aligned} \tag{7.18}$$

and

$$\begin{aligned}
[ER2']_j^n &= -\frac{\left[(4\epsilon - 1)(\Delta x)^2 - 3a^2(\Delta t)^2\right]}{48(1-\epsilon)}\frac{\partial^3 \tilde{u}}{\partial x^3} + \frac{a\Delta t\left[\epsilon(\Delta x)^2 - a^2(\Delta t)^2\right]}{24\epsilon(1-\epsilon)}\frac{\partial^4 \tilde{u}}{\partial x^4} \\
&+ \frac{1}{4\epsilon(1-\epsilon)}\left\{O\left[\Delta t(\Delta x)^3\right] + O\left[(\Delta t)^2(\Delta x)^2\right] + O\left[(\Delta t)^4\right]\right\} \\
&+ \frac{1}{4(1-\epsilon)}\left\{O\left[(\Delta x)^4\right] + O\left[\Delta t(\Delta x)^3\right] + O\left[(\Delta t)^2(\Delta x)^2\right] + \epsilon O\left[(\Delta x)^4\right]\right\}.
\end{aligned} \tag{7.19}$$

As in Eqs. (7.10)–(7.12), the derivatives in Eqs. (7.18) and (7.19) are evaluated with $x = x_j$ and $t = t^n$.

(e) Again, let $u = \tilde{u}(x,t)$ and $v = \tilde{v}(x,t)$ be a solution to Eqs. (2.22) and (7.16). Then, by definition, $[DE1]_j^n$ and $[DE2']_j^n$ are the truncation errors of Eqs. (7.1) and (7.2) with respect to the above PDE's [24, p.20]. Futhermore, we have $[PDE]_j^n = 0$ and $\tilde{w}_j^n = 0$. Thus

$$[DE1]_j^n = [ER1]_j^n, \qquad \text{and} \qquad [DE2']_j^n = [ER2']_j^n, \tag{7.20}$$

i.e., $[ER1]_j^n$ and $[ER2']_j^n$ given in Eqs.(7.18) and (7.19) are the truncation errors.

(f) Consider the special case in which ϵ does not vary with Δt and Δx. According to Eqs. (7.18), and (7.19), we have

$$[ER1]_j^n - \frac{\epsilon(\Delta x)^4}{48\Delta t}\left[\frac{\partial^4 \tilde{u}}{\partial x^4} + O(\Delta x)\right] \to 0 \quad \text{and} \quad [ER2']_j^n \to 0, \tag{7.21}$$

as $\Delta t, \Delta x \to 0$, regardless how the mesh is refined. Thus, $[ER1]_j^n \to 0$ as $\Delta t, \Delta x \to 0$ only if the mesh refinement is subjected to the condition:

$$(\Delta x)^4/\Delta t \to 0 \qquad \text{as} \qquad \Delta t, \Delta x \to 0. \tag{7.22}$$

Thus, the a-ϵ scheme is consistent with Eqs.(2.22) and (7.16) if the mesh refinement is subjected to Eq. (7.22).

Note that, by using Eqs. (7.8), (7.13) and (7.21), it can be shown that the a-ϵ scheme is consistent with Eq. (7.16) and the modified equation [25-27]:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + \frac{\epsilon(\Delta x)^4}{48\Delta t}\frac{\partial^4 u}{\partial x^4} + \frac{\epsilon}{\Delta t}O\left[(\Delta x)^5\right] = 0, \tag{7.23}$$

if Eq. (7.22) is not satisfied. Eq. (7.23) differs from Eq. (2.22) in the presence of a leading diffusion term and other higher-order terms. As a result, for a constant ϵ, the a-ϵ scheme becomes more diffusive as the ratio $(\Delta x)^4/\Delta t$ increases.

(g) In the general case in which $\epsilon = \epsilon(\Delta t, \Delta x)$, $\Delta t > 0$ and $\Delta x > 0$, consistency of the a-ϵ scheme is dependent on the behavior of $\epsilon(\Delta t, \Delta x)$, as $\Delta t, \Delta x \to 0$. As an example, let ϵ_2 and ϵ_1 be two constants such that $\epsilon_2 > \epsilon_1 > 0$. Let

$$\epsilon_2 > |\epsilon/\Delta t| > \epsilon_1 \qquad \text{as} \qquad \Delta t, \Delta x \to 0. \tag{7.24}$$

Then Eqs.(7.18) and (7.19) imply that the a-ϵ scheme is consistent with Eqs. (2.22) and (7.16). As another example, let $\epsilon_0 > 0$ be a constant, and

$$\frac{\epsilon \Delta x}{\Delta t} \to \epsilon_o \qquad \text{as} \qquad \Delta t/\Delta x \to 0. \tag{7.25}$$

Then the a-ϵ scheme is also consistent with Eqs. (2.22) and (7.16) if the mesh refinement is subjected to the condition

$$\Delta t/\Delta x \to 0 \qquad \text{as} \qquad \Delta t, \Delta x \to 0. \tag{7.26}$$

This completes the discussion on Eqs.(7.8) and (7.13). For either $\epsilon = 0$ or $\epsilon = 1$, the second term on the left side of Eq. (7.9) vanishes. For $\epsilon = 0$, which is a special case of the a-μ scheme with $\mu = 0$, it is shown in [1] how consistency of the a-ϵ scheme can be studied by recasting Eq. (7.9) into another form. By applying the same technique, consistency of the scheme with $\epsilon = 1$ can also be studied.

8. Numerical Results

In [1], numerical solutions of Eq. (2.1) generated by the MacCormack [4, p.102], the Leapfrog/DuFort-Frankel [4, p.161], and the a-μ schemes are compared with the corresponding analytical solutions for different values of physical coefficients, mesh parameters and total running times. These comparisons show that the a-μ scheme is far superior to the Leapfrog/DuFort-Frankel scheme in accuracy, and has a substantial advantage over the MacCormack scheme in both accuracy and stability.

In this section, accuracy of both the Euler and the Navier-Stokes solvers will be evaluated numerically using a shock tube problem suggested by Sod [28]. Because the a-ϵ scheme may be considered as a special case of the Euler solver, no separate numerical evaluation for the a-ϵ scheme will be given.

Let the specific heat ratio $\gamma = 1.4$. At $t = 0$, let (i) $(\rho, v, p) = (1, 0, 1)$, i.e., $(u_1, u_2, u_3) = (1, 0, 2.5)$, if $x < 0$, and (ii) $(\rho, v, p) = (0.125, 0, 0.1)$, i.e., $(u_1, u_2, u_3) = (0.125, 0, 0.25)$, if $x > 0$. For all $(j, n) \in \Omega$, let $x_j = j\Delta x$, and $t^n = n\Delta t$. Then (i)

$$\left((u_1)_j^0, (u_2)_j^0, (u_3)_j^0\right) = \begin{cases} (1, 0, 2.5), & \text{if } j = -1/2, -3/2, \ldots; \\ (0.125, 0, 0.25), & \text{if } j = 1/2, 3/2, \ldots, \end{cases} \tag{8.1}$$

and (ii) $(u_{mx})_j^0 = 0$, $j = \pm 1/2, \pm 3/2, \ldots$, for $m = 1, 2, 3$. Hereafter, we assume $n \geq 0$.

The above initial conditions coupled with several equations given in Secs. 4 and 5, imply that, for both the Euler and the Navier-Stokes solvers, $(u_m)_j^n$ is a constant and $(u_{mx})_j^n = 0$ in two separate regions that are defined by $j \leq -(n + 1/2)$ and $j \geq (n + 1/2)$, respectively. Thus, one needs to evaluate the above variables only if $(n + 1/2) > |j|$.

Without exception, $\Delta x = 0.01$ is assumed in this section. Also, all numerical results will be compared with the exact weak solution at $t = 0.2$. Because, at $t = 0.2$, the effect of the initial discontinuity at $t = 0$ is far from reaching the spatial regions defined by $x > 0.5$ and $x < -0.5$, respectively, numerical computations will be simplified by assuming that, for all n with $t^n \leq 0.2$, (i)

$$\left((u_1)_j^n, (u_2)_j^n, (u_3)_j^n\right) = \begin{cases} (1, 0, 2.5), & \text{if } x_j < -0.5; \\ (0.125, 0, 0.25), & \text{if } x_j > 0.5, \end{cases} \tag{8.2}$$

and (ii) $(u_{mx})_j^n = 0$ if $|x_j| > 0.5$. Because $\Delta x = 0.01$, the above assumptions imply that the computation domain can be limited to $|j| \leq 50$.

In the initial evaluation, we consider the Euler marching scheme defined by Eqs. (4.28) and (4.32). numerical results (triangles) obtained assuming $\Delta t = 0.004$ and $\epsilon = 1/2$ are compared with the exact solutions (solid lines) in Fig. 6. Because each marching step advances the solution from t to $t + \Delta t/2$, these results at $t = 0.2$ are obtained after 100 steps. Also it can be estimated that $CFL \doteq 0.88$ where CFL is defined to be the maximum

value of $(|v|+|c|)\Delta t/\Delta x$. Thus the numerical calculation is carried out within the stability limits given by Eq. (4.49). Note that the agreements between the numerical results and the exact solutions are excellent. Particularly, shock discontinuity is resolved almost within one mesh interval, and contact discontinuity is resolved in four mesh intervals. Also, there are only slight numerical overshoots and/or oscillations near these discontinuties.

According to the discussions given in Secs. 3, 4, and 6, the Euler solver behaves like the Leapfrog scheme if $\epsilon = 0$, and like the Lax scheme if $\epsilon = 1$. The former is free from numerical diffusion while the latter is highly diffusive. *The current scheme with* $\epsilon = 1/2$ *can be considered as a scheme midway between the above two celebrated schemes.*

Moreover, the last term on the right side of Eq. (4.32) vanishes if $\epsilon = 1/2$. The remaining term is simply a central-difference approximation for $(u_{mx})_j^n$.

Let Eq. (4.32) be modified with $(u_{mx}^c)_j^n$ being replaced by $(u_{mx}^{w_o})_j^n$ (see eqs. (4.51) and 4.54)). Again assuming that $\Delta t = 0.004$ and $\epsilon = 1/2$, the numerical results obtained with $\alpha = 1$, $\alpha = 2$, and $\alpha = 3$, respectively, are given in Figs. 7–9. The effectiveness of the above modification as a tool to surpress numerical wiggles near discontinuities is apparent. It was explained in Sec. 4 why this modification does not cause the smearing of shock discontinuities. Furthermore, the modification has no discernable effect on the smooth part of the solution. Because $(u_{mx}^{w_o})_j^n = (u_{mx}^c)_j^n$ if $\alpha = 0$, in the following discussion, it should be understood that the above modification is turned off if $\alpha = 0$.

Note that the results shown in Figs. 6–9 can be generated using the sample program listed at the end of the present paper. It is coded assuming $\epsilon = 0.5$. The value of the input parameter *ic* is equal to that of α.

Let $\alpha = 0$ and $\Delta t = 0.004$. The numerical results obtained with $\epsilon = 0.1$, $\epsilon = 0.3$, $\epsilon = 0.7$, and $\epsilon = 0.9$, respectively, are given in Figs. 10–13. Note that the case with $\epsilon = 0.5$ was presented in Fig. 6. For $\epsilon = 0.1$, because the scheme has very small numerical diffusion, pronounced wiggles appear in large regions near discontinuities. However, because of the same reason, *the smooth part of the solution is highly accurate.* For $\epsilon = 0.3$, the wiggles are less pronounced and appear in more limited regions. Also the smooth part of the solution becomes less accurate. As the value of ϵ increases, the wiggles disappear and the solution becomes more diffusive. The solution obtained with $\epsilon = 0.7$ is excellent except that, compared with the case with $\epsilon = 0.5$, it requires one more mesh interval to resolve the contact discontinuity. The results shown in Figs. 6 and 10–13 are consistent with the theoretical prediction that the Euler solver becomes progressively diffusive as the value of ϵ increases from 0 to 1.

Figs. 14 and 11 are both generated using the same conditions except that $\alpha = 2$ in the former, while $\alpha = 0$ in the latter. Note that the wiggles almost completely disappear in Fig. 14.

The above numerical results are all generated assuming $\Delta t = 0.004$. The numerical results shown in Fig. 15 are generated with $\Delta t = 0.002$, $\epsilon = 1/2$, and $\alpha = 0$. It is obtained after 200 steps and $CFL \doteq 0.44$. A comparison between Figs. 6 and 15 reveals that the current solver is more diffusive at a smaller *CFL*. Note that, by considering the

truncation error, it was shown in Sec. 7 that, for constant ϵ and Δx, the a-ϵ scheme becomes more diffusive as Δt decreases. A similar conclusion can also be reached by studying the amplification factors given in Eqs. (3.12) and (3.13). Because the Euler solver is a straightforward extension of the a-ϵ scheme, one would expect that the former also behaves similarly. Fig. 16 shares with Fig. 15 the same defining conditions except that $\alpha = 1$ for the former.

The numerical results shown in Figs. 17 and 18 are generated assuming $\Delta t = 0.0004$ (i.e., $CFL \doteq 0.088$). Note that it takes 1000 marching steps to advance the solution to $t = 0.2$. Other defining conditions for these figures are identical to those for Figs. 6 and 7, respectively. As expected, the results obtained with low CFL are much more diffusive than those obtained with CFL closer to 1 (see Figs. 6 and 7). Also, as the value of CFL decreases, the diffusive effect of replacing $\alpha = 0$ with $\alpha = 1$ becomes more discerable even in the smooth part of the solution. In other words, numerical diffusion introduced by replacing $\alpha = 0$ with $\alpha > 0$, is greater when CFL is small.

To modify the above Euler solver such that it can compensate for the observed effect of increasing numerical diffusion as Δt decreases, in the following discussions, we shall consider the more general marching scheme defined by Eqs. (4.28) and (4.36). The parameter $(\hat{\epsilon})_j^n$ in Eq. (4.36) will be dependent on the mesh position (j, n) and the ratio $\Delta t/\Delta x$. Moreover, the term $(u_{mx}^c)_j^n$ in Eq. (4.36) will be replaced by $(u_{mx}^w)_j^n$, which is defined in Eq. (4.59). The weight factor β will also be dependent on (j, n) and $\Delta t/\Delta x$.

To proceed, let

$$\zeta(x) \stackrel{\text{def}}{=} x \exp(1 - x), \qquad 0 \le x \le 1. \tag{8.3}$$

Because ζ is an increasing function within its domain, we have

$$\zeta(x) \le \zeta(1) = 1, \qquad 0 \le x \le 1. \tag{8.4}$$

For all $(j, n) \in \Omega$, let

$$(\hat{\epsilon})_j^n = b\, \zeta((\hat{\nu}_{max})_j^n), \tag{8.5}$$

and

$$(u_{mx}^w)_j^n = W\left((u_{mx-})_j^n, (u_{mx+})_j^n; \alpha, \sqrt{(\hat{\nu}_{max})_j^n}\right), \tag{8.6}$$

where $(\hat{\nu}_{max})_j^n$ is defined in Eq. (4.50), and b and α are constants that do not vary from one mesh point to another. Because $(\hat{\epsilon}_m)_j^n = (\hat{\epsilon})_j^n$, $m = 1, 2, 3$, is assumed in Eq. (4.36), Eqs. (4.49), (8.4) and (8.5) imply that (i) $(\hat{\nu}_{max})_j^n$ is in the domain of $\zeta(x)$, and (ii) $0 \le b \le 1$.

Note that $(\hat{\nu}_{max})_j^n$ is proportional to $\Delta t/\Delta x$. Thus, Eqs. (8.3) and (8.5) imply that $(\hat{\epsilon})_j^n$ is an increasing function of $\Delta t/\Delta x$, i.e., it decreases as Δt decreases if other parameters are held constant. Because numerical diffusion decreases as $(\hat{\epsilon})_j^n$ decreases, with other factors being equal, the replacement of a constant ϵ with $(\hat{\epsilon})_j^n$ has an effect in reducing numerical diffusion as Δt decreases. This effect will compensate for the observed opposite effect on numerical diffusion as Δt decreases with ϵ, Δx, *and the total running time being*

held constant. Furthermore, because $\zeta(x)/x \to 0$ as $x \to 0$, Eqs. (4.50) and (8.5) imply that

$$(\hat{\epsilon})_j^n \frac{\Delta x}{\Delta t} \to b\left(|\hat{v}_j^n| + |\hat{c}_j^n|\right) \qquad \text{as} \qquad \Delta t/\Delta x \to 0. \tag{8.7}$$

Eq. (8.7) is similar to the consistent condition given in Eq. (7.25).

Moreover, for a fixed α, $W(x_-, x_+; \alpha, \beta) \to (x_- + x_+)/2$ as $\beta \to 0$. This fact coupled with Eq. (4.53) implies that the numerical diffusion introduced as a result of replacing $(u_{mx}^c)_j^n$ with $(u_{mx}^w)_j^n$ will decrease as β decreases. Because $(\hat{\nu}_{max})_j^n$ is proportional to $\Delta t/\Delta x$, with other factors being equal, the replacement of $(u_{mx}^c)_j^n$ by $(u_{mx}^w)_j^n$ defined in Eq. (8.6), has an effect in reducing numerical diffusion as Δt decreases. This effect will compensate for the observed opposite effect on numerical diffusion as Δt decreases with α, β, Δx, *and the total running time being held constant*. Note that $W_o(x_-, x_+; \alpha)$ is a special case of $W(x_-, x_+; \alpha, \beta)$ with $\beta = 1$.

Assuming $\alpha = 1$ and $b = 0.5$, the numerical results shown in Figs. 19, 20 and 21 are generated with $\Delta t = 0.004\,(CFL \doteq 0.88)$, $\Delta t = 0.0004\,(CFL \doteq 0.088)$, and $\Delta t = 0.0001\,(CFL \doteq 0.022)$, respectively. Note that the results shown in Fig. 19 are almost identical to those shown in Fig. 7 which were generated assuming the same conditions but using a simpler marching scheme. However, the results shown in Fig. 20 are far less diffusive than their counterparts shown in Fig. 18. One can conclude from this comparison and the results shown in Fig. 21 that the current modified Euler solver is capable of generating accurate numerical solutions even for the case with a very small *CFL*.

In the above modified Euler scheme, $(\hat{\epsilon})_j^n$ and β are expressed as two special functions of $(\hat{\nu}_{max})_j^n$, respectively. They are only two among many possible choices. The investigation of other choices is a subject to be studied in the future.

The most general marching scheme presented in Sec. 4 is that defined by Eqs. (4.28) and (4.35). It requires several matrix multiplications at each mesh points and, therefore, is much more costly. Thus, its use is difficult to justify unless a substantial gain in accuracy can be made. How this most general marching scheme can be applied wisely is left for a future study.

This completes the numerical study of the Euler solver. We conclude this section with a numerical evaluation of the Navier-Stokes marching scheme defined by Eqs. (5.20) and (5.26). Again the initial conditions defined in Eq. (8.1) are assumed, and the numerical solutions are compared with *the exact weak solution of the Euler equations* at $t = 0.2$. The numerical results shown in Figs. 22–28 are generated assuming $\Delta t = 0.004$, $\Delta x = 0.01$, $\gamma = 1.4$, and $Pr = 0.72$. The value of the Prandtl number used here is that for air at standard conditions. The values of the Re_L for these figures are 2,000, 4,000, 6,000, 8,000, 10,000, 12,000, and 20,000, respectively.

From the results shown in these figures, one concludes that, for a high-Reynolds-number flow, the shock can be resolved within one mesh interval by the current Navier-Stokes solver. Also the contact discontinuity can be resolved within a few mesh intervals. *Note that these results are obtained without using any ad-hoc parameters or techniques.*

Because the Reynold number is inversely proportional to the physical viscosity, as expected, numerical overshoots and oscillations shown in these figures increase slightly as the values of the Reynolds number increase.

Furthermore, through repeated numerical experiments using different physical and mesh parameters, it is established that the current Euler solver is stable if, for all $(j,n) \in \Omega$,

$$0 \leq Re_L, \quad 0 \leq Pr, \quad \text{and} \quad (\hat{\nu}_{max})_j^n < 1 \tag{8.8}$$

However, because a Navier-Stokes problem is fundamentally an initial-value/boundary-value problem, the current explicit marching scheme obviously cannot model such a problem unless the boundary effect is small, i.e., when the contribution of the viscous terms to Eqs. (5.20) and (5.26) is small compared to that of the convection terms. In general, this implies that the current scheme is applicable only to high-Reynolds-number flows. Note that the Leapfrog/Dufort-Frankel and the a-μ schemes [1] also encounter a similar limitation in modelling Eq. (2.1).

Finally, note that the current Navier-Stokes solver with $Re_L = \infty$ (i.e., the physical viscosity vanishes) and $Pr = 0$ can be considered as a nonlinear extension of the inviscid a-μ scheme. Because the latter scheme is neutrally stable, generally one would expect that a nonlinear extension of such a scheme is unstable. However, it has been shown numerically that the current Navier-Stokes solver is stable even for the above limiting case as long as $(\hat{\nu}_{max})_j^n < 1$ for all $(j,n) \in \Omega$.

9. Conclusions and Discussions

Several key limitations of the finite difference, finite volume, finite element, and spectral methods were discussed in Sec 1. The method of space-time conservation element and solution element was conceived to overcome these limitations.

Using the a-μ scheme as an example, major differences between the current method and those mentioned above were explained in Sec. 2. This explicit scheme has the unusual property that its stability is limited only by the *CFL* condition, i.e., it is independent of μ. Also, it was shown that its amplification factors are identical to those of the Leapfrog scheme if $\mu = 0$, and to those of the DuFort-Frankel scheme if $a = 0$. These coincidences are rather unexpected because the a-μ scheme and the above classical schemes are derived from completely different perspectives, and the current scheme *does not* reduce to the above classical schemes in the limiting cases.

The inviscid a-μ scheme is neutrally stable and reversible in time. It is well known that a neutrally stable numerical analogue of Eq. (2.22) generally becomes unstable when it is extended to model the Euler equations. It is also obvious that a scheme that is reversible in time cannot model a physical problem that is irreversible in time, e.g., an inviscid flow problem involving shocks. Thus, the inviscid version was modified in Sec. 3 to form the a-ϵ scheme. This new scheme has the unusual property that numerical diffusion is controlled by an adjustable parameter ϵ. As a matter of fact, for all wavelengths, numerical diffusion can be *uniformly* bounded from above by an arbitrary small number by choosing a small enough ϵ. Stability of the a-ϵ scheme is limited by the *CFL* condition and $0 \leq \epsilon \leq 1$. Moreover, if $\epsilon = 0$, the amplification factors of the a-ϵ scheme are identical to those of the Leapfrog scheme, which has no numerical diffusion. On the other hand, if $\epsilon = 1$, they unexpectedly become identical to each other and to the amplification factor of the highly diffusive Lax scheme. Note that, because the Lax scheme is very diffusive and uses a mesh that is staggered in time, a two-level scheme using such a mesh is often associated with a highly diffusive scheme. The a-ϵ scheme, which also uses a mesh staggered in time, demonstrates that such a scheme could be free from numerical diffusion.

In Sec. 4, the a-ϵ scheme was extended to become an Euler solver. This solver has the unusual property that numerical diffusion at any mesh point (j, n) can be controlled by a set of local parameters $(\hat{\epsilon}_m)_j^n$, $m = 1, 2, 3$. As in the a-ϵ scheme, stability of the Euler solver is limited by the *CFL* condition and the requirement that, for all (j, n), $0 \leq (\hat{\epsilon}_m)_j^n \leq 1$, $m = 1, 2, 3$. Note that an Euler solver using a mesh staggered in time is usually highly diffusive for a small *CFL* number. It was shown in Sec. 8 that the current solver is an exception. It can generate highly accurate shock tube solutions with the *CFL* number ranging from 0.88 to 0.022.

In Sec. 5, the a-μ scheme was extended to become a Navier-Stokes solver. Stability of this *explicit* solver is also limited only by the *CFL* condition. Despite the fact that it does not use (i) *any* techniques related to the high-resolution upwind methods, and (ii) *any* ad hoc parameter, it was shown in Sec. 8 that the current solver is capable of generating

highly accurate shock tube solutions. Particularly, shock discontinuites can be resolved within one mesh interval.

A summary of the key results of the present work has been given. Behind these results is a continuous effort to maintain the simplicity, generality, and accuracy of the current method. This effort is summarized in the following remarks:

(a) *Simplicity*. The current numerical framework rests upon only two basic building blocks, i.e., the space-time conservation and solution elements. It uses only local discrete variables. Also, the set of discrete variables in any one of the numerical equations to be solved is associated with a single SE or a few immediately neighboring SE's. Thus, local flexibility is preserved and one needs only to deal with a very sparse matrix. Moreover, flux evaluation at an interface separating two CE's requires no interpolation or extrapolation. Nor does it require the use of an ad hoc flux model. Finally, partly because no characteristics-based techniques are used, a numerical scheme can be constructed by using only the simplest approximation techniques.

(b) *Generality*. A guiding principle in the design of the current method is to limit the use of special assumptions or techniques that would restrict its use in more general situations. Thus we do not use characteristics-based techniques, and we try to avoid using ad hoc techniques.

(c) *Accuracy*. Because (i) a physical solution of the conservation laws may involve shocks or high-gradient regions, and (ii) an accurate numerical simulation of such a solution is difficult to obtain without enforcing flux conservation, the current method requires that *a numerical solution satisfies (i) the differential form of the conservation laws uniformly within an SE, and (ii) the integral form over any space-time region that is the union of any combination of CE's*. In addition, accuracy of the current method is aided by treating both $(u_m)_j^n$ and $(u_{mx})_j^n$ as independent variables, instead of expressing $(u_{mx})_j^n$ as a finite-difference approximation involving $(u_m)_j^n$'s of neighboring mesh points. The latter approach may result in poor accuracy in a high-gradient region. Also, accuracy is enhanced by the fact that the flux at an interface separating two CE's is evaluated without interpolation or extrapolation. Moreover, because flux conservation is fundamentally a property in space-time, the current unified treatment of space and time may also contribute to a more accurate simulation of the conservation laws.

As a result of its simplicity and generality, the current framework is also *highly flexible*. This flexibility will be demonstrated in the following discussion on how to discretize steady-state problems using the current method. In this discussion, we shall also address the important issue of boundary-condition implementation.

As a vehicle of demonstration, we consider a dimensionless form of the 2-D steady incompressible Navier-Stokes equations with constant viscosity coefficient [4]. Without any loss of generality, we can assume that the mass density = 1. Let x and y be the first and second coordinates, respectively, of a 2-D Euclidean space E_2. Let u_1 and u_2 be the x- and y-velocities, respectively. Let u_3 be the static presssure. Let Re_L be the Reynolds

number. Let

$$f_1^x \stackrel{\text{def}}{=} u_1, \qquad f_1^y \stackrel{\text{def}}{=} u_2, \tag{9.1}$$

$$f_2^x \stackrel{\text{def}}{=} (u_1)^2 + u_3 - \frac{2}{Re_L}\frac{\partial u_1}{\partial x}, \tag{9.2}$$

$$f_2^y = f_3^x \stackrel{\text{def}}{=} u_1 u_2 - \frac{1}{Re_L}\left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\right), \tag{9.3}$$

and

$$f_3^y \stackrel{\text{def}}{=} (u_2)^2 + u_3 - \frac{2}{Re_L}\frac{\partial u_2}{\partial y}. \tag{9.4}$$

Then the Navier-Stokes equations can be expressed as

$$\frac{\partial f_m^x}{\partial x} + \frac{\partial f_m^y}{\partial y} = 0, \qquad m = 1, 2, 3. \tag{9.5}$$

The integral form of Eq. (9.5) in E_2 is Eq. (4.6) with $\vec{h}_m = (f_m^x, f_m^y)$, $m = 1, 2, 3$, being the mass, x-momentum, and y-momentum flux current density vectors, respectively. Note that, with the understanding that the current E_2 is no longer a space-time, $S(V)$ and $d\vec{s}$ have the same definitions as given in Sec. 2.

Consider the rectangular computational domain depicted in Fig. 29(a). The upper and lower boundaries are in contact with stationary walls, while the left and right boundaries are the inlet and exit planes, respectively. The domain is first divided into rectangular regions (see fig. 29(a)). The center of a rectangular region is denoted by its coordinates (x_j, y_k). Each retangular region is further divided into four triangular regions (see Fig. 29(a)). Let the interiors of the triangular regions to the east, north, west and south of the center (x_j, y_k) be solution elements, and denoted by SE$(j, k; 1)$, SE$(j, k; 2)$, SE$(j, k; 3)$, and SE$(j, k; 4)$, respectively (see Fig. 29(b)-(e)). The CE's will be defined later.

Let $q = 1, 2, 3, 4$. For any $(x, y) \in$ SE$(j, k; q)$, let $u_m(x, y)$, $f_m^x(x, y)$, $f_m^y(x, y)$, and $\vec{h}_m(x, y)$ be approximated by $u_m^*(x, y\,; j, k; q)$, $f_m^{x*}(x, y\,; j, k; q)$, $f_m^{y*}(x, y\,; j, k; q)$, and $\vec{h}_m^*(x, y\,; j, k; q)$, respectively. The last four functions will be defined shortly. Let $(x_j^q,\, y_k^q)$ be a point in SE$(j, k; q)$ (see Fig. 29(b)-(e)), and

$$\begin{aligned} u_m^*(x, y\,; j, k; q) &\stackrel{\text{def}}{=} (u_m)_{j,k}^q + (u_{mx})_{j,k}^q (x - x_j^q) + (u_{my})_{j,k}^q (y - y_k^q) \\ &+ \frac{1}{2}(u_{mxx})_{j,k}^q (x - x_j^q)^2 + \frac{1}{2}(u_{myy})_{j,k}^q (y - y_k^q)^2 + (u_{mxy})_{j,k}^q (x - x_j^q)(y - y_k^q), \end{aligned} \tag{9.6}$$

where $(u_m)_{j,k}^q$, $(u_{mx})_{j,k}^q$, $(u_{my})_{j,k}^q$, $(u_{mxx})_{j,k}^q$, $(u_{myy})_{j,k}^q$, and $(u_{mxy})_{j,k}^q$ are constants in SE$(j, k; q)$. They are considered to be the numerical analogues of the values of u_m, $\partial u_m/\partial x$, $\partial u_m/\partial y$, $\partial^2 u_m/\partial x^2$, $\partial^2 u_m/\partial y^2$, and $\partial^2 u_m/\partial x \partial y$ at $(x_j^q,\, y_k^q)$, respectively.

Similarly, for $m = 1, 2, 3$ and $q = 1, 2, 3, 4$, let

$$f_m^{x*}(x, y; j, k; q) \stackrel{\text{def}}{=} (f_m^x)_{j,k}^q + (f_{mx}^x)_{j,k}^q (x - x_j^q) + (f_{my}^x)_{j,k}^q (y - y_k^q) + \frac{1}{2}(f_{mxx}^x)_{j,k}^q (x - x_j^q)^2 + \frac{1}{2}(f_{myy}^x)_{j,k}^q (y - y_k^q)^2 + (f_{mxy}^x)_{j,k}^q (x - x_j^q)(y - y_k^q), \tag{9.7}$$

and

$$f_m^{y*}(x, y; j, k; q) \stackrel{\text{def}}{=} (f_m^y)_{j,k}^q + (f_{mx}^y)_{j,k}^q (x - x_j^q) + (f_{my}^y)_{j,k}^q (y - y_k^q) + \frac{1}{2}(f_{mxx}^y)_{j,k}^q (x - x_j^q)^2 + \frac{1}{2}(f_{myy}^y)_{j,k}^q (y - y_k^q)^2 + (f_{mxy}^y)_{j,k}^q (x - x_j^q)(y - y_k^q). \tag{9.8}$$

Also, because $\vec{h}_m = (f_m^x, f_m^y)$, let

$$\vec{h}_m^*(x, y; j, k; q) \stackrel{\text{def}}{=} (f_m^{x*}(x, y; j, k; q), f_m^{y*}(x, y; j, k; q)). \tag{9.9}$$

The expansion coefficients in Eqs. (9.7) and (9.8) can be defined in terms of those in Eq. (9.6). As an example, consider $(f_{2xx}^x)_{j,k}^q$. It is the numerical analogue of the value of $\partial^2 f_2^x / \partial x^2$ at (x_j^q, y_k^q). By using Eq. (9.2), $\partial^2 f_2^x / \partial x^2$ can be expressed in terms of u_1, u_3, and their derivatives. With the understanding that the numerical analogue of any derivative of u_m higher than second order is set to zero, one can obtain the numerical version of the above relation by replacing each derivative with its numerical analogue. Using this numerical version, $(f_{2xx}^x)_{j,k}^q$ can be defined in terms of the expansion coefficients in in Eq. (9.6). Note that $(f_{mx})_j^n$ and $(f_{mt})_j^n$ were defined in a similar fashion (see Eqs. (4.14)–(4.17)).

From the above discussion, one concludes that the only independent discrete variables needed to be solved are the expansion coefficients in Eq. (9.6). Because $m = 1, 2, 3$, there are 18 unknowns for each SE.

We assume that the numerical solution satisfies Eq. (9.5) uniformly within an SE, i.e., for all $(x, y) \in \text{SE}(j, k; q)$,

$$\frac{\partial f_m^{x*}(x, y; j, k; q)}{\partial x} + \frac{\partial f_m^{y*}(x, y; j, k; q)}{\partial y} = 0, \qquad m = 1, 2, 3. \tag{9.10}$$

Because Eq. (9.10) is equivalent to $\nabla \cdot \vec{h}_m^* = 0$, Gauss' divergence theorem implies that the total flux of $\vec{h}_m^*$ leaving the boundary of *any region within an SE* vanishes.

As a result of Eqs. (9.7) and (9.8), for each m the expression on the left side of Eq. (9.10) is a polynomial of first order in $(x - x_j^q)$ and $(y - y_k^q)$. Eq. (9.10) requires that all three coefficients of this polynomial vanish. Because $m = 1, 2, 3$, Eq. (9.10) represents nine

conditions for each SE. Thus to match the number of conditions with that of unknowns, nine more conditions per SE are needed. One of many ways to fulfill this need is described in the following discussion.

First we consider an interior SE, i.e., an SE with each of its three edges being an interface separating two SE's. Hereafter, an edge separating two SE's will be referred to as an interior edge; while an edge bordering with the boundary of the computational domain will be referred to as an boundary edge. An example of an interior SE is that depicted in Fig. 29(f). Each interior edge is divided into two subsections. Thus (i) the edge BC is divided into BA′ and A′C, (ii) the edge CA is divided into CB′ and B′A, and (iii) the edge AB is divided into AC′ and C′B. Moreover, for each m, we shall assume that the net flux of $\vec{h}^*_m$ entering each subsection from both sides vanishes. Because $m = 1, 2, 3$, there are six interface flux conservation conditions for each interior edge. Moreover, because an interior edge is bordered by two SE's, each SE can be allocated only three net interface flux conditions for each interior edge. Because an interior SE has three interior edges, the need for nine extra conditions is fulfilled by the above interface flux conditions.

Next consider a boundary SE, i.e., an SE with at least one boundary edge. The edges DE and FD of the SE depicted in Fig. 29(g) are interior edges, while the edge EF is a part of a stationary wall. Again we divide the interior edges into two subsections, i.e., (i) DE is divided into DF′ and F′E, and (ii) FD is divided into FE′ and E′D. The interface flux conditions imposed on DE or FD are similar to those described earlier for other interior edges. As will be explained shortly, three boundary conditions will be imposed on a boundary edge, such as EF. Because three net interface conditions are allocated to each interior edge, adding the number of boundary conditions to the number of interface conditions results in the nine extra conditions each boundary SE needs. This conclusion is valid even if the SE has more than one boundary edge.

With the above preparations, the CE's may be defined as follows: An interior SE such as that depicted in Fig. 29(f) can be divided into four triangular regions, i.e., AC′B′, BA′C′, CB′A′, and A′B′C′. The union of each of these regions and its boundary, by definition, is a CE. On the other hand, a boundary SE such as that depicted in Fig. 29(g) can be divided into a triangular region DF′E′ and a quadrilateral region EFE′F′. The union of each of these two regions and its boundary, by definition, is also a CE.

As a result of the above definition, an interface separating two CE's may be of two different types. Those of the first type are located within an SE. They are exemplified by A′B′ in Fig. 29(f), and E′F′ in Fig. 29(g). Because $\vec{h}^*_m$ is continuous on the neighborhood of an interface of this type, the net flux of $\vec{h}^*_m$ entering such an interface vanishes. Interfaces of the second type are subsections separating two SE's. They are exemplified by AC′ in Fig. 29(f), and DF′ in Fig. 29(g). By assumption, the flux of $\vec{h}^*_m$ entering an interface of this type also vanishes. Because the interior of a CE is within an SE, according to a statement made following Eq. (9.10), the total flux of $\vec{h}^*_m$ leaving the boundary of a CE vanishes. Combining the results established above, one concludes that the total flux of $\vec{h}^*_m$ leaving the union of any combination of CE's vanishes. Note that the SE's and CE's defined here are substantially different from those depicted in Figs. 2 and 3. Furthermore,

it should be emphasized that it is possible to define two different sets of CE's with the same SE's and numerical conditions.

In the above discussions, we have assumed that three boundary conditions are imposed on each boundary edge. Using the boundary edge EF depicted in Fig. 29(g) as an example, we shall describe how these boundary conditions may be implemented.

The edge EF is a part of a stationary wall. Thus, we assume that

$$u_1 = 0, \qquad \text{and} \qquad u_2 = 0, \tag{9.11}$$

on EF. Because EF is aligned with the x-axis, we also have

$$\frac{\partial u_1}{\partial x} = 0, \tag{9.12}$$

on EF. Moreover, Eq. (9.12) coupled with the equation in Eq. (9.5) with $m = 1$ implies that

$$\frac{\partial u_2}{\partial y} = 0, \tag{9.13}$$

on EF. With the aid of Eqs. (9.11)–(9.13), Eqs. (9.1), (9.2) and (9.4) imply that

$$f_1^x = f_1^y = f_2^x - f_3^y = 0, \tag{9.14}$$

on EF. As a result, we may assume that *the simple average of each of* f_1^{x*}, f_1^{y*} *and* $(f_2^{x*} - f_3^{y*})$ *over the length of EF vanishes.* The above assumption imposes three boundary conditions on EF. Note that the net numerical mass flux entering the upper wall through EF vanishes if the average of f_1^{y*} over EF vanishes.

At this juncture, it should be emphasized that the boundary conditions proposed above, as in the case of the interface flux conservation conditions, are conditions over a domain. For the special case under consideration, the domain is the entire length of EF. In general, however, the domain may be a subsection of EF. As an example, the edge EF may be divided into two subsections. We may require that (i) the average of f_1^{x*} over each of these subsections vanishes, and (ii) the average of f_1^{y*} over the entire length of EF vanishes. The resulting three conditions may replace the three conditions proposed earlier. Obviously, there are many other alternatives. Futhermore, as need arises, one can easily impose more boundary conditions over a boundary edge.

In the above numerical discretization, $u_m(x, y)$ is approximatd by a polynomial of second order within an SE (see Eq. (9.6)). Next we briefly consider other cases in which polynomials of first order and third order are used.

Let the numerical version of $u_m(x, y)$ be a polynomial of first order. Then there are three expansion coefficients, i.e., three independent unknowns, for each $m = 1, 2, 3$. Thus, there are nine independent unknowns for each SE. Let the numerical solution satisfy Eq. (9.5) uniformly within an SE, i.e., an equation similar to Eq. (9.10) is imposed for

each m. Because the expression on the left side of Eq. (9.10) would become a constant if polynomials of first order were used in Eqs. (9.6)–(9.8), the last requirement represents one condition per SE for each m. Thus, to match the number of conditions with that of unknowns, six extra conditions per SE are needed.

In Fig. 30, the computational domain depicted in Fig. 29(a) is divided into triangular regions and rhombic regions. The interior of each triangular region is a boundary SE while the interior of each rhombic region is an interior SE. For each $\vec{h}_m^*$, let one interface flux conservation condition be imposed on each interior edge. In addition, as shown earlier, one may impose three boundary conditions over each boundary edge. Using arguments given previously, it is easy to show that these interface and boundary conditions result in the needed six extra conditions per SE. Let the union of each SE and its boundary be a CE. Then one concludes that, for each m, the total flux of $\vec{h}_m^*$ leaving the union of any combination of CE's vanishes.

Next we consider the case in which the numerical version of $u_m(x, y)$ is a polynomial of third order. Then there are ten expansion coefficients, i.e., ten independent unknowns, for each $m = 1, 2, 3$. Thus, there are 30 independent unknowns per SE. Again we assume that the numerical solution satisfies Eq. (9.5) uniformly within an SE. Because (i) the expression on the left side of Eq. (9.10) would become a polynomial of second order if polynomials of third order were used in Eqs. (9.6)–(9.8), and (ii) there are six expansion coefficients for each polynomial of second order involving two unknowns, the last requirement represents six conditions per SE for each m. Thus, to match the number of conditions with that of unknowns, 12 extra conditions per SE are needed. Because the number of extra conditions needed per SE is twice that needed in the special case we have just discussed, it is obvious that the current need can be fulfilled if (i) for each $\vec{h}_m^*$, two interface flux conditions are imposed on each interior edge depicted in Fig. 30, and (ii) six boundary conditions are imposed on each boundary edge. To give a definition of CE's, in Fig. 30, an interior SE is divided into 16 subregions, and a boundary SE is divided into six subregions. Each of these subregions can be considered as a CE.

Several variants of numerical discretization for the same flow problem have been described. Each variant represents a system of nonlinear equations, and is implicit in nature. Each can be solved by a variety of solution procedures. In [3], using Newton's method and another variant of discretization, a new efficient procedure was developed for the solution of incompressible, laminar channel flow. It was shown that, for a flow with $Re_L = 100$, an accurate solution can be obtained by using as few as six SE's across the channel.

Appendix A. An Alternative Stability Analysis for the Lax and Leapfrog/DuFort-Frankel Schemes

With the use of the regular mesh depicted in Fig. 31, the Lax scheme for solving Eq. (2.22) can be expressed as

$$\frac{u_{j'}^{n'+1} - \left(u_{j'+1}^{n'} + u_{j'-1}^{n'}\right)/2}{\Delta t'} + a\,\frac{u_{j'+1}^{n'} - u_{j'-1}^{n'}}{2\Delta x'} = 0, \qquad (A.1)$$

where $j', n' = 0, \pm 1, \pm 2, \ldots$. The system of equations represented by Eq. (A.1) can be divided into two sets completely independent from each other. The first set involves only the variables associated with those mesh points marked by dots in Fig. 31, and the second set, by crosses. Thus, the solution to Eq. (A.1) contains two decoupled solutions. Traditionally the von Neumann stability analysis for the Lax scheme is performed without taking into account this decoupling nature. Consider a solution to Eq. (A.1) in which $u_{j'}^{n'} = 1$ for all mesh points (j', n') that are marked by dots, and $u_{j'}^{n'} = -1$ for all other (j', n'). In reality, this solution represents the union of two completely decoupled *constant* solutions. However, at any time level, the combined solution is represented by a Fourier component of the shortest wavelength ($= 2\Delta x'$) in the traditional analysis. Therefore, two decoupled *constant* solutions may be wrongly perceived as a *rapidly-varying* solution. For the above reason, we shall consider each decoupled solution separately in the following von Neumann stability analysis.

Let $n = n'/2$, $j = j'/2$, $\Delta x = 2\Delta x'$, and $\Delta t = 2\Delta t'$. Then the mesh depicted in Fig. 31 is identical to that depicted in Fig. 2(a) except that those mesh points marked by crosses in Fig. 31 have no counterparts in Fig. 2(a). As a result, the decoupling nature of Eq. (A.1) will be removed if the Lax scheme is expressed using the staggered mesh depicted in Fig. 2(a), i.e., for all $(j, n) \in \Omega$,

$$\frac{u_j^n - \left(u_{j+1/2}^{n-1/2} + u_{j-1/2}^{n-1/2}\right)/2}{\Delta t/2} + a\,\frac{u_{j+1/2}^{n-1/2} - u_{j-1/2}^{n-1/2}}{\Delta x} = 0. \qquad (A.2)$$

With the aid of Eq. (2.13), Eq. (A.2) can be simplified as

$$u_j^n = (1/2)\left[(1+\nu)u_{j-1/2}^{n-1/2} + (1-\nu)u_{j+1/2}^{n-1/2}\right]. \qquad (A.3)$$

By applying Eq. (A.3) successively, one has

$$u_j^{n+1} = (1/4)\left[(1+\nu)^2 u_{j-1}^n + 2(1-\nu^2)u_j^n + (1-\nu)^2 u_{j+1}^n\right]. \qquad (A.4)$$

In contrast to Eq. (2.19), Eq. (A.4) implies that u_j^{n+1} *does not* approach u_j^n as $\Delta t \to 0$. Moreover, by substituting

$$u_j^n = [G(\nu,\theta)]^n\, e^{ij\theta} \qquad (i \stackrel{\text{def}}{=} \sqrt{-1}, \quad -\pi < \theta \le \pi) \qquad (A.5)$$

into Eq. (A.4), one concludes that the amplification factor of the Lax scheme is given by

$$G(\nu,\theta) = [\cos(\theta/2) - i\nu\sin(\theta/2)]^2 . \qquad (A.6)$$

A comparison among Eqs. (3.12), (3.15), and (A.6) reveals that $G_+^{(2)} = G_-^{(2)} = G(\nu,\theta)$ when $\epsilon = 1$.

Because u_j^{n+1} does not approach u_j^n as $\Delta t \to 0$. It follows from Eq. (A.5) that $G(\nu,\theta)$ cannot approach 1 as $\nu \to 0$. As a matter of fact, $G(\nu,\theta) \to \cos^2(\theta/2)$ as $\nu \to 0$. In turn, this implies that the Lax scheme is highly diffusive when $|\nu|$ is small.

With the use of the regular mesh depicted in Fig. 31, the Leapfrog/DuFort-Frankel scheme for solving Eq. (2.1) can be expressed as

$$\frac{u_{j'}^{n'+1} - u_{j'}^{n'-1}}{2\Delta t'} + a\,\frac{u_{j'+1}^{n'} - u_{j'-1}^{n'}}{2\Delta x'} - \mu\,\frac{u_{j'+1}^{n'} + u_{j'-1}^{n'} - u_{j'}^{n'+1} - u_{j'}^{n'-1}}{(\Delta x')^2} = 0. \qquad (A.7)$$

where $j', n' = 0, \pm 1, \pm 2, \ldots$. Even though Eq. (A.7) is a three-level scheme while Eq. (A.1) is a two-level scheme, they have the same decoupling nature. The decoupling of Eq. (A.7) can be removed if the scheme is expressed with respect to the staggered mesh depicted in Fig. 2(a), i.e., for all $(j,n) \in \Omega$,

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + a\,\frac{u_{j+1/2}^{n-1/2} - u_{j-1/2}^{n-1/2}}{\Delta x} - \mu\,\frac{u_{j+1/2}^{n-1/2} + u_{j-1/2}^{n-1/2} - u_j^n - u_j^{n-1}}{(\Delta x/2)^2} = 0. \qquad (A.8)$$

With the aid of Eq. (2.13), Eq. (A.8) can be simplified as

$$(1+\xi)u_j^n = (1-\xi)u_j^{n-1} + (\nu+\xi)u_{j-1/2}^{n-1/2} - (\nu-\xi)u_{j+1/2}^{n-1/2}, \qquad (A.9)$$

Eq. (A.9) can also be expressed in a two-level form, i.e.,

$$\vec{u}(j,n) = L_+\vec{u}(j-1/2,\, n-1/2) + L_-\vec{u}(j+1/2,\, n-1/2). \qquad (A.10)$$

Here

$$\vec{u}(j,n) \stackrel{\text{def}}{=} \begin{pmatrix} u_j^n \\ u_{j+1/2}^{n-1/2} \end{pmatrix} \qquad (A.11)$$

for all $(j,n) \in \Omega$ with $n > 0$, and

$$L_+ \stackrel{\text{def}}{=} \begin{pmatrix} \dfrac{\nu+\xi}{1+\xi} & \dfrac{1-\xi}{1+\xi} \\ 0 & 0 \end{pmatrix}, \quad \text{and} \quad L_- \stackrel{\text{def}}{=} \begin{pmatrix} \dfrac{-(\nu-\xi)}{1+\xi} & 0 \\ 1 & 0 \end{pmatrix}. \qquad (A.12)$$

By applying Eq. (A.10) successively, one has

$$\vec{u}(j, n+1) = (L_+)^2\vec{u}(j-1, n) + (L_+L_- + L_-L_+)\vec{u}(j,n) + (L_-)^2\vec{u}(j+1, n). \quad (A.13)$$

To perform the von Neumann stability analysis for Eq. (A.13), let

$$\vec{u}(j,n) = \vec{u}^{\star}(n,\theta)e^{ij\theta} \qquad (i \stackrel{\text{def}}{=} \sqrt{-1}, \quad -\pi < \theta \leq \pi) \quad (A.14)$$

where $\vec{u}^{\star}(n,\theta)$ is a 2×1 column matrix. Substituting Eq. (A.14) into Eq. (A.13), one obtains

$$\vec{u}^{\star}(n+1,\theta) = [L(\nu,\xi,\theta)]^2\ \vec{u}^{\star}(n,\theta) \quad (A.15)$$

where

$$L(\nu,\xi,\theta) \stackrel{\text{def}}{=} e^{-i\theta/2}L_+ + e^{i\theta/2}L_-. \quad (A.16)$$

According to Eq. (A.15), $[L(\nu,\xi,\theta)]^2$ is the amplification matrix. Substituting Eq. (A.12) into Eq. (A.16), one has

$$L(\nu,\xi,\theta) = \begin{pmatrix} \dfrac{2[\xi\cos(\theta/2) - i\nu\sin(\theta/2)]}{1+\xi} & \dfrac{(1-\xi)e^{-i\theta/2}}{1+\xi} \\ e^{i\theta/2} & 0 \end{pmatrix}. \quad (A.17)$$

The amplification factors $A_\pm$ given in Eq. (2.21) are the eigenvalues of the amplification matrix $[L(\nu,\xi,\theta)]^2$.

References

[1] Chang, S.C. and To, W.M., "A New Numerical Framework for Solving Conservation Laws–The Method of Space-Time Conservation Element and Solution Element," NASA TM 104495, August, 1991.

[2] Chang, S.C. and To, W.M., "A Brief Description of a New Numerical Framework for Solving Conservation Laws–The Method of Space-Time Conservation Element and Solution Element," *Proceedings of the Thirteenth International Conference on Numerical Methods in Fluid Dynamics*, Rome, Italy, 1992, M. Napolitano and F. Sabetta, eds., Lecture Notes in Physics 414, Springer-Verlag, pp. 396-400. Also published as NASA TM 105757.

[3] Scott, J.R. and Chang, S.C., "A New Flux Conserving Newton's Method Scheme for the Two-Dimensional, Steady Navier-Stokes Equations," NASA TM 106160, June, 1993.

[4] Anderson, D.A., Tannehill, J.C. and Pletcher, R.H., *Computational Fluid Mechanics and Heat Transfer* (Hemisphere, 1984).

[5] Baker, A.J., *Finite Element Computational Fluid Mechanics* (Hemisphere, 1983).

[6] Canuto, C., Hussaini, M.Y., Quarteroni, A. and Zang, T.A., *Spectral Methods in Fluid Dynamics* (Springer-Verlag New York Inc., 1988).

[7] Vinokur, M., "An Analysis of Finite-Difference and Finite-Volume Formulation of Conservation Laws," J. Comput. Phys., 81, 1989,pp. 1-52.

[8] LeVeque, R.J., *Numerical Methods for Conservation Laws* (Birkhäuser Verlag, 1990).

[9] Roe, P.L., "Approximate Riemann Solvers, Parameter Vectors and Difference Schemes," J. Comput. Phys., 43, 1981, pp. 357-372.

[10] van Leer, B., "Flux Vector Splitting for the Euler Equations," Lecture Notes in Physics 170, 1982, pp. 501-512.

[11] Osher, S. and Chakravarthy, S., "Upwind Schemes and Boundary Conditions with Applications to Euler Equations in General Coordinates," J. Comput. Phys., 50, 1983, pp. 447-481.

[12] van Leer, B., "Toward the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection," J. Comput. Phys., 23, 1977, pp. 276-299.

[13] Smith, M.J. and Stoker, R.W., "Extension of CFD Techniques to Computational Aeroacoustics (CAA): Comparative Evaluation," AIAA Paper 93-0150, Reno, Nevada, January 1993.

[14] Chang, S.C., "On An Origin of Numerical Diffusion: Violation of Invariance under Space-Time Inversion," *Proceedings of 23rd Conference on Modeling and Simulation*, April 30–May 1, 1992, Pittsburgh, PA, USA, William G. Vogt and Marlin H. Mickle eds., Part 5, pp. 2727-2738. Also published as NASA TM 105776.

[15] Roe, P.L., "A Survey of Upwind Differencing Techniques," *Proceedings of the Eleventh International Conference on Numerical Methods in Fluid Dynamics*, 1988, Lecture Notes in Physics 323, Springer-Verlage, 1989, pp. 69-78.

[16] Warming, R.F., Beam, R.M. and Hyett B.J., "Diagonalization and Simultaneous Symmetrization of the Gas-Dynamics Matrices," Math. of Comput., 29, 1975, pp. 1037-1045.

[17] Rusanov, V.V., "The characteristics of General Equations of Gas Dynamics," Zhurnal vychislitel 'noi mathematiki mathematicheskoi fiziki, 3, 1963, pp. 508-527.

[18] Nessyahu, H. and Tadmor, E., "Non-oscillatory Central Differencing for Hypobolic Conservation Laws," J. Comput. Phys., 87, 1990, pp. 408-463.

[19] Leonard, B.P., "Universal Limiter for Transient Interpolation Modeling of the Advective Transport Equations: The ULTIMATE Conservative Difference Scheme," NASA TM 100916, September, 1988.

[20] Courant R. and Hilbert D., *Methods of Mathematical Physics*, Vol. II (Interscience, 1962).

[21] van Leer, B., "Toward the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection," J. Compu. Phys., 23, 1977, pp. 276-298.

[22] van Albada, G.D., van Leer, B. and Roberts, W.W., "A Comparative Study of Computational Methods in Cosmic Gas Dynamics," Astronom. and Astrophys.,108, 1982, pp. 76-84.

[23] Noble, B. and Daniel, J.W., Linear Algebra and Its Applications, 2nd edition (Prentice-Hall, 1977).

[24] Richtmyer, R.D. and Morton, K.W., *Difference Methods for Initial-value Problems*, 2nd edition (Interscience, 1967).

[25] Warming, R.F. and Hyett, B.J., "The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods," J. Compu. Phys., 14, 1974, pp. 159-179.

[26] Griffiths, D.F. and Sanz-Serna, J.M., "On the Scope of the Method of Modified Equations," SIAM J. Sci. Stat. Comput., 7, 1988, pp. 994-1008.

[27] Chang, S.C., "A Critical Analysis of the Modified Equation Technique of Warming and Hyett," J. Compu. Phys., 86, 1990, pp. 107-126.

[28] Sod, G.A., "A Survey of Several Finite Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws," J. Comput. Phys., 27, 1978, pp. 1-31.

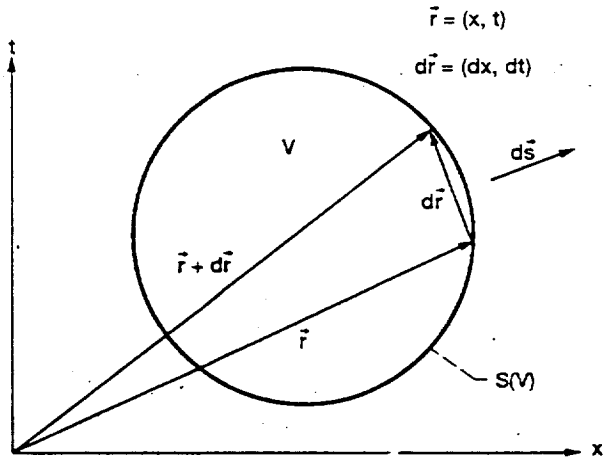
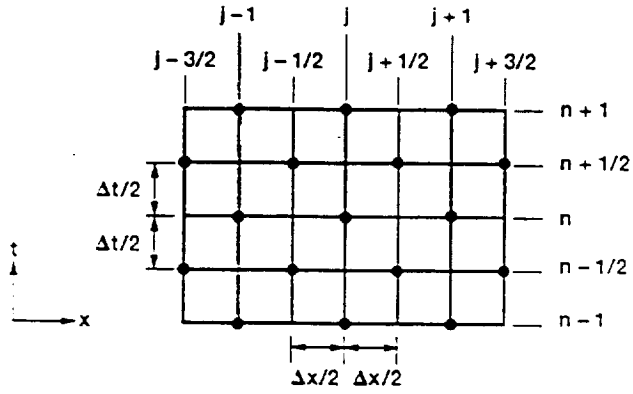
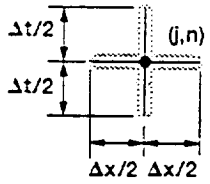


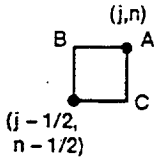
Figure 1.—A surface element $d\vec{s}$ and a line segment $d\vec{r}$ on the boundary S(V) of a volume V in a space-time E_2.



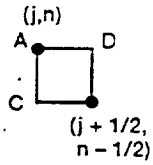
(a) The relative positions of SEs and CEs.



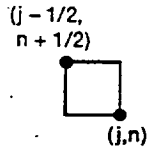
(b) SE (j,n).



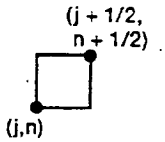
(c) $CE_{-}(j,n)$.



(d) $CE_{+}(j,n)$.

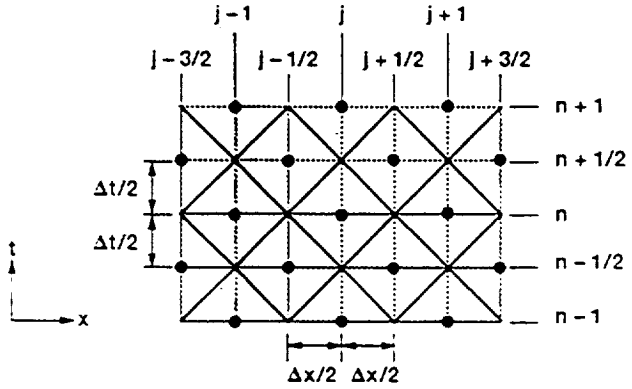


(e) $CE_{+}(j - 1/2, n + 1/2)$.

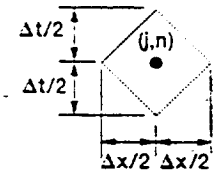


(f) $CE_{-}(j + 1/2, n + 1/2)$.

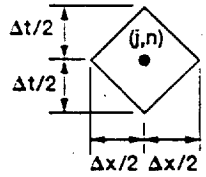
Figure 2.—The SEs and CEs of type I.



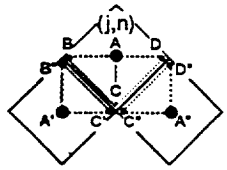
(a) The relative positions of SEs and CEs.



(b) SE (j,n).



(c) CE (j,n).



(d) Three neighboring CEs.

Figure 3.—The SEs and CEs of type II.

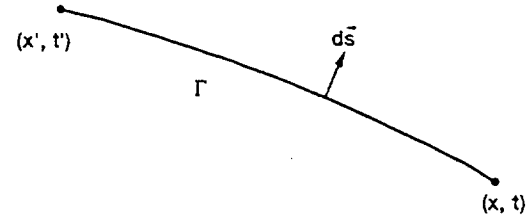
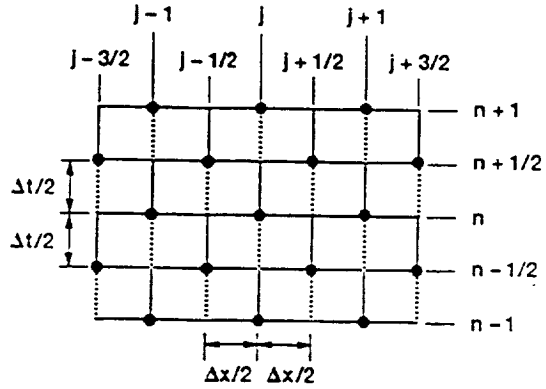
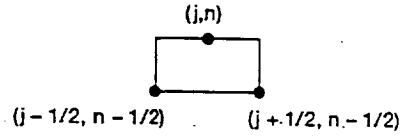


Figure 4.—A simple curve Γ joining (x, t) and (x', t').



(a) The relative positions of CEs and mesh points.



(b) CE (j,n).

Figure 5.—The mesh and CEs of the a-ϵ scheme.

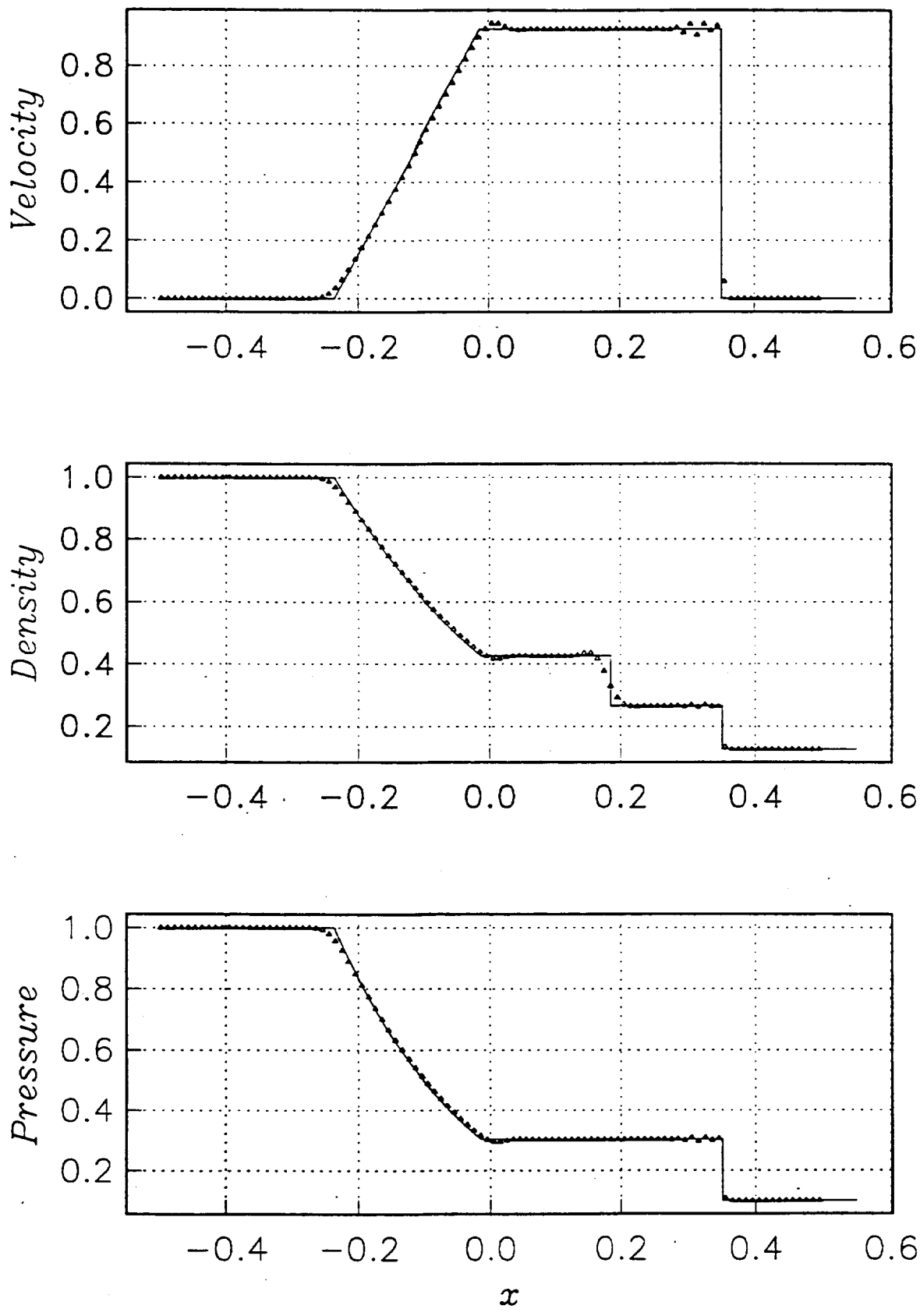


Figure 6.– The Euler solution ($\epsilon = 0.5$, $\alpha = 0$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

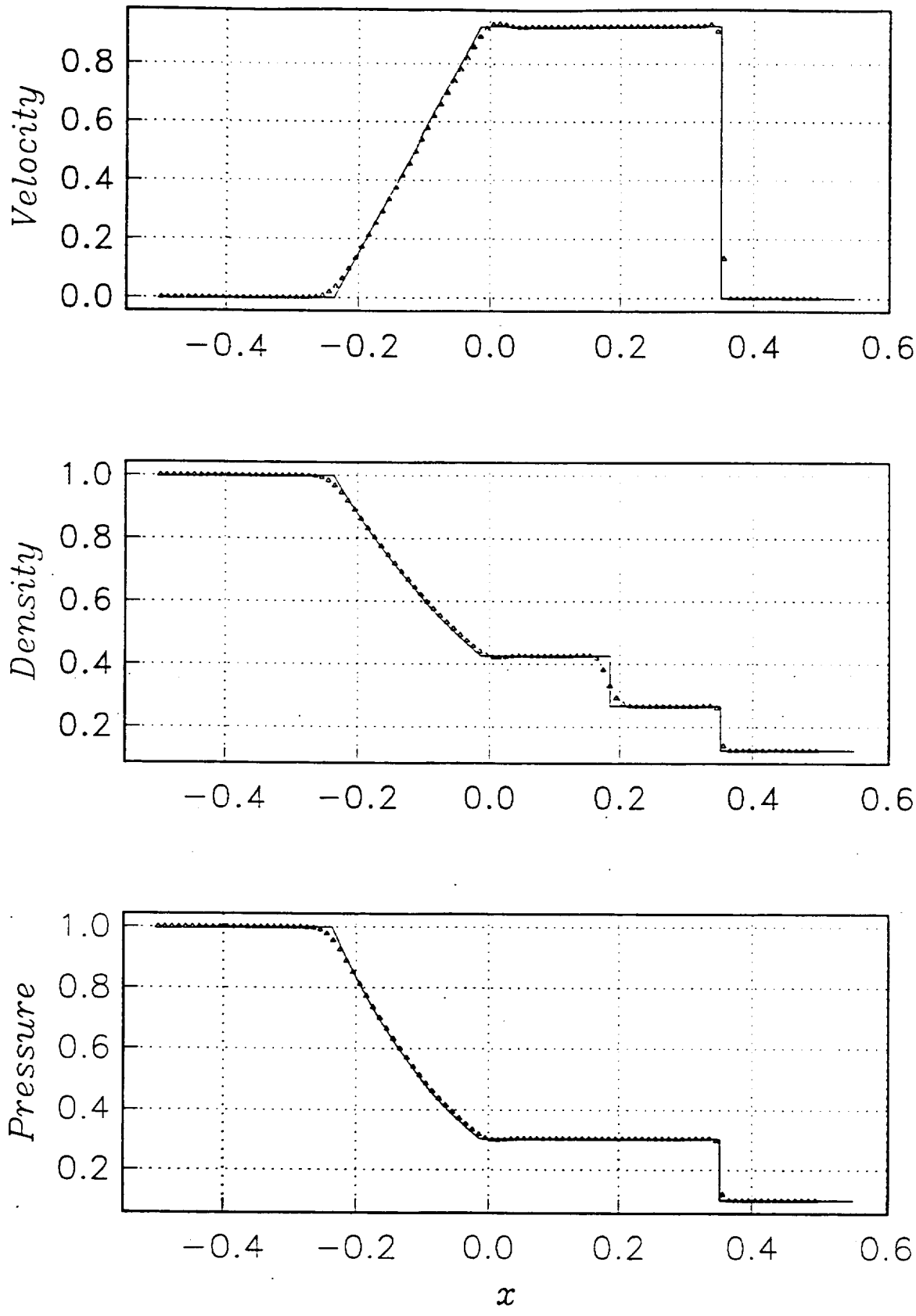


Figure 7.– The Euler solution ($\epsilon = 0.5$, $\alpha = 1$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

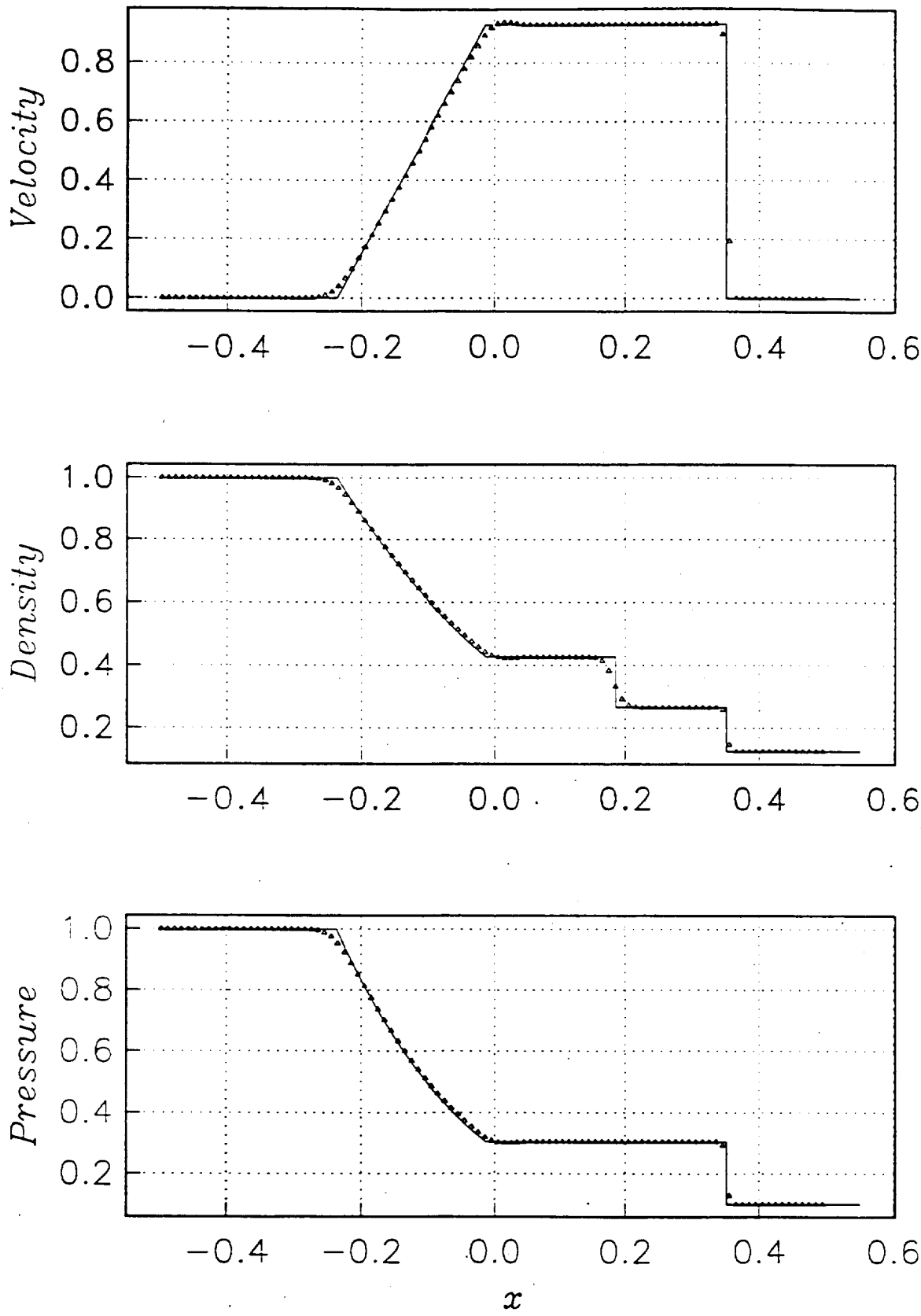


Figure 8.– The Euler solution ($\epsilon = 0.5$, $\alpha = 2$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

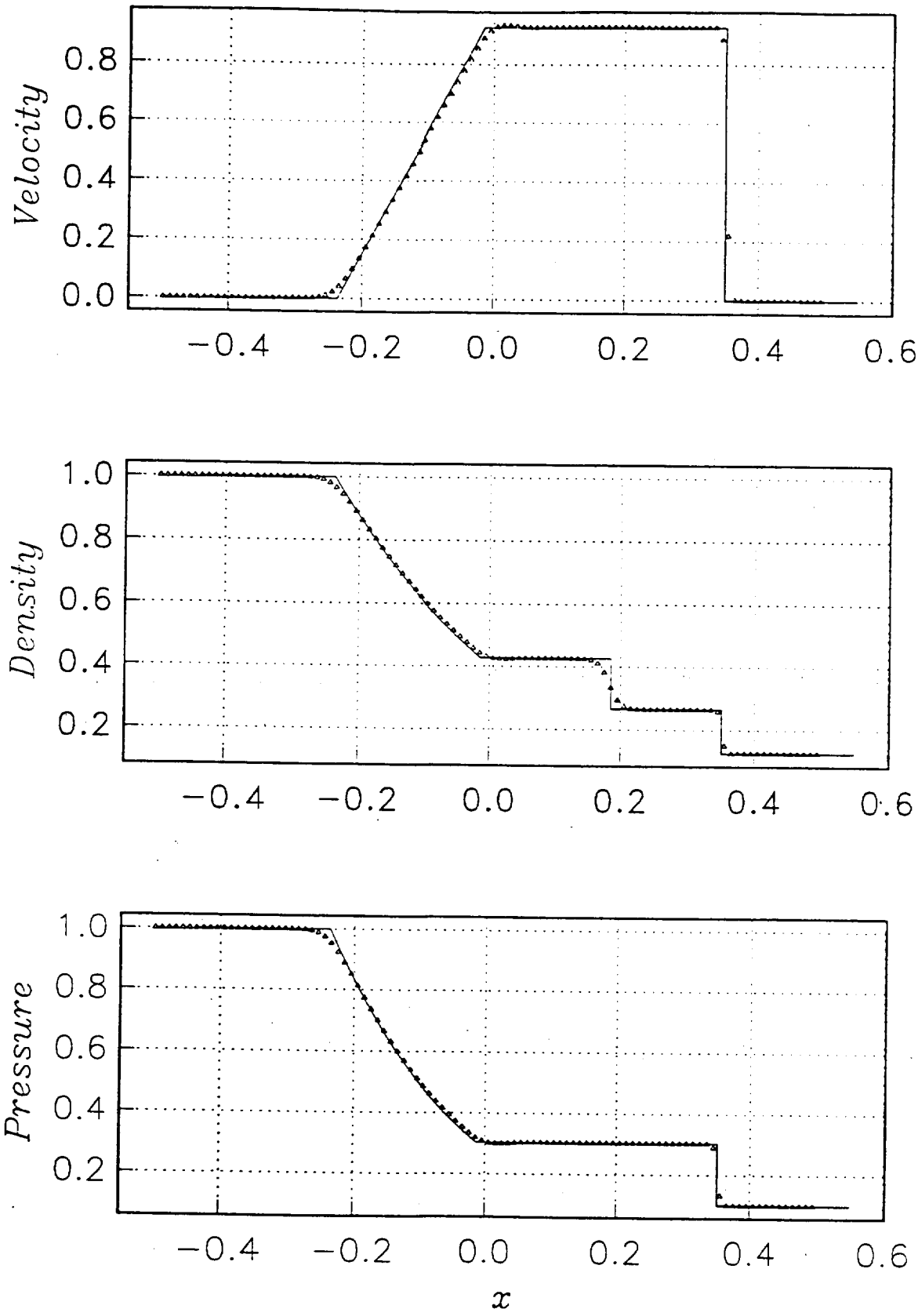


Figure 9.– The Euler solution ($\epsilon = 0.5$, $\alpha = 3$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

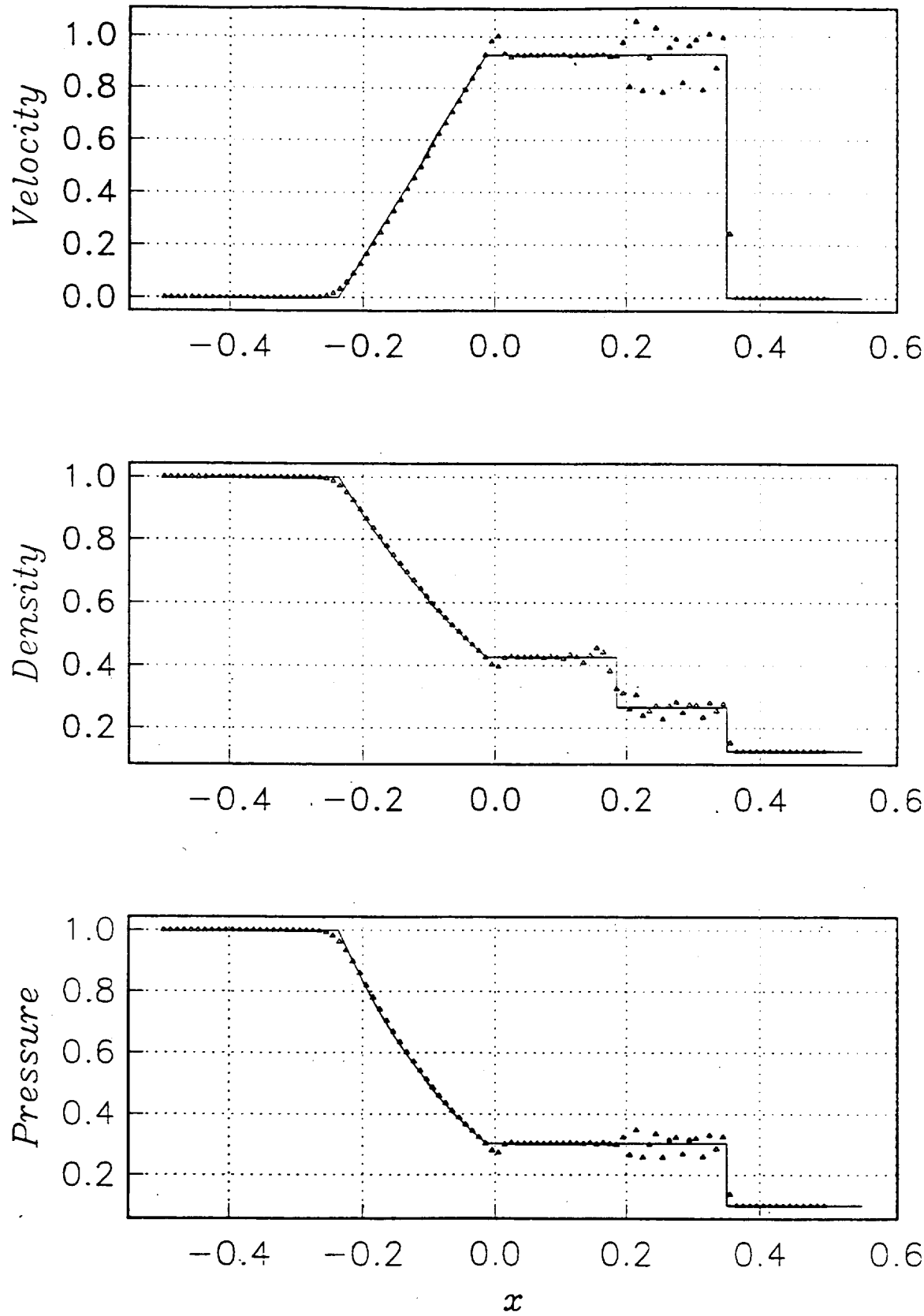


Figure 10.– The Euler solution ($\epsilon = 0.1$, $\alpha = 0$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

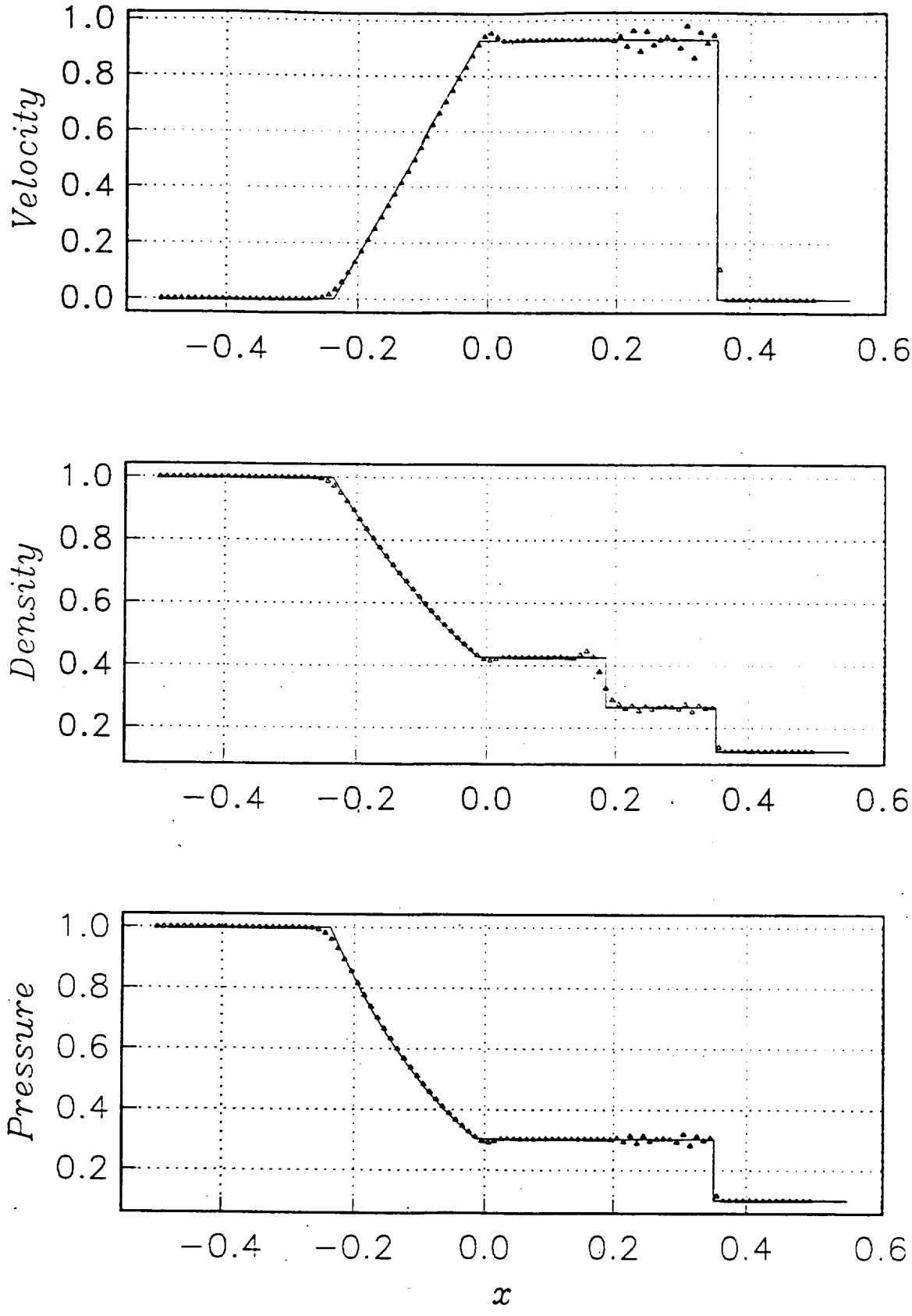


Figure 11.– The Euler solution ($\epsilon = 0.3$, $\alpha = 0$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

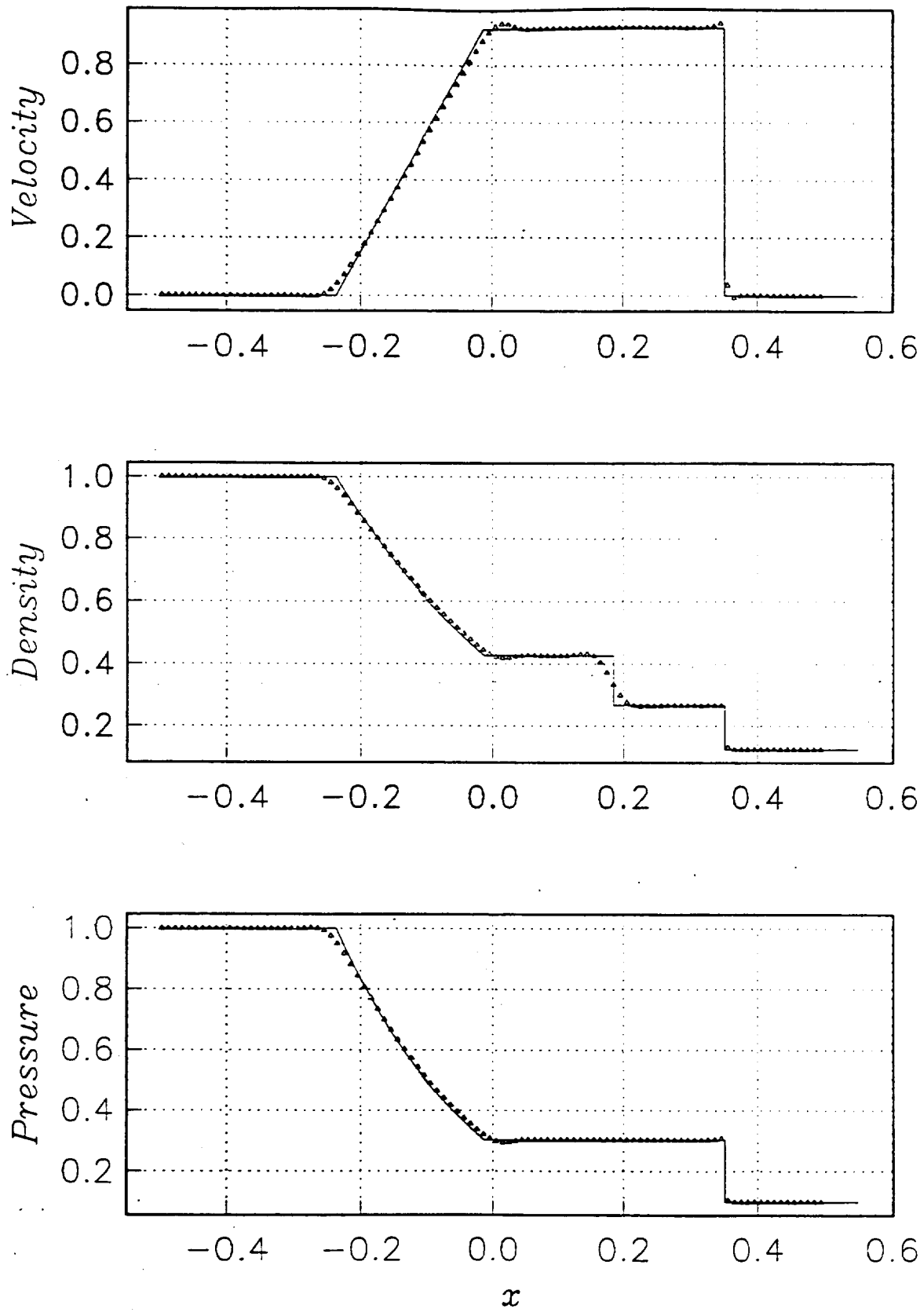


Figure 12.– The Euler solution ($\epsilon = 0.7$, $\alpha = 0$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

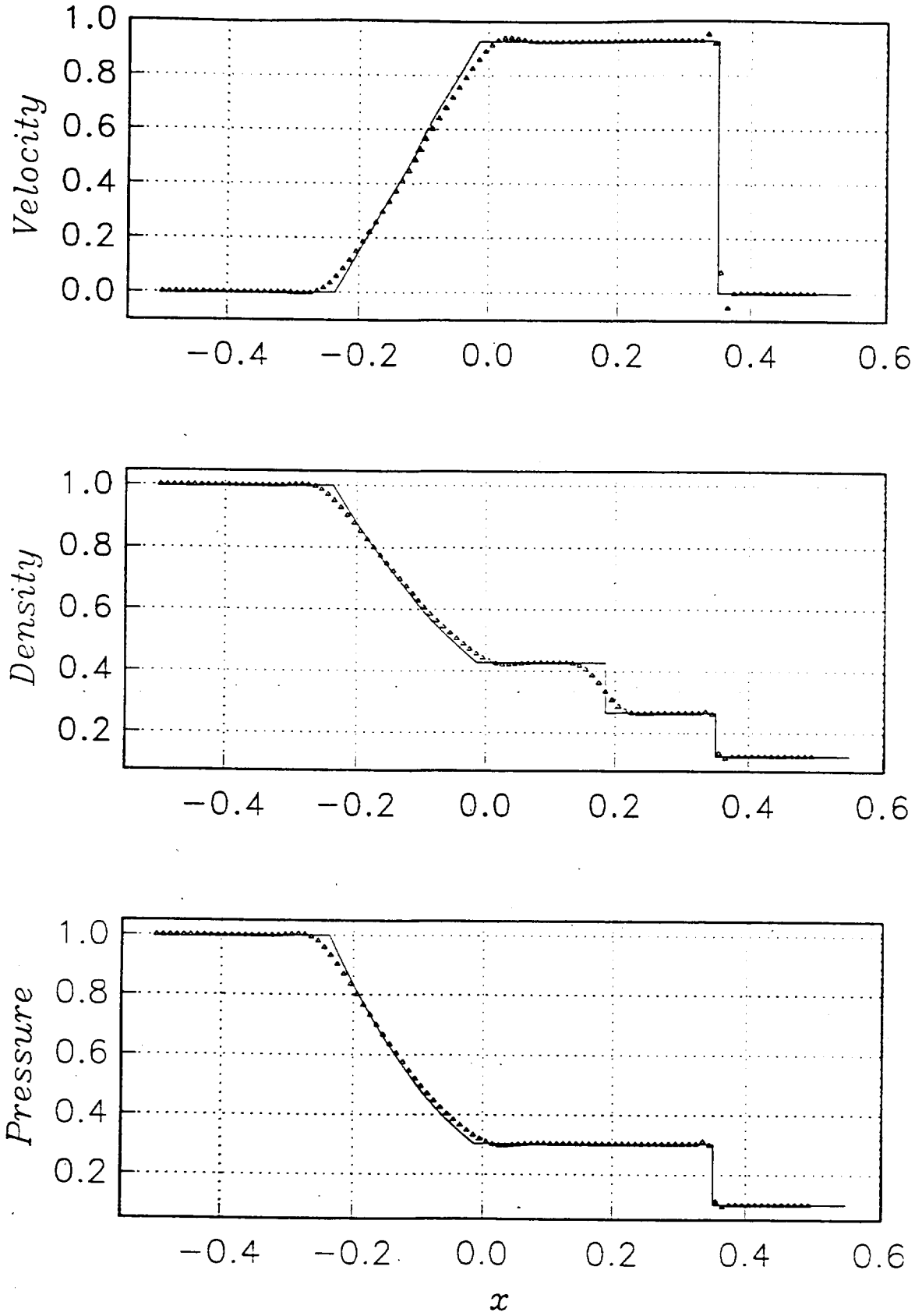


Figure 13.– The Euler solution ($\epsilon = 0.9$, $\alpha = 0$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

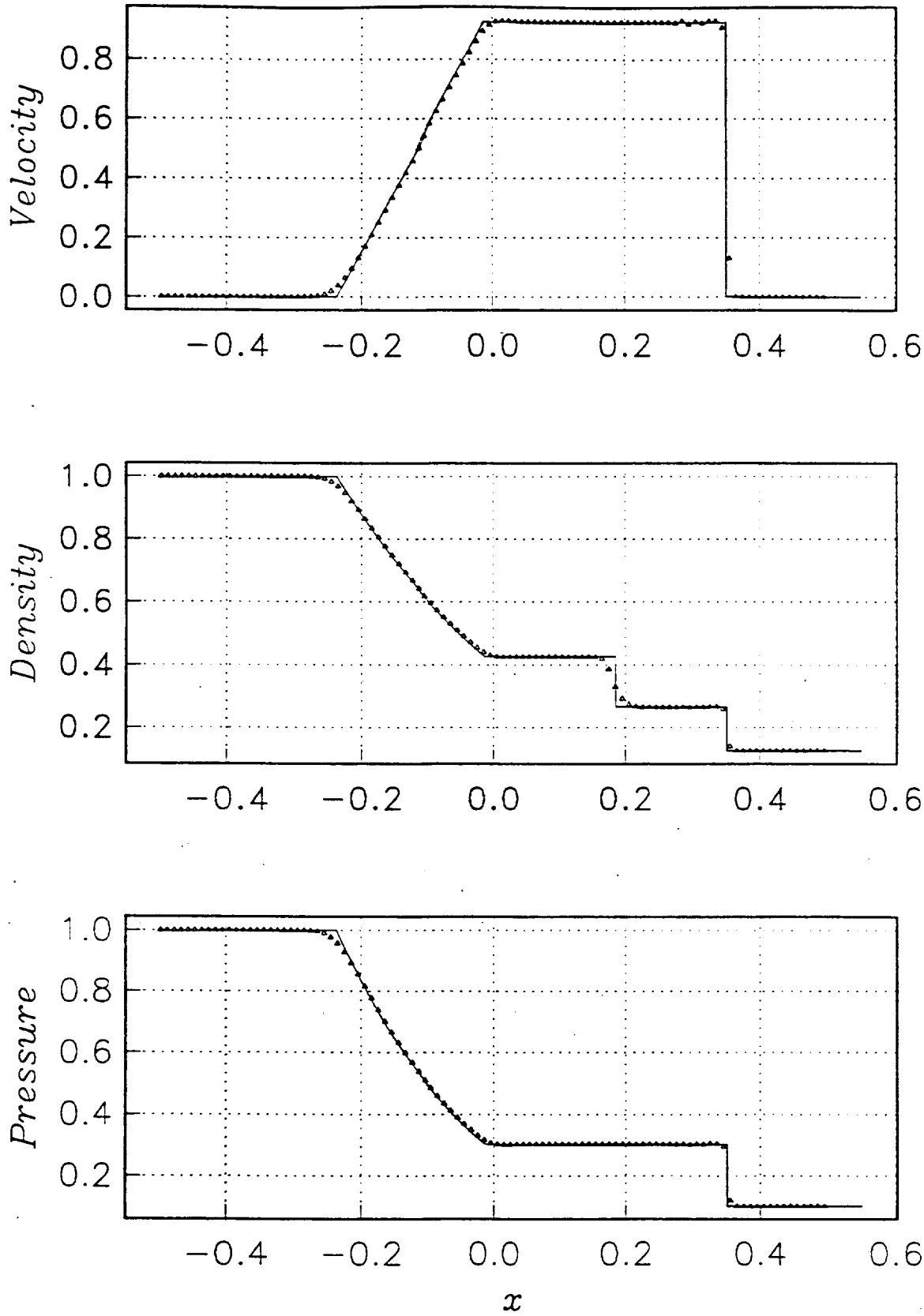


Figure 14.– The Euler solution ($\epsilon = 0.3$, $\alpha = 2$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

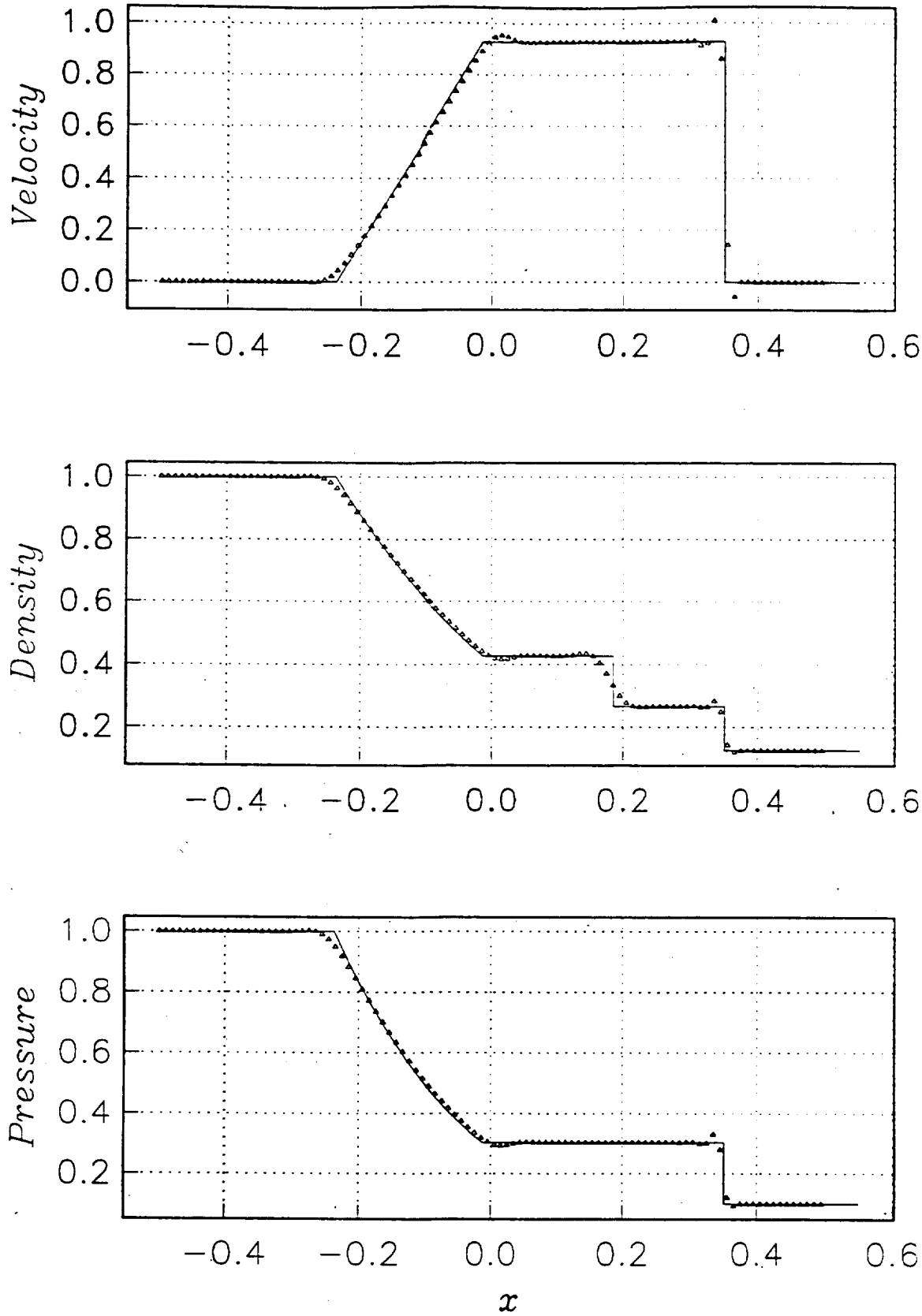


Figure 15.– The Euler solution ($\epsilon = 0.5$, $\alpha = 0$, $\Delta t = 0.002$, $CFL \doteq 0.44$).

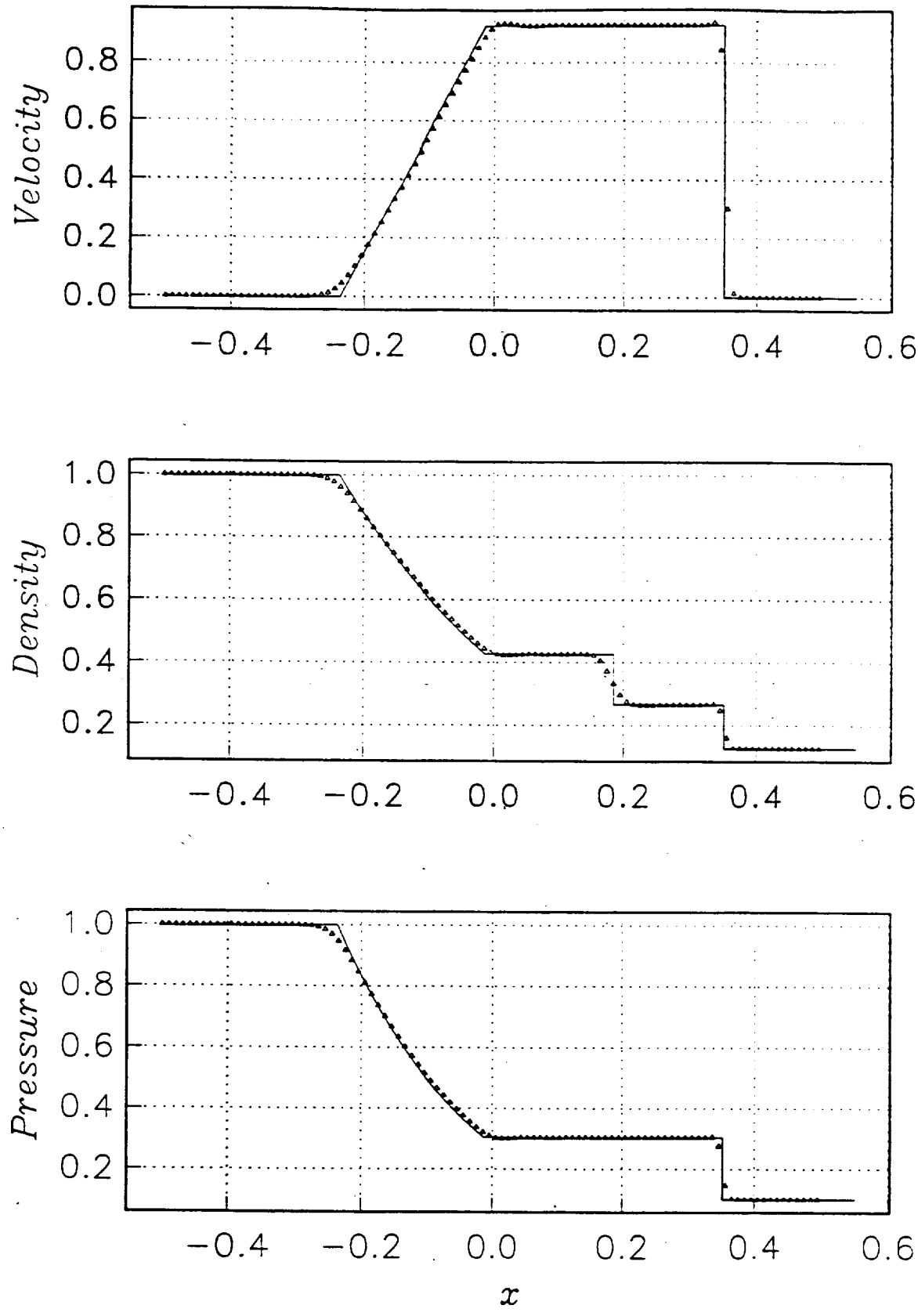


Figure 16.- The Euler solution ($\epsilon = 0.5$, $\alpha = 1$, $\Delta t = 0.002$, $CFL \doteq 0.44$).

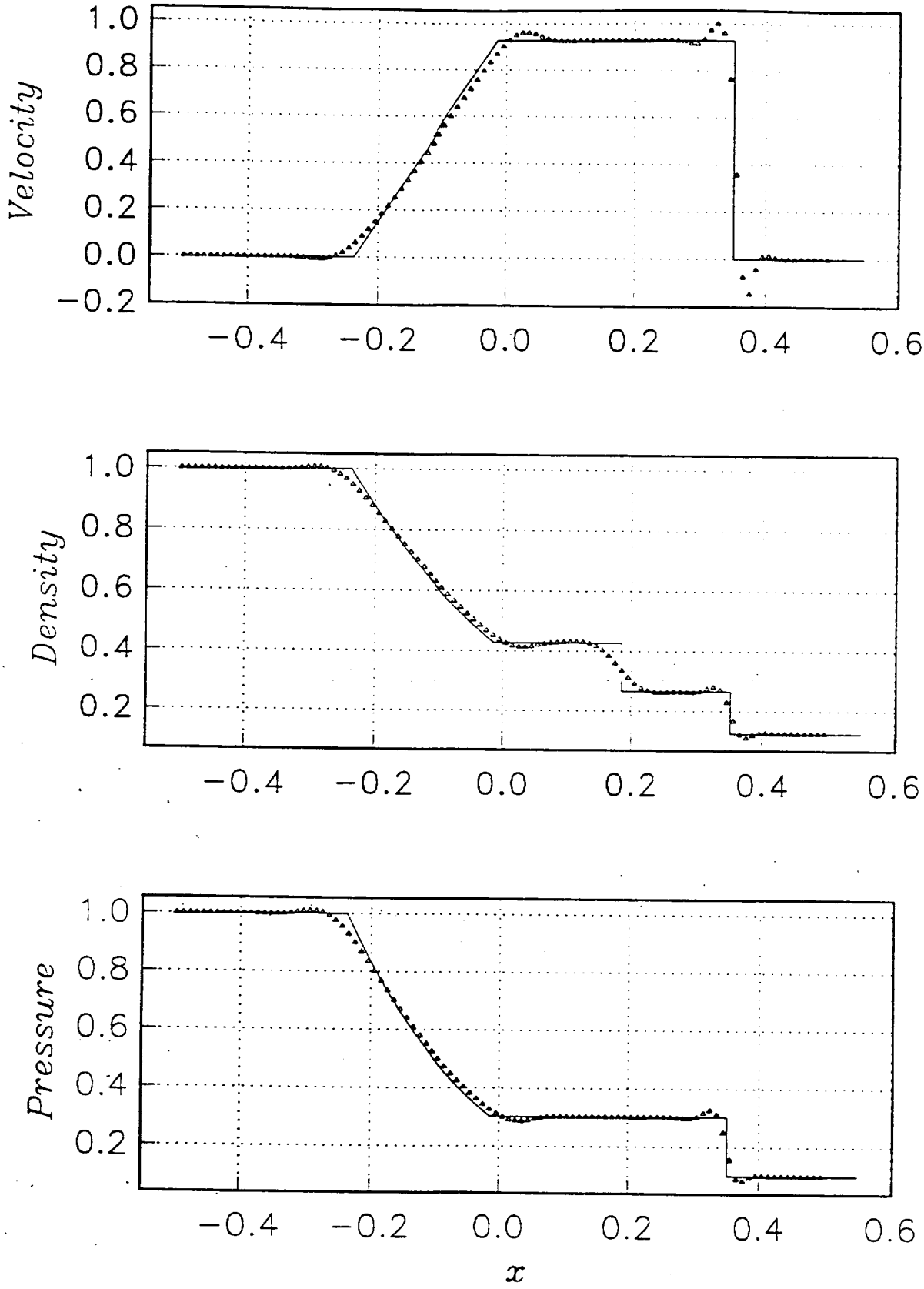


Figure 17.– The Euler solution ($\epsilon = 0.5$, $\alpha = 0$, $\Delta t = 0.0004$, $CFL \doteq 0.088$).

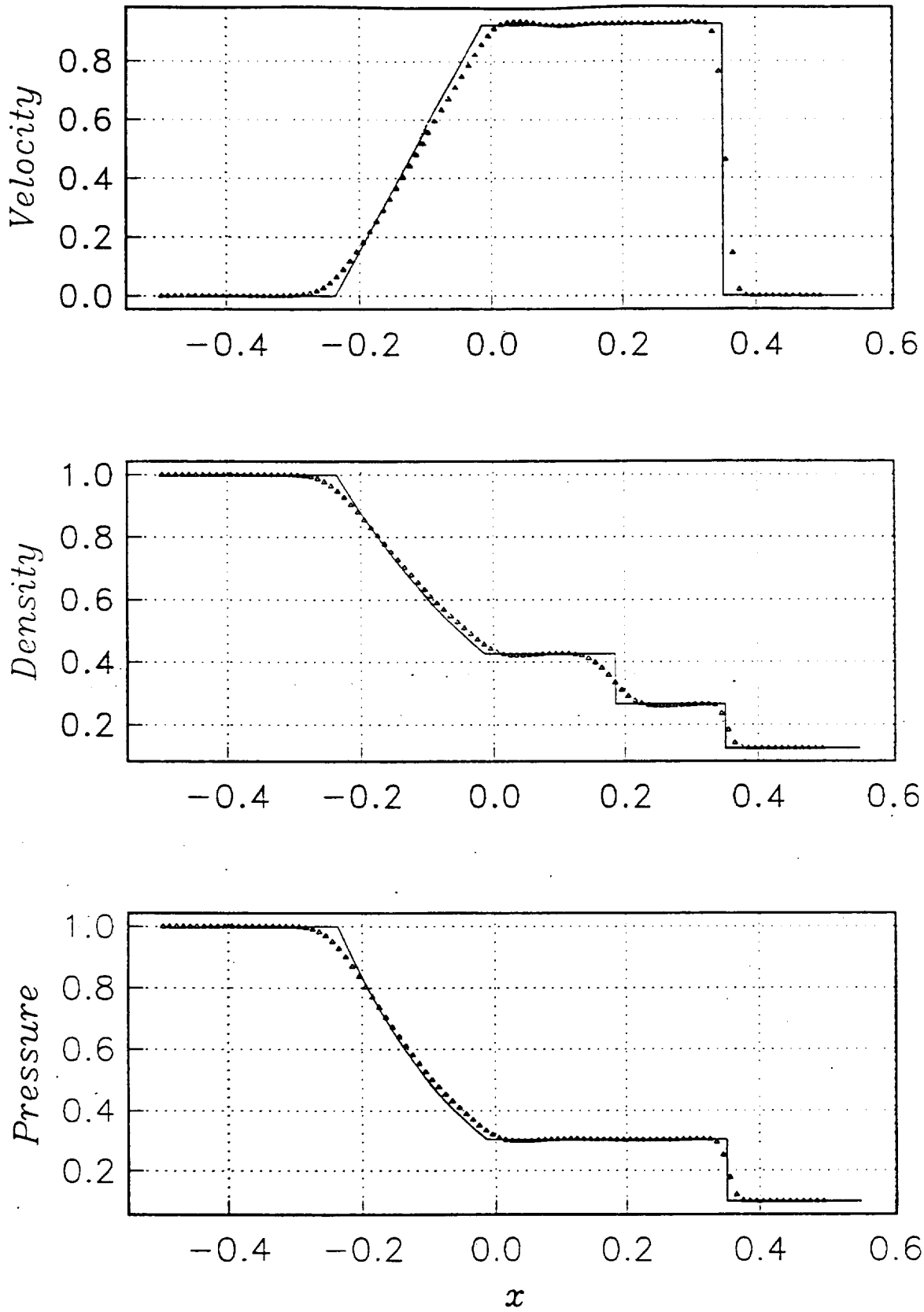


Figure 18.– The Euler solution ($\epsilon = 0.5$, $\alpha = 1$, $\Delta t = 0.0004$, $CFL \doteq 0.088$).

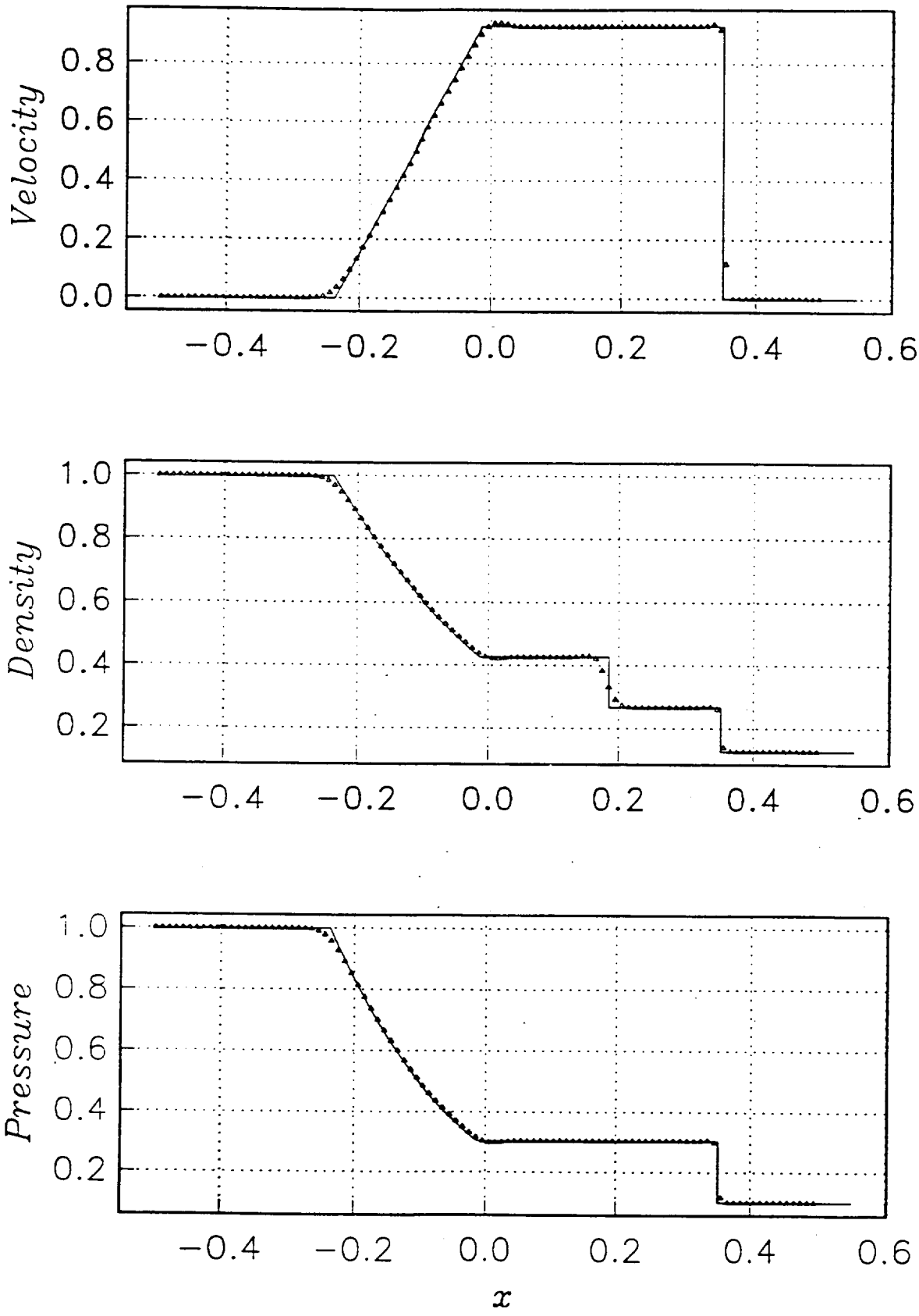


Figure 19.– The Euler solution ($b = 0.5$, $\alpha = 1$, $\Delta t = 0.004$, $CFL \doteq 0.88$).

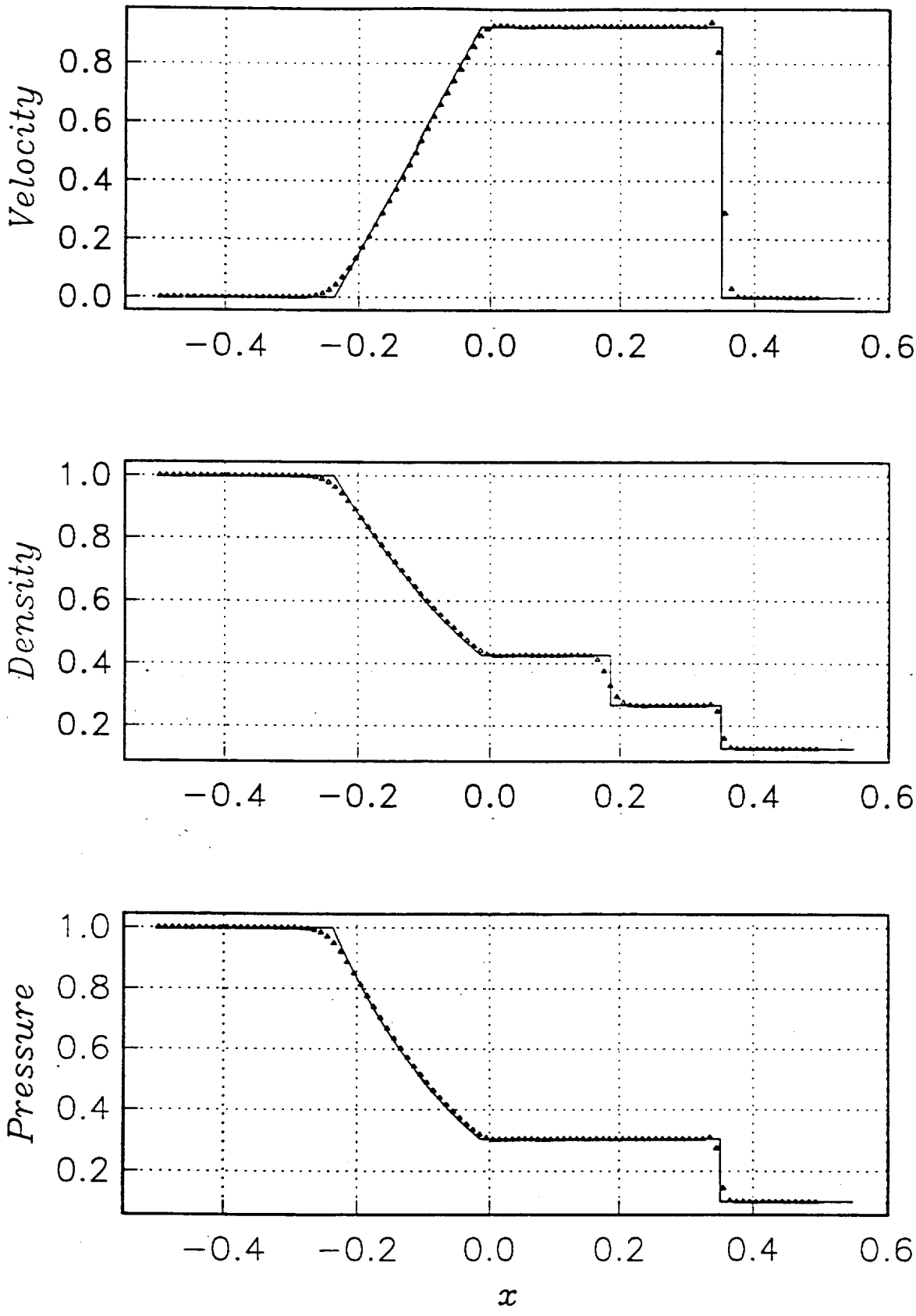


Figure 20.- The Euler solution ($b = 0.5$, $\alpha = 1$, $\Delta t = 0.0004$, $CFL \doteq 0.088$).

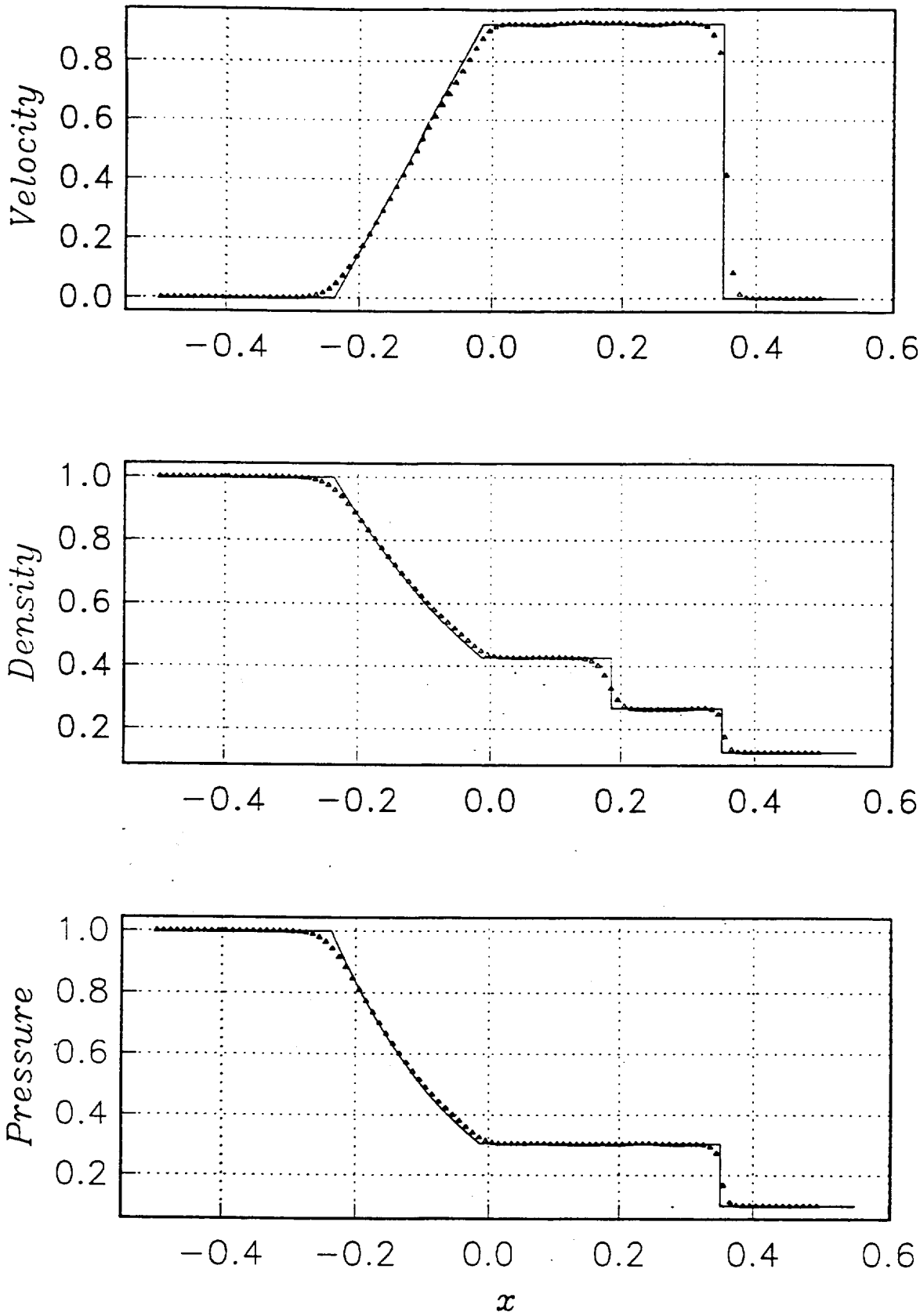


Figure 21.– The Euler solution ($b = 0.5$, $\alpha = 1$, $\Delta t = 0.0001$, $CFL \doteq 0.022$).

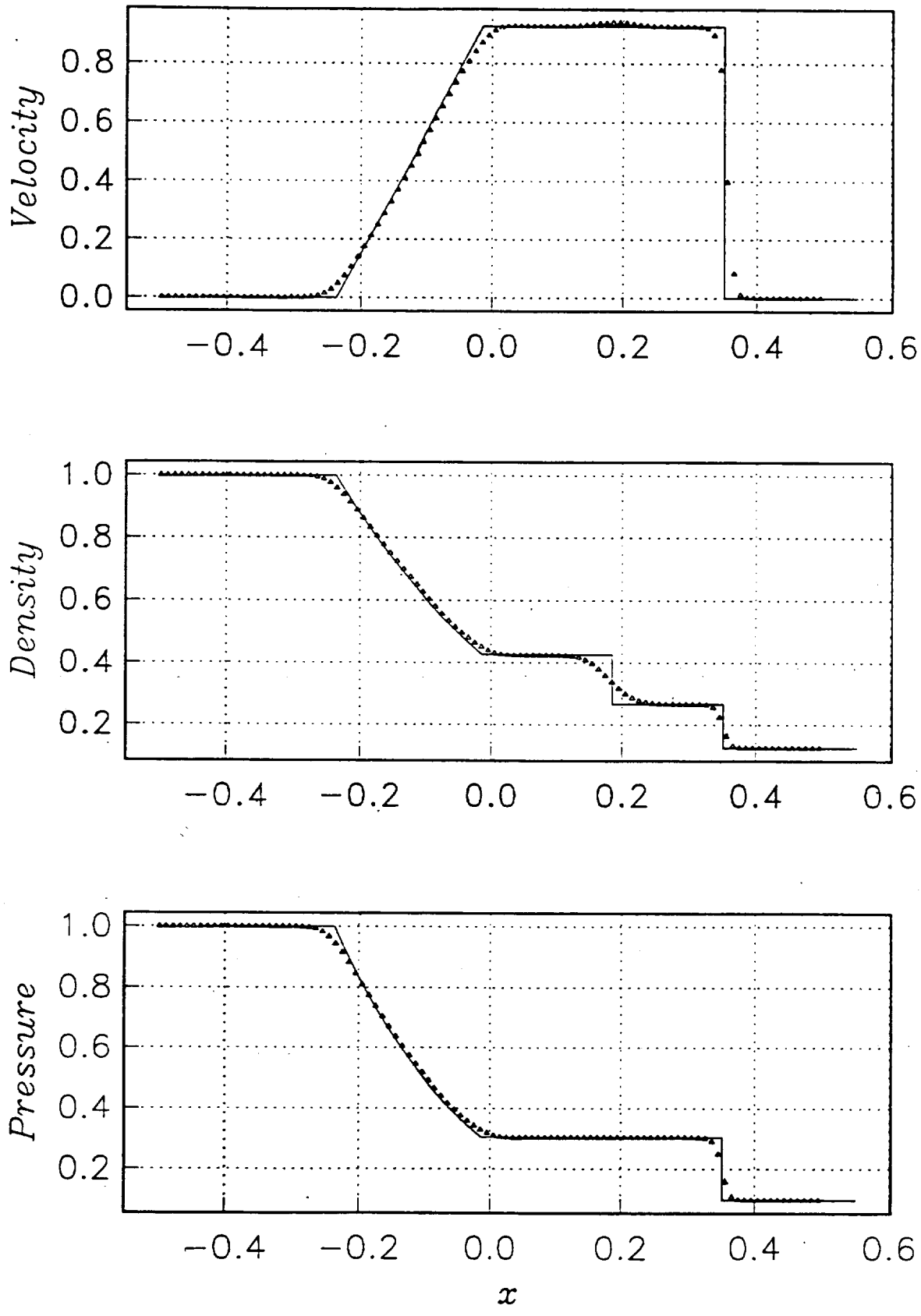


Figure 22.- The Navier–Stokes solution ($Re_L = 2000$).

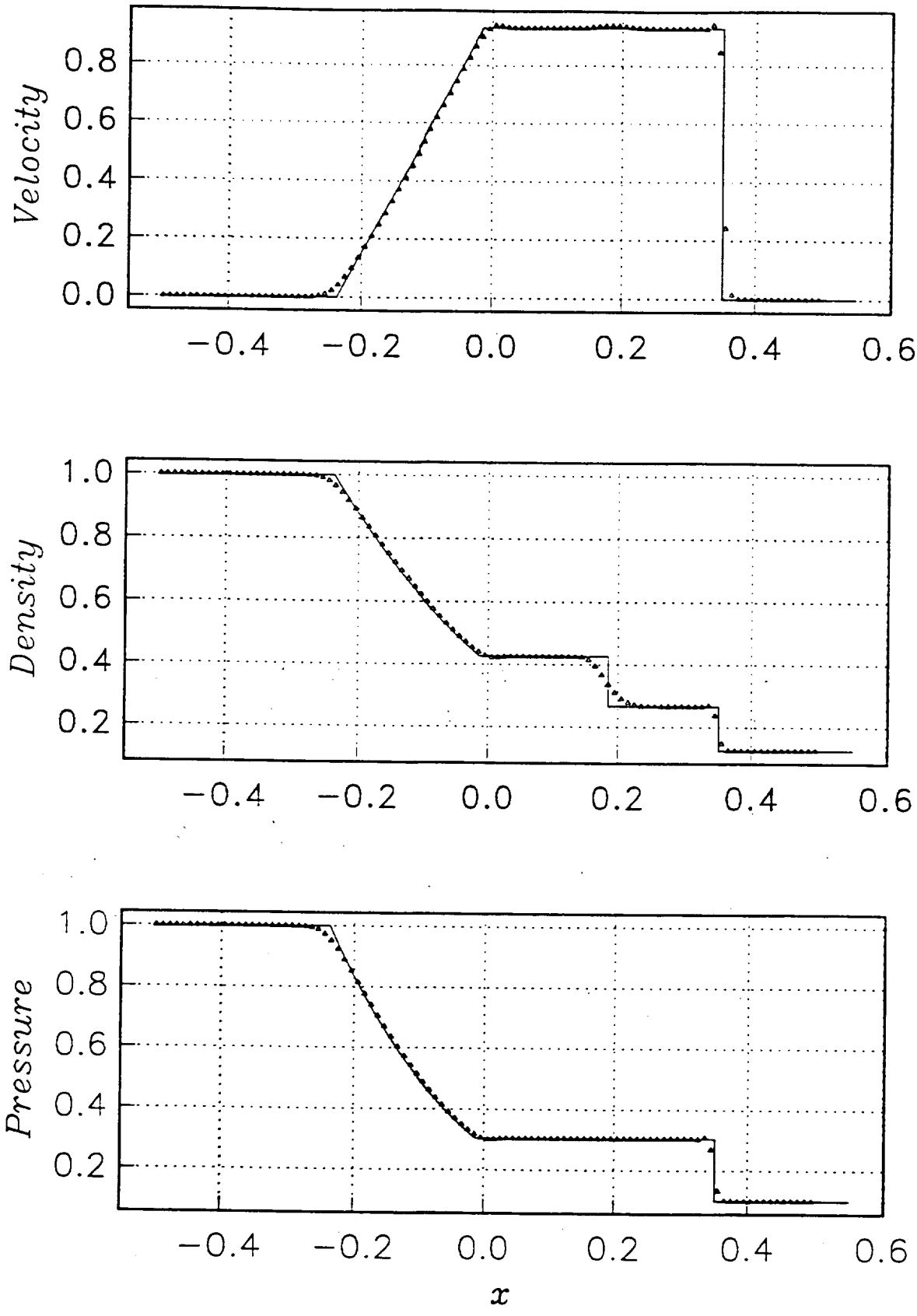


Figure 23.- The Navier-Stokes solution ($Re_L = 4000$).

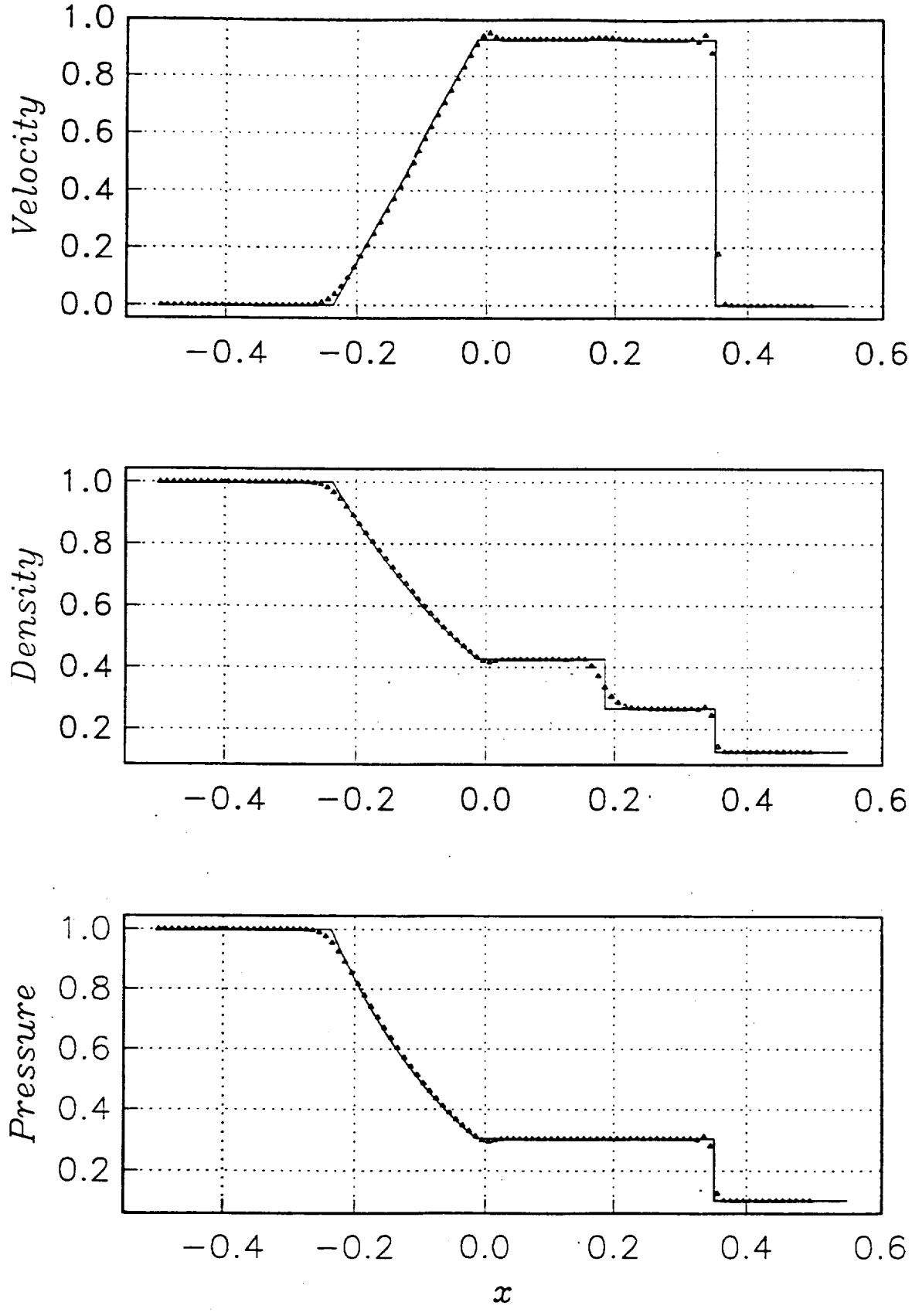


Figure 24.– The Navier-Stokes solution ($Re_L = 6000$).

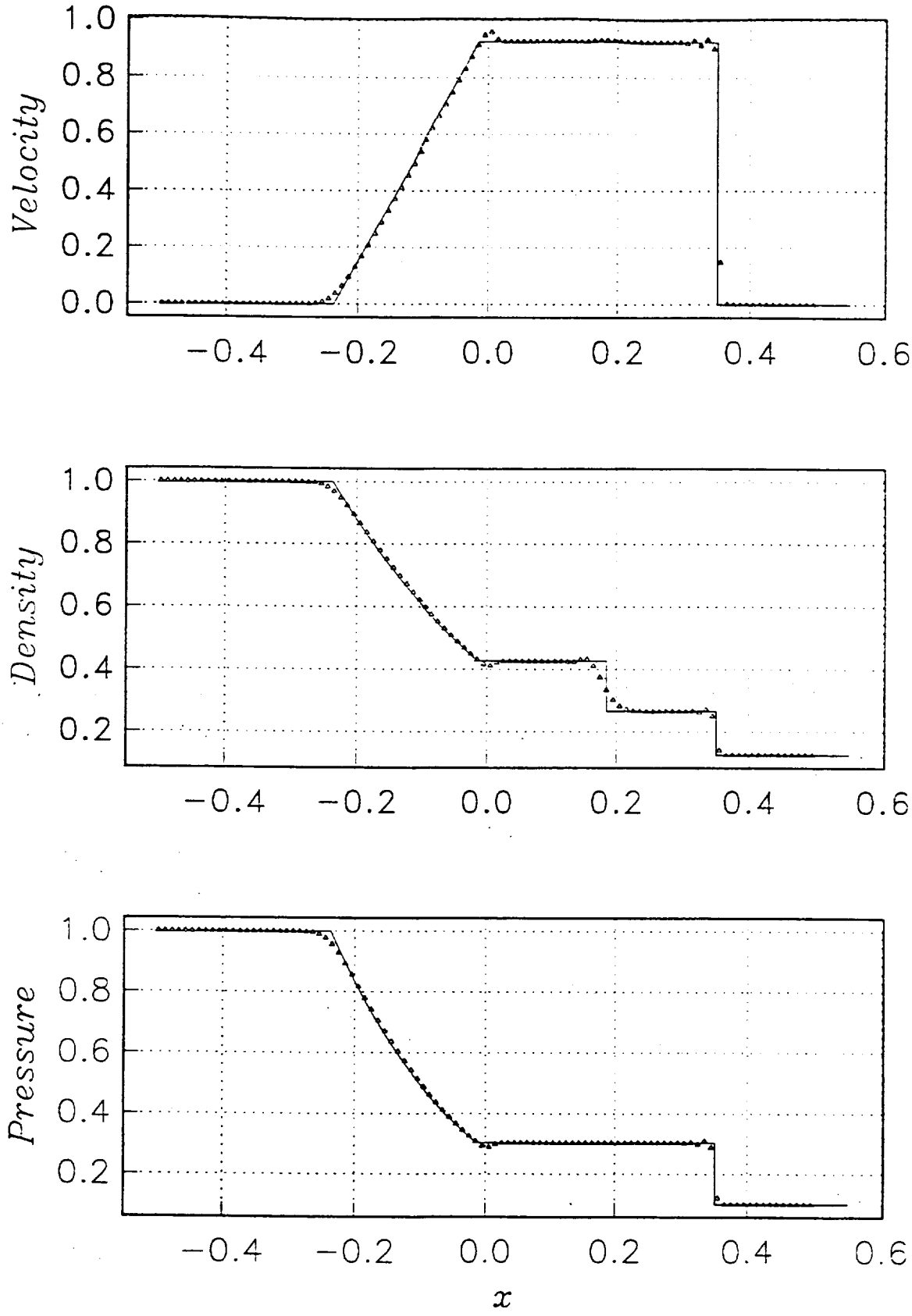


Figure 25.- The Navier-Stokes solution ($Re_L = 8000$).

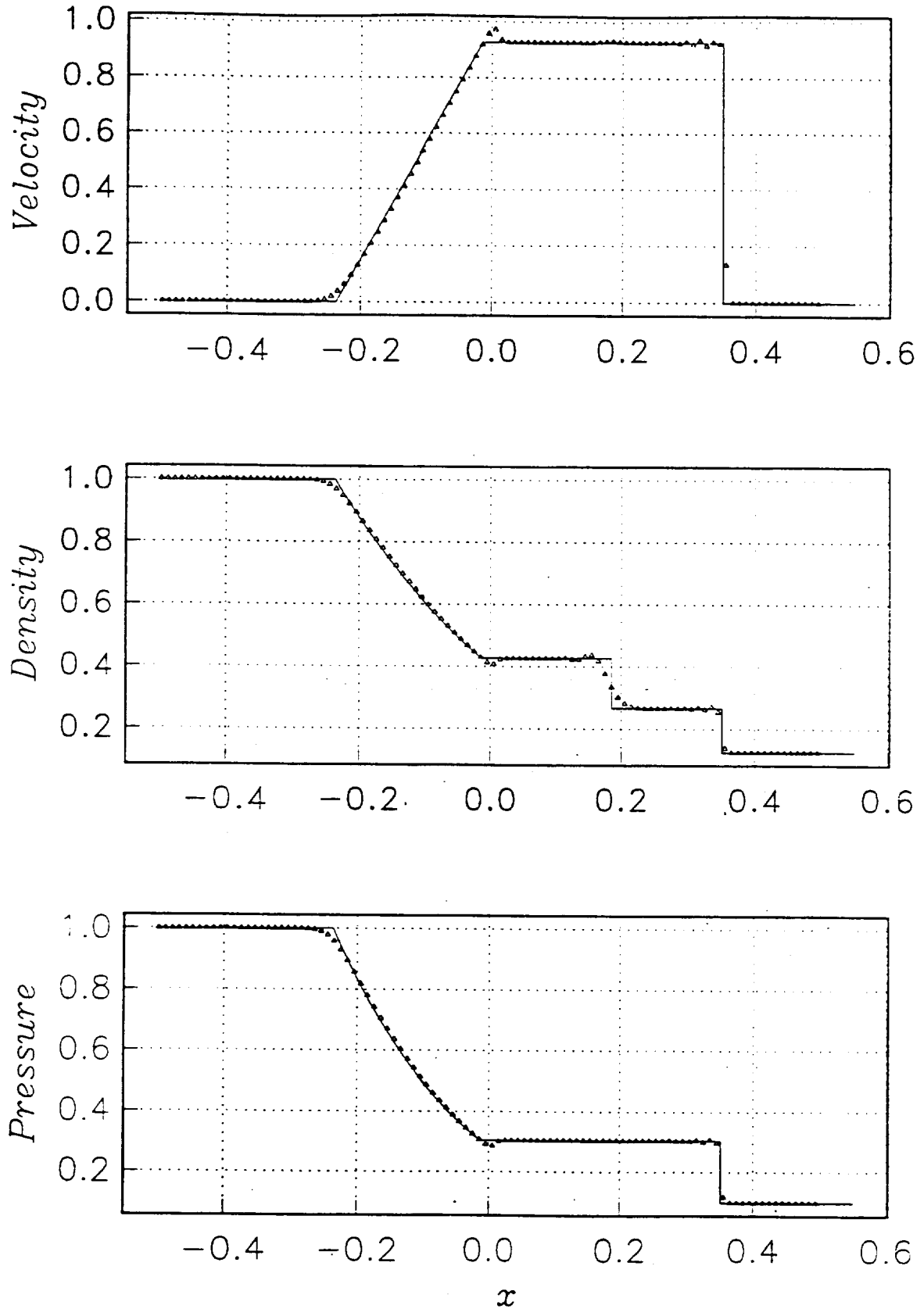


Figure 26.– The Navier–Stokes solution ($Re_L = 10000$).

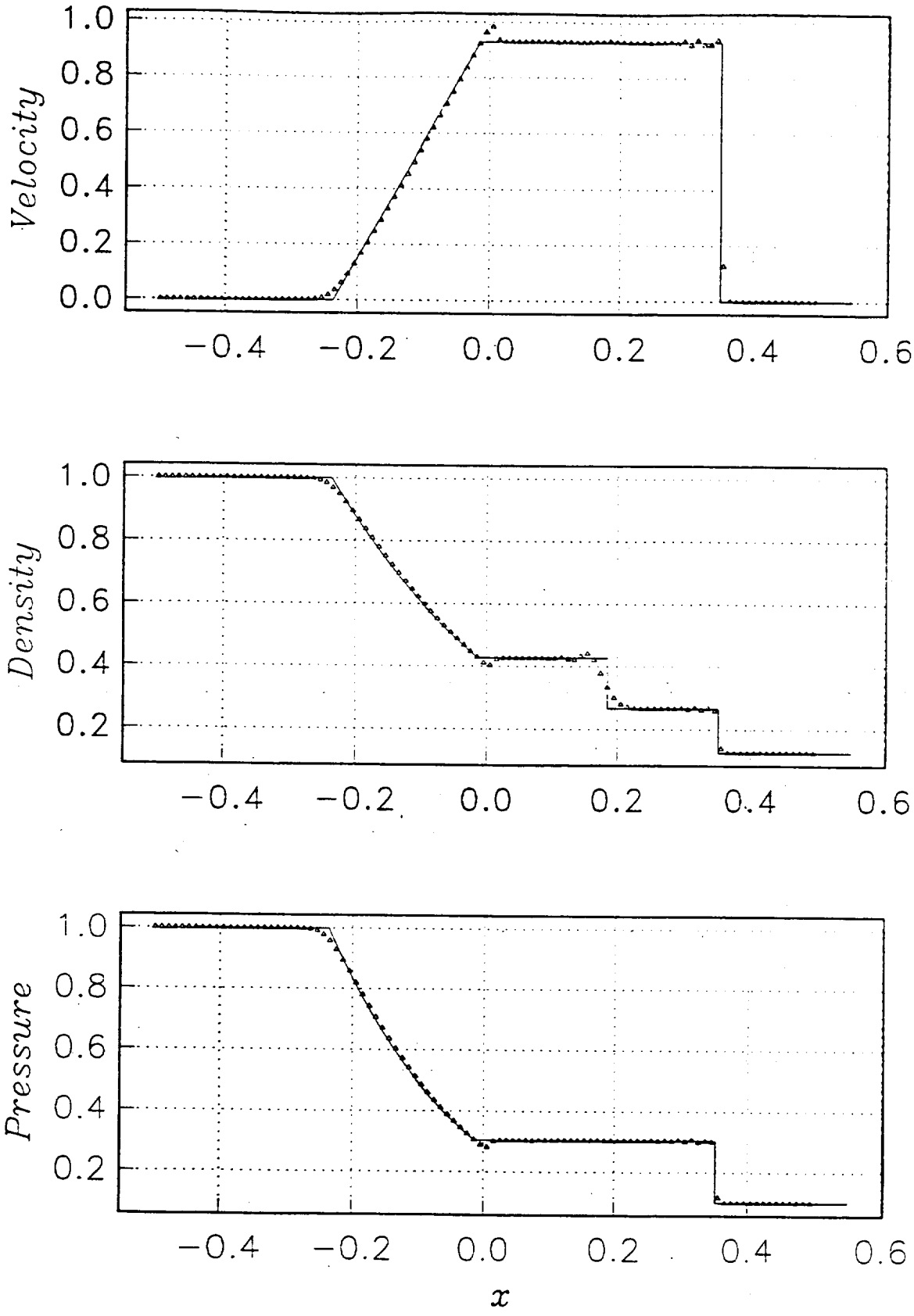


Figure 27.– The Navier–Stokes solution ($Re_L = 12000$).

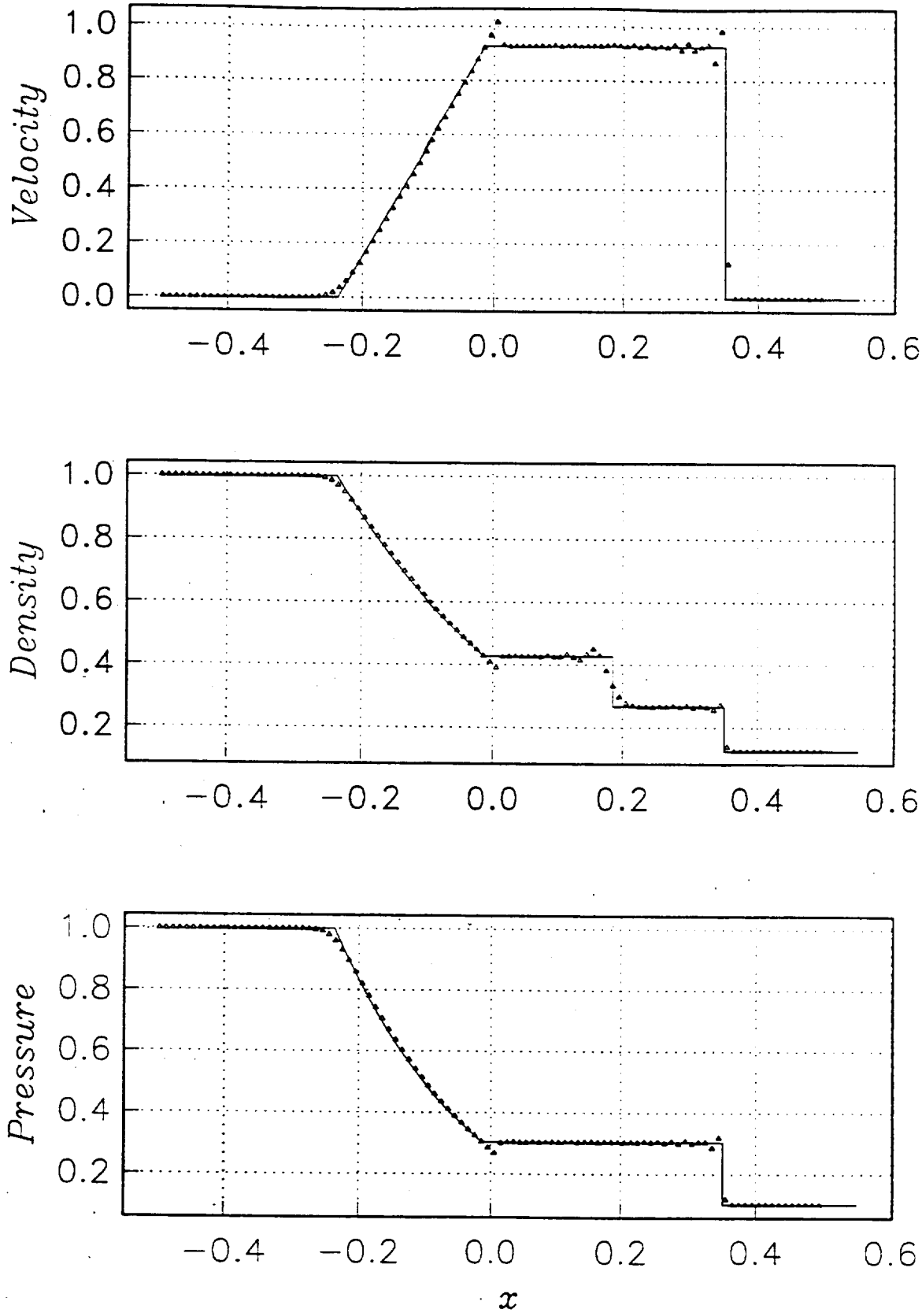
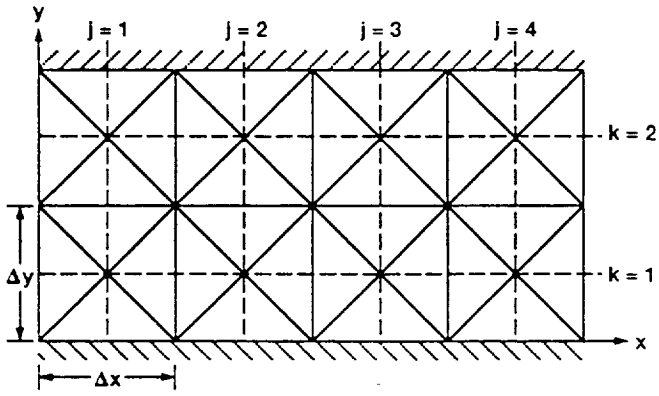
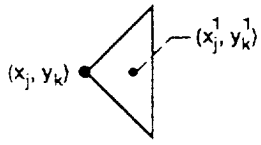


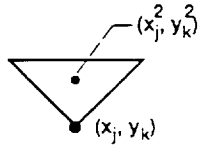
Figure 28.– The Navier-Stokes solution ($Re_L = 20000$).



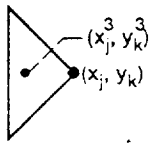
(a) the computational domain.



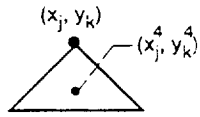
(b) SE (j, k; 1).



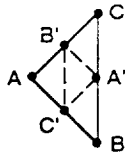
(c) SE (j, k; 2).



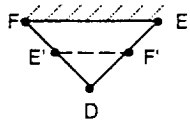
(d) SE (j, k; 3).



(e) SE (j, k; 4).



(f) An interior SE.



(g) A boundary SE.

Figure 29.—The SE's and CE's for a flow problem (first construction).

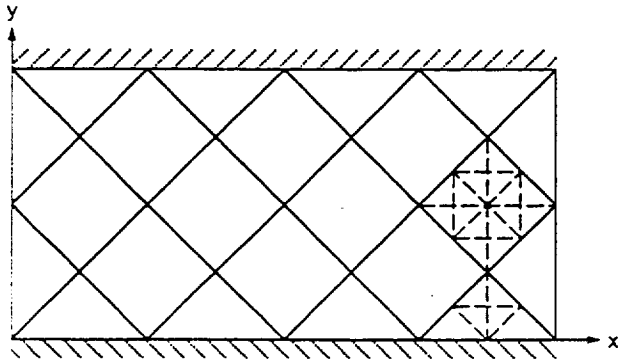


Figure 30.—The SE's and CE's for a flow problem (second construction).

$n' = 0, \pm 1, \pm 2, \ldots$

$j' = 0, \pm 1, \pm 2, \ldots$

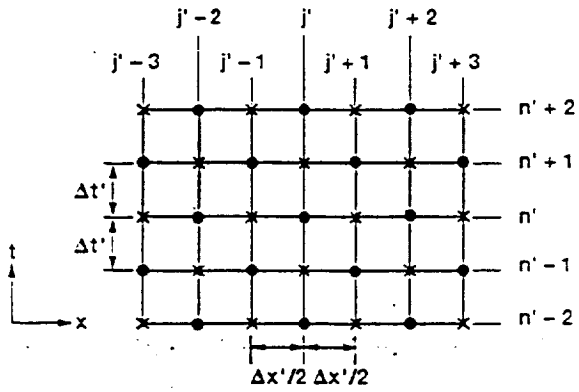


Figure 31.–A regular space-time mesh.

A sample program for solving the shock tube problem

```
      implicit real*8(a-h,o-z)
      dimension  q(3,1000), qn(3,1000), qx(3,1000), qt(3,1000),
     *           s(3,1000), vxl(3), vxr(3), xx(1000)
c
      it = 100
      dt = 0.4d-2
      dx = 0.1d-1
      ga = 1.4d0
      rhol = 1.d0
      ul = 0.d0
      pl = 1.d0
      rhor = 0.125d0
      ur = 0.d0
      pr = 0.1d0
      ic = 1
c
      hdt = dt/2.d0
      tt = hdt*dfloat(it)
      qdt = dt/4.d0
      hdx = dx/2.d0
      qdx = dx/4.d0
      dtx = dt/dx
      a1 = ga - 1.d0
      a2 = 3.d0 - ga
      a3 = a2/2.d0
      a4 = 1.5d0*a1
      q(1,1) = rhol
      q(2,1) = rhol*ul
      q(3,1) = pl/a1 + 0.5d0*rhol*ul**2
      itp = it + 1
      do 5 j = 1,itp
      q(1,j+1) = rhor
      q(2,j+1) = rhor*ur
      q(3,j+1) = pr/a1 + 0.5d0*rhor*ur**2
      do 5 i = 1,3
      qx(i,j) = 0.d0
5     continue
c
      open (unit=8,file='for008')
      write (8,10) tt,it,ic
      write (8,20) dt,dx,ga
      write (8,30) rhol,ul,pl
      write (8,40) rhor,ur,pr
c
      m = 2
      do 400 i = 1,it
      do 100 j = 1,m
      w2 = q(2,j)/q(1,j)
      w3 = q(3,j)/q(1,j)
```


C-2

```
      f21 = -a3*w2**2
      f22 = a2*w2
      f31 = a1*w2**3 - ga*w2*w3
      f32 = ga*w3 - a4*w2**2
      f33 = ga*w2
      qt(1,j) = -qx(2,j)
      qt(2,j) = -(f21*qx(1,j) + f22*qx(2,j) + a1*qx(3,j))
      qt(3,j) = -(f31*qx(1,j) + f32*qx(2,j) + f33*qx(3,j))
      s(1,j) = qdx*qx(1,j) + dtx*(q(2,j) + qdt*qt(2,j))
      s(2,j) = qdx*qx(2,j) + dtx*(f21*(q(1,j) + qdt*qt(1,j)) +
     *  f22*(q(2,j) + qdt*qt(2,j)) + a1*(q(3,j) + qdt*qt(3,j)))
      s(3,j) = qdx*qx(3,j) + dtx*(f31*(q(1,j) + qdt*qt(1,j)) +
     *  f32*(q(2,j) + qdt*qt(2,j)) + f33*(q(3,j) + qdt*qt(3,j)))
100   continue
      mm = m - 1
      do 200 j = 1,mm
      do 200 k =1,3
      qn(k,j+1) = 0.5d0*(q(k,j) + q(k,j+1) + s(k,j) - s(k,j+1))
      vxl(k) = (qn(k,j+1) - q(k,j) - hdt*qt(k,j))/hdx
      vxr(k) = (q(k,j+1) + hdt*qt(k,j+1) - qn(k,j+1))/hdx
      qx(k,j+1) = (vxl(k)*(dabs(vxr(k)))**ic + vxr(k)*(dabs(vxl(k)))
     *      **ic)/((dabs(vxl(k)))**ic + (dabs(vxr(k)))**ic + 1.d-60)
200   continue
      do 300 j = 2,m
      do 300 k = 1,3
      q(k,j) = qn(k,j)
300   continue
      m = m + 1
400   continue
c
      t2 = dx*dfloat(itp)
      xx(1) = -0.5d0*t2
      do 500 j = 1,itp
      xx(j+1) = xx(j) + dx
500   continue
      do 600 j = 1,m
      x = q(2,j)/q(1,j)
      z = a1*(q(3,j) - 0.5d0*x**2*q(1,j))
      write (8,50) xx(j),q(1,j),x,z
600   continue
c
      close (unit=8)
10    format(' t = ',g14.7,' it = ',i4,' ic = ',i4)
20    format(' dt = ',g14.7,' dx = ',g14.7,' gamma = ',g14.7)
30    format(' rhol = ',g14.7,' ul = ',g14.7,' pl = ',g14.7)
40    format(' rhor = ',g14.7,' ur = ',g14.7,' pr = ',g14.7)
50    format(' x =',f8.4,' rho =',g14.7,' u =',g14.7,' p =',g14.7)
      stop
      end
```

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 1993	Technical Memorandum

4. TITLE AND SUBTITLE

New Developments in the Method of Space-Time Conservation Element and Solution-Element – Applications to the Euler and Navier-Stokes Equations

5. FUNDING NUMBERS

WU–505–62–52

6. AUTHOR(S)

Sin-Chung Chang

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135–3191

8. PERFORMING ORGANIZATION REPORT NUMBER

E–7943

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, D.C. 20546–0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA TM–106226

11. SUPPLEMENTARY NOTES

Prepared for the second U.S. National Congress on Computational Mechanics sponosred by the U.S. Association for Computational Mechanics, August 16–18, 1993. Responsible person, Sin-Chung Chang, (216) 433–5874.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Categories 34 and 64

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

A new numerical framework for solving conservation laws is being developed. This new approach differs substantially in both concept and methodology from the well-established methods – i.e., finite difference, finite volume, finite element, and spectral methods. It is conceptually simple and designed to avoid several key limitations to the above traditional methods. An explicit model scheme for solving a simple 1-D unsteady convection-diffusion equation is constructed and used to illuminate major differences between the current method and those mentioned above. Unexpectedly, its amplification factors for the pure convection and pure diffusion cases are identical to those of the Leapfrog and the DuFort-Frankel schemes, respectively. Also, this *explicit* scheme and its Navier-Stokes extension have the unusual property that their stabilities are limited only by the CFL condition. Moreover, despite the fact that it does not use any flux-limiter or slope-limiter, the Navier-Stokes solver is capable of generating highly accurate shock tube solutions with shock discontinuities being resolved within one mesh interval. An accurate Euler solver also is constructed through another extension. It has many unusual properties, e.g., numerical diffusion at all mesh points can be controlled by a set of local parameters.

14. SUBJECT TERMS

Space-time; Conservation element; Solution element

15. NUMBER OF PAGES
97

16. PRICE CODE
A05

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

National Aeronautics and
Space Administration

Lewis Research Center
Cleveland, Ohio 44135

FOURTH CLASS MAIL

ADDRESS CORRECTION REQUESTED