# A New Approach for Constructing Highly Stable High Order CESE Schemes

*Sin-Chung Chang*
*Glenn Research Center, Cleveland, Ohio*

# NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NASA Aeronautics and Space Database and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at *http://www.sti.nasa.gov*

- E-mail your question via the Internet to *help@sti.nasa.gov*

- Fax your question to the NASA STI Help Desk at 443–757–5803

- Telephone the NASA STI Help Desk at 443–757–5802

- Write to:
  NASA Center for AeroSpace Information (CASI)
  7115 Standard Drive
  Hanover, MD 21076–1320

# A New Approach for Constructing Highly Stable High Order CESE Schemes

*Sin-Chung Chang*
*Glenn Research Center, Cleveland, Ohio*

National Aeronautics and
Space Administration

Glenn Research Center
Cleveland, Ohio 44135

August 2010

*Level of Review*: This material has been technically reviewed by technical management.

Available from

# A NEW APPROACH FOR CONSTRUCTING HIGHLY STABLE HIGH ORDER CESE SCHEMES

Sin-Chung Chang

NASA Glenn Research Center, Cleveland, OH 44135

E-mail: sin-chung.chang@nasa.gov

## Abstract

A new approach is devised to construct high order CESE schemes which would avoid the common shortcomings of traditional high order schemes including: (a) susceptibility to computational instabilities; (b) computational inefficiency due to their local implicit nature (i.e., at each mesh points, need to solve a system of linear/nonlinear equations involving all the mesh variables associated with this mesh point); (c) use of large and elaborate stencils which complicates boundary treatments and also makes efficient parallel computing much harder; (d) difficulties in applications involving complex geometries; and (e) use of problem-specific techniques which are needed to overcome stability problems but often cause undesirable side effects. In fact it will be shown that, with the aid of a conceptual leap, one can build from a given 2nd-order CESE scheme its 4th-, 6th-, 8th-,... order versions which **have the same stencil and same stability conditions of the 2nd-order scheme, and also retain all other advantages of the latter scheme**. A sketch of multidimensional extensions will also be provided.

## 1. Introduction

The space-time conservation element and solution element (CESE) method is a high-resolution and genuinely multidimensional method for solving conservation laws [1–63]. Its nontraditional features include: (i) a unified treatment of space and time; (ii) the introduction of conservation elements (CEs) and solution elements (SEs) as the vehicles for enforcing space-time flux conservation; (iii) a time marching strategy that has a space-time staggered stencil at its core and, as such, fluxes at an interface can be evaluated without using any interpolation or extrapolation procedure (which, in turn, leads to the method's ability to capture shocks without using Riemann solvers); (iv) the requirement that each scheme be built from a nondissipative core scheme and, as a result, the numerical dissipation can be controlled effectively; and (v) the fact that mesh values of the physical dependent variables and their spatial derivatives are considered as independent mesh variables to be solve for simultaneously. Note that CEs are non-overlapping space-time subdomains introduced such that (i) the computational domain can be filled by these subdomains; and (ii) flux conservation can be enforced over each of them and also over the union of any combination of them. On the other hand, SEs are space-time subdomains introduced such that (i) the boundary of each CE can be divided into component parts with each of them belonging to a unique SE; and (ii) within a SE, any physical flux vector is approximated using simple smooth functions. In general, a CE does not coincide with a SE.

Without using flux-splitting, dimensional-splitting, mesh-alignment or other special techniques, since its inception in 1991 [1], the unstructured-mesh compatible 2nd-order CESE method has been used to obtain numerous accurate 1D, 2D and 3D steady and unsteady flow solutions with Mach numbers ranging from 0.0028 to 10 [15]. The physical phenomena modeled include traveling and interacting shocks, acoustic waves, vortex shedding, viscous flows, detonation waves, cavitation, flows in fluid film bearings, heat conduction with melting and/or freezing, MHD vortex, hydraulic jump, crystal growth, chromatographic problems and solar wind [3–63]. In particular, its unexpected simple non-reflecting boundary conditions [12,50] and rather unique capability to resolve both strong shocks and small disturbances (e.g., acoustic waves) simultaneously [19,21,22] make the CESE method an effective tool for attacking computational aeroacoustics (CAA) problems. Note that the fact that the second-order CESE method is capable of solving CAA problems accurately is an exception to the commonly-held belief that a second-order scheme is not adequate for modeling CAA

problems. Also note that, while numerical dissipation is needed for shock capturing, it may also result in annihilation of small disturbances. Thus a solver that can handle both strong shocks and small disturbances simultaneously must be able to overcome this difficulty.

In spite of its nontraditional features and robust capabilities, the core ideas of the CESE method are simple. In fact, all of its key features are the results of a pursuit driven by these simple ideas. The first and foremost is the belief that the method must be solid in physics. As such, in the CESE development, conservation laws are enforced locally and globally in their natural space-time unity forms for 1D, 2D and 3D cases. Moreover, because *direct* physical interaction generally occurs only among the immediate neighbors, use of the simplest stencil also becomes a common CESE feature. Obviously, this requirement has the effect of simplifying boundary-condition implementation.

The second idea emerges from the realization that stability and accuracy are two competing issues in time-accurate computations, i.e., too much numerical dissipation will degrade accuracy while too little of it will cause instability. In other words, to meet both accuracy and stability requirements, computation must be performed away from the edge ("cliff") of instability but not too far from it. This represents a real dilemma in numerical method development. As an example, high order schemes generally have higher accuracy and lower numerical dissipation. However, they are susceptible to computational instabilities. In fact, in complicated real-world applications, not only they seldom live up to their nominal order of accuracy— generally they possess only first-order accuracy when shocks are present, stability of these schemes often is difficult to maintain without resorting to problem-specific treatments. To confront this issue head-on, in the CESE development, generally it is required that a solver be built from a nondissipative (i.e., neutrally stable) core scheme. By definition, computations involving a neutrally stable scheme are performed right on the edge of instability and therefore the numerical results generated are nondissipative. As such numerical dissipation can be controlled effectively if the deviation of a solver from its nondissipative core scheme can be adjusted using some built-in parameters.

Moreover, because an accurate viscous flow simulation requires that the numerical dissipation be much smaller than the physical dissipation which decreases as the Reynolds number increases, *in principle, an accurate and robust solver for high-Reynolds-number flows must be able to cut numerical dissipation as the Reynolds number increases*. Obviously this requirement can only be met by a solver built from a nondissipative core solver.

Other CESE ideas are: (i) the flux at an interface be evaluated in a simple and consistent manner; (ii) *genuinely* multidimensional schemes be built as simple, consistent and straightforward extensions of 1D schemes so that *multidimensional shocks can be captured without using any mesh alignment technique*; (iii) triangular and tetrahedral meshes be used in 2D and 3D cases, respectively, so that the method is compatible with the simplest unstructured meshes and thus can be easily used to solve problems with complex geometries; and (iv) logical structures and approximation techniques used be as simple as possible, and special techniques that has only limited applicability and may cause undesirable side effects be avoided. Fortunately for the CESE development, realization of the above lesser ideas (i)–(iv) follows naturally from the first two core ideas.

Building on its initial successes, efforts to refine and improve the CESE method have continued in the past few years. As an example, it was shown in [5] that the numerical dissipation of a dissipative extension of a CESE core scheme may increase to an intolerable level as the value of the $CFL$ number decreases from its maximum stability bound. As such, in a case with a large $CFL$ number disparity (e.g., a simulation with a highly non-uniform spatial mesh and a spatially independent time step), the performance sensitivity with respect to the $CFL$ number can lead to a solution that is highly dissipative in a region where the local $CFL$ number $\ll 1$. Even though a remedy was suggested in [5], a simple and robust solution to this problem had not arrived until a family of new Courant number insensitive schemes [25,31,46–49,51,54,55,58] was developed with fresh insights. Note that these new schemes have one important advantage, i.e., **all variables at each mesh point can be evaluated explicitly without resorting to solving a system of linear/nonlinear algebraic equations involving these local mesh variables even in applications where systems of multidimensional nonlinear PDEs are solved**.

As another example, even though they are accurate enough to solve CAA problems, second-order CESE

solvers are not capable of resolving fine flow structures within a boundary layer without using relatively fine meshes and/or meshes with large aspect ratios. To overcome this limitation, two neutrally stable 4th-order schemes, referred to as the $a(3)$ and $a(4)$ schemes [59–61], were developed in 2006 and intended to serve as the core schemes of other high order CESE schemes. Unfortunately, the dissipative extensions of these new schemes turned out to have the same intractable stability problem that afflicts traditional high order schemes. As such it became obvious to the author that overcoming this problem requires a fresh new prospective.

It will be shown in this paper that, with the aid of a conceptual leap, one can build from a given 2nd-order CESE scheme its 4th-, 6th-, 8th-,... order versions which **have the same stability conditions of the 2nd-order scheme and also retain all other advantages of the latter scheme**. Thus the high order CESE schemes so constructed would avoid the common shortcomings of traditional high order schemes including: (a) susceptibility to computational instabilities; (b) computational inefficiency due to their local implicit nature (i.e., at each mesh point, need to solve a system of linear/nonlinear equations involving all the mesh variables associated with this mesh point); (c) use of large and elaborate stencils which complicates boundary treatments and also makes efficient parallel computing much harder; (d) difficulties in applications involving complex geometries; and (e) use of problem-specific techniques which are needed to overcome stability problems but often cause undesirable side effects.

Because the new high order CESE schemes are built from 2nd-order CESE schemes, the remainder of the paper includes an extensive review of 2nd-order CESE schemes (Sec. 2–4). It is followed by a thorough discussion of several linear and nonlinear high order CESE schemes (Sec. 5–7). A sketch of multidimensional extensions is provided in Sec. 8. The numerical results are presented in Sec. 9. Conclusions and discussions are given in Sec. 10. As a final note, in Sec. 10, we also present a hybrid 3rd-order scheme built from the Lax-Wendroff scheme and a 2nd-order CESE scheme.

## 2. Review of second-order 1D CESE schemes

For simplicity, we review existing 2nd-order CESE solvers for the simple partial differential equation (PDE)

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{2.1}$$

where the advection speed $a \neq 0$ is a constant. Let $x_1 = x$, and $x_2 = t$ be the coordinates of a two-dimensional Euclidean space $E_2$. Then, because Eq. (2.1) can be expressed as $\nabla \cdot \vec{h} = 0$ with $\vec{h} \stackrel{\text{def}}{=} (au, u)$, Gauss' divergence theorem in the space-time $E_2$ implies 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 $E_2$, and (ii) $d\vec{s} = d\sigma\,\vec{n}$ with $d\sigma$ and $\vec{n}$, respectively, being the area and the unit outward normal vector of a surface element on $S(V)$. Note that: (i) because $\vec{h} \cdot d\vec{s}$ is the *space-time* flux of $\vec{h}$ leaving the region $V$ through the surface element $d\vec{s}$, Eq. (2.2) simply states that the total *space-time* flux of $\vec{h}$ leaving $V$ through $S(V)$ vanishes; (ii) in $E_2$, $d\sigma$ is the length of a line segment on the simple closed curve $S(V)$; and (iii) all mathematical operations can be carried out as though $E_2$ were an ordinary two-dimensional Euclidean space.

To proceed, let $\Omega$ denote the set of all space-time staggered mesh points in $E_2$ (dots in Fig. 2(a)), where $n = 0, \pm 1/2, \pm 1, \pm 3/2, \pm 2, \ldots$, and, for each $n$, $j = n \pm 1/2, n \pm 3/2, n \pm 5/2, \ldots$. Each $(j, n) \in \Omega$ is associated with a solution element, i.e., $\mathrm{SE}(j, n)$. By definition, $\mathrm{SE}(j, n)$ is 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. By definition, the end points of the line segments referred to above are excluded from $\mathrm{SE}(j, n)$ so that two SEs will not overlap.

Eq. (2.2) will be simulated numerically assuming that, for any $(x, t) \in \mathrm{SE}(j, n)$, $u(x, t)$ and $\vec{h}(x, t)$, respectively, are approximated by

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

3

and

$$\vec{h}^*(x,t\,;j,n) \stackrel{\text{def}}{=} \left(au^*(x,t\,;j,n),\, u^*(x,t\,;j,n)\right) \qquad (x,t) \in \text{SE}(j,n) \qquad (2.4)$$

Note that: (i) $u_j^n$, $(u_x)_j^n$, and $(u_t)_j^n$ are constants in $\text{SE}(j,n)$, and (in a rough sense) they can be considered as the numerical analogues of the values of $u$, $\partial u/\partial x$, and $\partial u/\partial t$ at the mesh point $(j,n)$, respectively, (ii) $(x_j, t^n)$ are the coordinates of the mesh point $(j,n)$ with $x_j = j\Delta x$ and $t^n = n\Delta t$, and (iii) Eq. (2.4) is the numerical analogue of the definition $\vec{h} = (au, u)$.

Let $u = u^*(x,t\,;j,n)$ satisfy Eq. (2.1) within $\text{SE}(j,n)$. Then one has

$$(u_t)_j^n = -a\,(u_x)_j^n \qquad (2.5)$$

As a result, Eq. (2.3) reduces to

$$u^*(x,t\,;j,n) = u_j^n + (u_x)_j^n\left[(x - x_j) - a\,(t - t^n)\right] \qquad (2.6)$$

i.e., $u_j^n$ and $(u_x)_j^n$ are the only independent mesh variables associated with $(j,n)$. Note that, *for the purpose of uniquely defining the flux over a boundary of a CE, $u^*(x,t\,;j,n)$ is defined here only for $(x,t) \in \text{SE}(j,n)$* for each $(j,n) \in \Omega$. In some later developments, this function will also be used in applications other than flux evaluation such as construction of finite-difference approximations. **Hereafter, for such extended applications, it is to be understood that $u^*(x,t\,;j,n)$ is defined even for a point $(x,t) \notin \text{SE}(j,n)$.**

Let $E_2$ be divided into non-overlapping rectangular regions (see Fig. 2(a)) referred to as basic conservation elements (BCEs). As depicted in Figs. 2(c)–2(e), (i) each $(j,n) \in \Omega$ is assigned with two BCEs, i.e., $\text{CE}_-(j,n)$ and $\text{CE}_+(j,n)$; (ii) each BCE has one and only one pair of diagonally opposite vertices which belong to $\Omega$; (iii) the space-time $E_2$ can be filled by $\text{CE}_-(j,n)$ and $\text{CE}_+(j,n)$, $(j,n) \in \Omega$; and (iv) $\text{CE}(j,n)$, which is the union of $\text{CE}_-(j,n)$ and $\text{CE}_+(j,n)$, is referred to as a compounded conservation element (CCE).

Given the above preliminaries, we are ready to describe the existing CESE solvers for Eq. (2.1).

## 2.1. The $a$ scheme

Note that, among the line segments forming the boundary of $\text{CE}_-(j,n)$, $\overline{AB}$ and $\overline{AD}$ belong to $\text{SE}(j,n)$, while $\overline{CB}$ and $\overline{CD}$ belong to $\text{SE}(j-1/2, n-1/2)$. Similarly, the boundary of $\text{CE}_+(j,n)$ belongs to either $\text{SE}(j,n)$ or $\text{SE}(j+1/2, n-1/2)$. As a result, by imposing the two local flux conservation conditions at each $(j,n) \in \Omega$, i.e.,

$$\oint_{S(\text{CE}_+(j,n))} \vec{h}^* \cdot d\vec{s} = 0 \qquad (j,n) \in \Omega \qquad (2.7)$$

and

$$\oint_{S(\text{CE}_-(j,n))} \vec{h}^* \cdot d\vec{s} = 0 \qquad (j,n) \in \Omega \qquad (2.8)$$

and using Eqs. (2.4) and (2.6), one can obtain two equations for two unknowns $u_j^n$ and $(u_x)_j^n$. In fact, with the aid of

$$\nu \stackrel{\text{def}}{=} a\frac{\Delta t}{\Delta x} \quad \text{and} \quad (u_{\bar{x}})_j^n \stackrel{\text{def}}{=} \frac{\Delta x}{4}(u_x)_j^n \qquad (j,n) \in \Omega \qquad (2.9)$$

it can be shown that (see Comment (a) given at the end of this subsection) Eqs. (2.7) and (2.8) are equivalent to

$$(1-\nu)\left[u + (1+\nu)u_{\bar{x}}\right]_j^n = (1-\nu)\left[u - (1+\nu)u_{\bar{x}}\right]_{j+1/2}^{n-1/2} \qquad (j,n) \in \Omega \qquad (2.7a)$$

and

$$(1+\nu)\left[u - (1-\nu)u_{\bar{x}}\right]_j^n = (1+\nu)\left[u + (1-\nu)u_{\bar{x}}\right]_{j-1/2}^{n-1/2} \qquad (j,n) \in \Omega \qquad (2.8a)$$

respectively. To simplify notation, in the above and hereafter we adopt a convention that can be explained using an expression on the left side of Eq. (2.7a) as an example, i.e.,

$$\left[u + (1+\nu)u_{\bar{x}}\right]_j^n = u_j^n + (1+\nu)(u_{\bar{x}})_j^n$$

4

By adding Eqs. (2.7a) and (2.8a) together, one has

$$u_j^n = \frac{1}{2}\left\{(1-\nu)\left[u-(1+\nu)u_{\bar{x}}\right]_{j+1/2}^{n-1/2} + (1+\nu)\left[u+(1-\nu)u_{\bar{x}}\right]_{j-1/2}^{n-1/2}\right\} \qquad (j,n)\in\Omega \qquad (2.10)$$

Let $1-\nu^2 \neq 0$, i.e., $1-\nu \neq 0$ and $1+\nu \neq 0$. Then Eqs. (2.7a) and (2.7b) can be divided by $(1-\nu)$ and $(1+\nu)$, respectively. By subtracting the resulting equations from each other, one has

$$(u_{\bar{x}})_j^n = (u_{\bar{x}}^a)_j^n \qquad (j,n)\in\Omega \qquad (2.11)$$

with

$$(u_{\bar{x}}^a)_j^n \stackrel{\text{def}}{=} \frac{1}{2}\left\{[u-(1+\nu)u_{\bar{x}}]_{j+1/2}^{n-1/2} - [u+(1-\nu)u_{\bar{x}}]_{j-1/2}^{n-1/2}\right\} \qquad (j,n)\in\Omega \qquad (2.12)$$

It has been shown that Eqs. (2.7) and (2.8) imply (i) Eq. (2.10) for all $\nu$; and (ii) Eq. (2.11) if $|\nu| \neq 1$. On the other hand, it can be shown that Eqs. (2.10) and (2.11) are equivalent to

$$[u+(1+\nu)u_{\bar{x}}]_j^n = [u-(1+\nu)u_{\bar{x}}]_{j+1/2}^{n-1/2} \qquad (j,n)\in\Omega \qquad (2.7b)$$

and

$$[u-(1-\nu)u_{\bar{x}}]_j^n = [u+(1-\nu)u_{\bar{x}}]_{j-1/2}^{n-1/2} \qquad (j,n)\in\Omega \qquad (2.8b)$$

for all $\nu$. Because (i) Eqs. (2.7b) and (2.8b) imply Eqs. (2.7a) and (2.8a), respectively; and (ii) Eqs. (2.7a) and (2.8a) are equivalent to Eqs. (2.7) and (2.8), respectively, one concludes that, for all $\nu$, the last two conservation conditions are enforced by the $a$ scheme which, *by definition, is formed by Eqs. (2.10) and (2.11) for all $\nu$.*

Note that, because

$$\frac{\partial u}{\partial \bar{x}} = \frac{\Delta x}{4}\frac{\partial u}{\partial x}$$

if $\bar{x} \stackrel{\text{def}}{=} x/(\Delta x/4)$, the *normalized* parameter $(u_{\bar{x}})_j^n$ can be interpreted as the value at $(j,n)$ of the derivative of $u$ with respective to the normalized coordinate $\bar{x}$. Also note that: (i) the superscript symbol "a" in the parameter $(u_{\bar{x}}^a)_j^n$ is introduced to remind the reader that Eq. (2.11) is valid for the $a$ scheme, and (ii) the $a$ scheme is second-order accurate in space and time for both $u_j^n$ and the non-normalized parameter $(u_x)_j^n$ [1,5].

The review of the $a$ scheme is concluded with the following remarks:

(a) Evaluation of Eqs. (2.7) and (2.8) can be facilitated by the following observations: because $u^*(x,t\,;j,n)$ is linear in $x$ and $t$, the total flux of $\vec{h}^*$ leaving $\mathrm{CE}_-(j,n)$ or $\mathrm{CE}_+(j,n)$ through any of the four line segments that form its boundary is equal to the scalar product of the vector $\vec{h}^*$ evaluated at the midpoint of the line segment and the "surface" vector (i.e., the unit outward normal vector multiplied by the length) of the line segment.

Note that, by definition, points $B$ and $D$ depicted in Fig. 2(c) do not belong to either $\mathrm{SE}(j,n)$ or $\mathrm{SE}(j-1/2,n-1/2)$. This fact, however, does not pose a problem for flux evaluation over $S\left(\mathrm{CE}_-(j,n)\right)$ because the values of $\vec{h}^*$ at isolated points do not contribute to the flux of $\vec{h}^*$ over a finite line segment. A similar comment applies to points $D$ and $F$ depicted in Fig. 2(d).

(b) Even though they have completely different origins and structures (one is a two-level scheme with two independent variables per mesh point while another is a three-level scheme with a single independent variable per mesh point), the $a$ scheme and the classical leapfrog scheme are deeply related. In fact, it can be shown that *each of the two variables $u_j^n$ and $(u_x)_j^n$ of the $a$ scheme by itself satisfies the same leapfrog scheme automatically [56]. Moreover, after sorting through the differences in the notations used in typical CESE papers and traditional literatures, and also taking into account the decoupled nature of a leapfrog solution, it is shown in [5, pp.301 and 313–315] that the two amplification factors*

5

of the $a$ scheme are identical to those of the neutrally stable leapfrog scheme. Thus, the $a$ scheme is also neutrally stable (i.e., nondissipative) if $|\nu| < 1$ (see the additional discussions given in Sec. 2.2).

(c) Even though it is introduced to model a single PDE (i.e., Eq. (2.1)) with a single dependent variable $u$, the $a$ scheme is formed by two coupled discrete equations (i.e., Eqs. (2.10) and (2.11)) involving two *independent* numerical variables $u_j^n$ and $(u_x)_j^n$. As such one would expect that the $a$ scheme be consistent with a pair of PDEs with one of them being Eq. (2.1). Indeed, it is shown in [1] that, regardless how $\Delta t > 0$ and $\Delta x > 0$ are refined, Eqs. (2.10) and (2.11) are always consistent with Eq. (2.1) and $v - \partial u/\partial x = 0$ with the understanding that $v$ being the analytical version of $(u_x)_j^n$.

(d) Because (i) the vector $\vec{h}^*$ at any surface element lying on any interface separating two neighboring BCE is evaluated using the information from a single SE, and (ii) the unit outward normal vector on the surface element pointing outward from one of these two neighboring BCEs is exactly the negative of that pointing outward from another BCE, one concludes that the flux leaving one of these BCEs through the interface is the negative of that leaving another BCE through the same interface. As a result of this flux cancelation, the local flux conservation relations Eqs. (2.7) and (2.8) lead to a global flux conservation relation, i.e., *the total flux of $\vec{h}^*$ leaving the boundary of any space-time region that is the union of any combination of BCEs will also vanish.* In particular, because $CE(j,n)$ is the union of $CE_-(j,n)$ and $CE_+(j,n)$,

$$\oint_{S(\text{CE}_{(j,n)})} \vec{h}^* \cdot d\vec{s} = 0 \qquad (j,n) \in \Omega \qquad (2.13)$$

must follow from Eqs. (2.7) and (2.8). In fact, *it can be shown that Eq. (2.13) is equivalent to Eq. (2.10).*

(e) In addition to the nondissipative $a$ scheme, as will be shown, there is a family of its dissipative extensions in which *only the less stringent conservation condition Eq. (2.13) is assumed* [5]. Because Eq. (2.13) is equivalent to Eq. (2.10), for each of these extensions, $u_j^n$ is still evaluated using Eq. (2.10) while $(u_{\bar{x}})_j^n$ is evaluated using an equation different from Eq. (2.11).

## 2.2. The $a$-$\epsilon$ scheme and the $c$ scheme

Consider any $(j,n) \in \Omega$. Then $(j \pm 1/2, n - 1/2) \in \Omega$. Let

$$u'^n_{j\pm1/2} \stackrel{\text{def}}{=} u^* \left( x_{j\pm1/2}, t^n; j \pm 1/2, n - 1/2 \right) = [u + (\Delta t/2)u_t]^{n-1/2}_{j\pm1/2} \qquad (2.14)$$

The validity of the last equality sign in Eq. (2.14) follows from Eqs. (2.5) and (2.6). Thus $u'^n_{j\pm1/2}$ is the first-order Taylor approximation at the point $(x_{j\pm1/2}, t^n)$ (which $\notin \text{SE}(j \pm 1/2, n - 1/2)$) evaluated using the mesh variables at $(j \pm 1/2, n-1/2)$. Moreover, with the aid of Eqs. (2.6) and (2.9), one can recast Eq. (2.14) in the form

$$u'^n_{j\pm1/2} = (u - 2\nu\, u_{\bar{x}})^{n-1/2}_{j\pm1/2} \qquad (2.15)$$

Note that, by definition, $(j \pm 1/2, n) \notin \Omega$ if $(j,n) \in \Omega$. Thus $u'^n_{j\pm1/2}$ is associated with a mesh point $\notin \Omega$. The reader is warned that similar situations may occur in the rest of this paper.

Because $u'^n_{j\pm1/2}$ is a first-order Taylor's approximation of $u$ at $(j \pm 1/2, n)$,

$$(u^c_{\bar{x}})_j^n \stackrel{\text{def}}{=} \frac{\Delta x}{4} \left( \frac{u'^n_{j+1/2} - u'^n_{j-1/2}}{\Delta x} \right) \equiv \frac{u'^n_{j+1/2} - u'^n_{j-1/2}}{4} \qquad (2.16)$$

is a central-difference approximation of $\partial u/\partial x$ at $(j,n)$, normalized by the same factor $\Delta x/4$ that appears in Eq. (2.9). Note that: (i) the superscript "$c$" is used to remind the reader of the central-difference origin of the term $(u^c_{\bar{x}})_j^n$; and (ii) by using Eqs. (2.15), (2.16) and (2.12), one has

$$(u^c_{\bar{x}})_j^n = \frac{1}{4} \left[ (u - 2\nu u_{\bar{x}})^{n-1/2}_{j+1/2} - (u - 2\nu u_{\bar{x}})^{n-1/2}_{j-1/2} \right] \qquad (2.17)$$

6

and

$$(u_{\bar{x}}^c - u_{\bar{x}}^a)_j^n = \frac{1}{4}\left[(u + 2u_{\bar{x}})_{j-1/2}^{n-1/2} - (u - 2u_{\bar{x}})_{j+1/2}^{n-1/2}\right] \qquad (2.18)$$

The $a$-$\epsilon$ scheme is formed by Eq. (2.10) and

$$(u_{\bar{x}})_j^n = (u_{\bar{x}}^a)_j^n + 2\epsilon(u_{\bar{x}}^c - u_{\bar{x}}^a)_j^n \qquad (2.19)$$

where $\epsilon$ is an adjustable real number. Note that: (i) with the aid of Eqs. (2.12) and (2.18), Eq. (2.19) can be rewritten as

$$(u_{\bar{x}})_j^n = \frac{1}{2}\left\{[(1-\epsilon)u + (2\epsilon - 1 - \nu)u_{\bar{x}}]_{j+1/2}^{n-1/2} - [(1-\epsilon)u + (1 - \nu - 2\epsilon)u_{\bar{x}}]_{j-1/2}^{n-1/2}\right\} \qquad (2.19a)$$

(ii) the $a$-$\epsilon$ scheme reduces to the $a$ scheme when $\epsilon = 0$; (iii) for the case $\epsilon = 1/2$, Eq. (2.19) reduces to

$$(u_{\bar{x}})_j^n = (u_{\bar{x}}^c)_j^n \qquad (2.20)$$

and (iv) because $(u_{\bar{x}}^c)_j^n$ represents a central-difference approximation, *hereafter, to simplify its frequent references, the special $a$-$\epsilon$ scheme with $\epsilon = 1/2$ will be referred to as the c scheme.*

As a preliminary to later developments, we offer the following remarks on the $a$-$\epsilon$ scheme:

(a) At each $(j, n) \in \Omega$, Eqs. (2.10) and (2.11) imply Eqs. (2.7) and (2.8). Because Eq. (2.19) does not reduce to Eq. (2.11) except in the special case $\epsilon = 0$, at each $(j, n) \in \Omega$, generally the $a$-$\epsilon$ scheme satisfies only the single conservation condition Eq. (2.13) (which is equivalent to Eq. (2.10)) rather than the two conservation conditions Eqs. (2.7) and (2.8). However, because $(u_{\bar{x}}^a)_j^n$ generally is present on the right side of Eq. (2.19), implementation of the $a$-$\epsilon$ scheme generally will still be burdened with the cost of solving two conservation conditions at each mesh point. *The exception occurs only for the special case $\epsilon = 1/2$ (i.e., the c scheme) in which Eq. (2.19) reduces to Eq. (2.20).* As it turns out, implementation of an Euler version of the $c$ scheme does not require inverting any system of equations while that involving an Euler version of any other $a$-$\epsilon$ scheme ($\epsilon \neq 1/2$) generally requires inverting, per mesh point and per time step, a system of several linear equations. As an example, because the 3D Euler equations is formed by a system of five PDEs with each of them being associated with three independent spatial derivatives at each mesh point, a 3D Euler $a$-$\epsilon$ ($\epsilon \neq 1/2$) scheme requires inverting a system of 15 ($= 5 \times 3$) equations. Partly because of this cost consideration, extensions of the $c$ scheme have been used extensively.

(b) For the $a$-$\epsilon$ scheme, it is shown in [5] that the principal and spurious amplification factors per marching step (i.e., per $\Delta t/2$—recall that the solution is advanced by a time period $\Delta t/2$ per marching step) are $\lambda_+$ and $\lambda_-$, respectively, with

$$\lambda_{\pm}(\epsilon, \nu, \theta) \stackrel{\text{def}}{=} \epsilon\cos(\theta/2) - i\nu\sin(\theta/2) \pm \sqrt{(1-\epsilon)\left[(1-\epsilon)\cos^2(\theta/2) + (1-\nu^2)\sin^2(\theta/2)\right]} \qquad (2.21)$$

Here (i) $i \stackrel{\text{def}}{=} \sqrt{-1}$, and (ii) $\theta$, $-\infty < \theta < +\infty$, is the phase angle variation of a Fourier component over the length $\Delta x$ (i.e., $\theta = k\Delta x$ where $k$ denotes the phase angle variation per unit length). In addition, it can be shown that the $a$-$\epsilon$ scheme is stable $\Leftrightarrow$ (i.e., "if and only if") either

$$0 \le \epsilon \le 1 \qquad \text{and} \qquad |\nu| < 1 \qquad (2.22a)$$

or

$$\epsilon = 1 \qquad \text{and} \qquad |\nu| = 1 \qquad (2.22b)$$

(Note: Because Eq. (2.11) is a result of Eqs. (2.7) and (2.8) only if $|\nu| \neq 1$, the $a$ and $a$-$\epsilon$ schemes were *not* defined for the case $|\nu| = 1$ in [5]. As such the case Eq. (2.22a) was not presented in [5]). Also one can show that the $a$-$\epsilon$ scheme becomes more dissipative as the value of $\epsilon$ increases from 0 to 1 [5].

7

Note that, unless specified otherwise, in the remainder of the paper the ranges of $\epsilon$ and $\nu$ respectively, are defined by Eq. (2.22a,b). Moreover, note that: (i) the mesh points used in the $a$-$\epsilon$ scheme are so arranged that *those at any two time levels are staggered in space-time unless these two levels are separated by a whole integer multiple of $\Delta t$*, and (ii) by applying Eqs. (2.10) and (2.19) successively over *two consecutive marching steps*, one can derive a compounded form (Eq. (3.7) in [5]) of the $a$-$\epsilon$ scheme in which the mesh variables at any $(j, n+1) \in \Omega$ are determined in terms of those at $(j-1, n)$, $(j, n)$ and $(j+1, n)$—i.e., the space-time staggering nature of the original scheme is eliminated from the compounded form. As such, (i) Hereafter two consecutive marching steps will be referred to as a *marching cycle*, and (ii) in the following accuracy evaluation of the numerical amplification factors, we choose to consider the amplification factors per marching cycle (i.e., per $\Delta t$).

Note that, for several reasons which are related to the space-time staggered nature of the meshes used in the CESE method (see pp. 313-315 in [5]), the same spatial-interval length and time-step size denoted by $\Delta x$ and $\Delta t$ respectively in traditional literatures, are denoted by $\Delta x/2$ and $\Delta t/2$ respectively here. Because the same definition $\theta \overset{\text{def}}{=} k\Delta x$ is used here and in traditional literatures, *for the same Fourier component (i.e., the same value of $k$), the corresponding value of $\theta$ in traditional literatures is only one half of that in the current paper, i.e., the symbol $\theta/2$ that appears on the right side of Eq. (2.21) would be replaced by $\theta$ in traditional literatures.*

(c) Let $k$ be a constant. Then $u = e^{ik(x-at)}$ represents a plane wave solution to Eq. (2.1). For this solution, $k$ is the phase angle variation per unit length. Thus, by using the relations $\theta = k\Delta x$ and $\nu = a\Delta t/\Delta x$, one concludes that

$$\text{the analytical amplification factor per } \Delta t \overset{\text{def}}{=} \frac{e^{ik[x-a(t+\Delta t)]}}{e^{ik(x-at)}} = e^{-ika\Delta t} = e^{-i\nu\theta} \tag{2.23}$$

(d) The amplification factors of the $a$-$\epsilon$ scheme per $\Delta t$ (i.e., per two consecutive marching steps) are given by $(\lambda_\pm)^2$. According to Eq. (2.21), $[\lambda_\pm(0, \nu, \theta)]^2$, the amplification factors of the $a$ scheme (which corresponds to the case $\epsilon = 0$) per $\Delta t$, have the following properties:

$$|[\lambda_\pm(0, \nu, \theta)]^2| = 1 \qquad \text{if} \quad |\nu| \leq 1 \tag{2.24}$$

$$\lim_{\nu \to \pm 1} [\lambda_+(0, \nu, \theta)]^2 = e^{\mp i\theta} \tag{2.25}$$

$$\lim_{\nu \to \pm 1} [\lambda_-(0, \nu, \theta)]^2 = e^{\pm i\theta} \tag{2.26}$$

and

$$[\lambda_\pm(0, 0, \theta)]^2 = 1 \tag{2.27}$$

On the other hand, $e^{-i\nu\theta}$, the analytical amplification factor per $\Delta t$, has the following properties:

$$|e^{-i\nu\theta}| = 1 \tag{2.28}$$

$$\lim_{\nu \to \pm 1} e^{-i\nu\theta} = e^{\mp i\theta} \tag{2.29}$$

and

$$e^{-i\nu\theta} = 1 \qquad \text{if} \qquad \nu = 0 \tag{2.30}$$

For the $a$ scheme, Eqs. (2.24)–(2.30) imply that: (i) the two amplification factor of the scheme, and the analytical amplification factor all have the same constant absolute value ($= 1$) if $|\nu| \leq 1$, i.e., the scheme is nondissipative if $|\nu| \leq 1$; (ii) in the limit of $|\nu| \to 1$ (i.e., $\nu \to 1$ or $\nu \to -1$), the principal amplification factor is identical to the analytical amplification factor and, thus, the former has no dissipative or dispersive error in this limit; (iii) also in the limit of $|\nu| \to 1$, *the phase angle associated with the spurious amplification factor is exactly the negative of that associated with the analytical amplification factor*—thus the spurious amplification factor has a large dispersive error in

8

this limit except when $|\theta| \ll 1$ (i.e., when the wavelengths of the errors $\gg 1$); and (iv) when $\nu = 0$, the two amplification factors of the scheme, and the exact amplification factor are all equal to 1 and, thus, the two amplification factors of the scheme have no dissipative or dispersive error if $\nu = 0$. Because the accuracy of a scheme is essentially hinged on the behaviors of the principal amplification factor [1], according to the facts stated above, *the a scheme tends to become very accurate when $|\nu|$ approaches 1 or 0*. However, the dispersive errors associated with the nondissipative spurious amplification factor (which could be introduced at $t = 0$ as a result of an inaccurate initial-value specification [1]) *may appear in a solution as persistent numerical wiggles when $|\nu|$ approaches 1*.

(e) According to Eq. (2.21), $[\lambda_\pm(1/2, \nu, \theta)]^2$, the amplification factors of the $c$ scheme (which corresponds to the case $\epsilon = 1/2$) per $\Delta t$, have the following properties:

$$\lim_{\nu \to \pm 1}[\lambda_+(1/2, \nu, \theta)]^2 = e^{\mp i\theta} \tag{2.31}$$

$$\lim_{\nu \to \pm 1}[\lambda_-(1/2, \nu, \theta)]^2 = -\sin^2(\theta/2) \tag{2.32}$$

and

$$[\lambda_\pm(1/2, 0, \theta)]^2 = \frac{1}{2}\left[1 \pm \cos(\theta/2)\sqrt{2 - \cos^2(\theta/2)}\right] \tag{2.33}$$

Because (i) $e^{i\pi} = -1$ and (ii)

$$1 - \left[\cos(\theta/2)\sqrt{2 - \cos^2(\theta/2)}\right]^2 \equiv \left[1 - \cos^2(\theta/2)\right]^2 \geq 0$$

i.e.,

$$1 \geq \left|\cos(\theta/2)\sqrt{2 - \cos^2(\theta/2)}\right|$$

Eqs. (2.32) and (2.33) can be recast as

$$\lim_{\nu \to \pm 1}[\lambda_-(1/2, \nu, \theta)]^2 = \sin^2(\theta/2)e^{i\pi} \tag{2.32a}$$

and

$$[\lambda_\pm(1/2, 0, \theta)]^2 = \frac{1}{2}\left|1 \pm \cos(\theta/2)\sqrt{2 - \cos^2(\theta/2)}\right| \tag{2.33a}$$

respectively. By comparing Eqs. (2.29) and (2.30) with Eqs. (2.31), (2.32a) and (2.33a), one arrives at the following conclusions for the $c$ scheme: (i) in the limit of $|\nu| \to 1$, the principal amplification factor is identical to the analytical amplification factor and, thus, the former has no dissipative or dispersive errors in this limit; (ii) also in the limit of $|\nu| \to 1$, the spurious amplification factor generally has large dissipative and dispersive errors; and (iii) when $\nu = 0$, both the principal and spurious amplification factors generally have large dissipative errors but no dispersive errors. According to the facts stated above, like the $a$ scheme, the $c$ scheme also tends to become very accurate when $|\nu|$ approaches 1. *However, unlike the a scheme, the errors associated with the spurious amplification factor of the c scheme do die out rapidly when $|\nu|$ approaches 1 if the value of $|\theta|$ is not too close to $\pi$ (see Eq. (2.32a)), i.e., if the wavelength of the discrete Fourier component is not close to $2\Delta x$—the shortest possible wavelength. Also, in sharp contrast to the a scheme, the c scheme becomes highly dissipative when $\nu$ approaches 0.*

From the above discussions, one concludes that:

(a) The advantages of the $a$ scheme include: (i) it is nondissipative for all $\theta$ if $|\nu| \leq 1$; and (ii) when the value of $|\nu|$ is close to 0 or 1, the scheme is very accurate. On the other hand, its disadvantages include: (i) because it is nondissipative, its extensions for nonlinear equations generally are unstable; (ii) when the value of $|\nu|$ is close to 1, the dispersive errors associated with the nondissipative spurious amplification

9

factors will appear in a solution as persistent numerical wiggles; and (iii) comparing with the $c$ scheme, it costs more to implement.

(b) The advantages of the $c$ scheme include: (i) when the value of $|\nu|$ is close to 1, it is very accurate and the errors associated with the spurious amplification factor generally die out rapidly; (ii) because of its dissipative property, its extensions for nonlinear equations can be stable; and (iii) in terms of ease of implementation and computer cost, it is much more superior than any other $a$-$\epsilon$ scheme. *On the other hand, the $c$ scheme has a serious disadvantage, i.e., it is very dissipative when $\nu$ approaches* 0.

Based on the remarks given above, it becomes clear that a new solver of Eq. (2.1) possesses all the advantages of the $a$ and $c$ schemes but none of their disadvantages if it is formed by Eq. (2.10) and a new equation in which *(i) $(u_{\bar{x}})_j^n$ is evaluated using a simple central-differencing procedure similar to that used to obtain $(u_{\bar{x}}^c)_j^n$; and (ii) the $(u_{\bar{x}})_j^n$ so obtained is identical to $(u_{\bar{x}}^c)_j^n$ in the limit of $|\nu| \to 1$ and identical to $(u_{\bar{x}}^a)_j^n$ in the limit of $\nu \to 0$.* In other words, such an ideal solver (i) is comparable to the $c$ scheme in ease of implementation; (ii) becomes the $c$ scheme in the limit of $|\nu| \to 1$; and (iii) becomes the $a$ scheme in the limit of $\nu \to 0$. In Sec. 3, it will be shown how such a solver can be constructed.

## 2.3. The $w$-$\alpha$ scheme—a special wiggle-suppressing scheme

If discontinuities are present in a numerical solution, any $a$-$\epsilon$ scheme such as the $c$ scheme is not equipped to suppress numerical wiggles that generally appear near these discontinuities. To serve as a preliminary for future development, here we shall briefly review an extension of the $c$ scheme which was introduced as a remedy for this deficiency [5].

To proceed, let

$$(u_{\bar{x}-})_j^n \overset{\text{def}}{=} \frac{\Delta x}{4}\left(\frac{u_j^n - u_{j-1/2}'^n}{\Delta x/2}\right) \equiv \frac{u_j^n - u_{j-1/2}'^n}{2} \tag{2.34}$$

and

$$(u_{\bar{x}+})_j^n \overset{\text{def}}{=} \frac{\Delta x}{4}\left(\frac{u_{j+1/2}'^n - u_j^n}{\Delta x/2}\right) \equiv \frac{u_{j+1/2}'^n - u_j^n}{2} \tag{2.35}$$

i.e., $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$ are normalized numerical analogues of $\partial u/\partial x$ at $(j, n)$ evaluated from the left and the right, respectively. By using Eqs. (2.16), (2.34), and (2.35), it can be shown that

$$(u_{\bar{x}}^c)_j^n = \frac{1}{2}\left(u_{\bar{x}-} + u_{\bar{x}+}\right)_j^n \tag{2.36}$$

i.e., $(u_{\bar{x}}^c)_j^n$ is the simple average of $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$. As such, the $c$ scheme can be extended by replacing $(u_{\bar{x}}^c)_j^n$ in Eq. (2.20) with an weighted average of $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$. In other words, the resulting extension is formed by Eq. (2.10) and

$$(u_{\bar{x}})_j^n = (w_-)_j^n (u_{\bar{x}-})_j^n + (w_+)_j^n (u_{\bar{x}+})_j^n \tag{2.37}$$

where $(w_-)_j^n$ and $(w_+)_j^n$, the weight factors associated with $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$ respectively, must satisfy the condition

$$(w_-)_j^n + (w_+)_j^n = 1 \tag{2.38}$$

at all $(j, n) \in \Omega$. In addition, it is assumed that the expression on the right side of Eq. (2.37) represents an interpolation (rather than an extrapolation) of $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$. This assumption $\Leftrightarrow$

$$(w_-)_j^n \geq 0 \quad \text{and} \quad (w_+)_j^n \geq 0 \tag{2.39}$$

For real variables $x_-$, $x_+$, and $\alpha \geq 0$, let $W_-$ and $W_+$ be the functions defined by: (i) $W_-(x_-, x_+; \alpha) = W_+(x_-, x_+; \alpha) = 1/2$ if $x_- = x_+ = 0$; and (ii)

$$W_-(x_-, x_+; \alpha) = \frac{|x_+|^\alpha}{|x_-|^\alpha + |x_+|^\alpha} \qquad (|x_-| + |x_+| > 0) \tag{2.40}$$

10

and

$$W_+(x_-, x_+; \alpha) = \frac{|x_-|^\alpha}{|x_-|^\alpha + |x_+|^\alpha} \qquad (|x_-| + |x_+| > 0) \qquad (2.41)$$

if either $x_- \neq 0$ or $x_+ \neq 0$. Furthermore, let

$$(w_\pm)_j^n = W_\pm((u_{\bar{x}-})_j^n, (u_{\bar{x}+})_j^n, \alpha) \qquad (2.42)$$

Then $(w_-)_j^n$ and $(w_+)_j^n$ so defined satisfy Eqs. (2.38) and (2.39) and have the property that

$$(w_-)_j^n = (w_+)_j^n = 1/2 \quad \text{if} \quad \alpha = 0 \quad \text{or} \quad |(u_{\bar{x}-})_j^n| = |(u_{\bar{x}+})_j^n| \qquad (2.43)$$

Note that: (i) to avoid dividing by zero, in practice either a small positive number such as $10^{-20}$ is added to each of the denominators in Eqs. (2.40) and (2.41), or one must make sure that the definition $W_-(x_-, x_+; \alpha) = W_+(x_-, x_+; \alpha) = 1/2$ instead of Eqs. (2.40) and (2.41) is applied when $x_- = x_+ = 0$; and (ii) the special cases of Eqs. (2.40) and (2.41) with $\alpha = 1$ and $\alpha = 2$ are used in the slope-limiter proposed by van Leer [64] and van Albada *et al.* [65].

An extension of the $c$ scheme is formed by Eqs. (2.10) and (2.37) with $(w_-)_j^n$ and $(w_+)_j^n$ being defined by Eq. (2.42). Because it involves an weighted average which is dependent on a parameter $\alpha$, hereafter the scheme is referred to as the $w$-$\alpha$ scheme. In a straightforward manner, the $w$-$\alpha$ scheme has been extended to become the Euler $w$-$\alpha$ scheme. The details are given in [5].

Assuming $\alpha > 0$ and $|(u_{\bar{x}-})_j^n| \neq |(u_{\bar{x}+})_j^n|$, then Eqs. (2.37), (2.40), and (2.41) imply that, of $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$, the one with smaller absolute value is assigned an weight factor $> 1/2$. This observation coupled with Eqs. (2.36)–(2.39) leads to the conclusion that, of $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$, $(u_{\bar{x}})_j^n$ will have an algebraic value closer to the one with smaller absolute value if $(u_{\bar{x}})_j^n$ is evaluated as an weighted average of $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$ according to Eq. (2.37). As a result, $(u_{\bar{x}})_j^n$ so evaluated has a smaller absolute value than that evaluated using Eq. (2.20). In turn, numerical wiggles or overshoots can be annihilated by the additional numerical dissipation introduced as a result of this local "flattening" of $(u_{\bar{x}})_j^n$. It has been shown numerically that the extension is stable if $|\nu| < 1$ and $\alpha \geq 0$. Moreover, as a result of Eqs. (2.20), (2.36), (2.37) and (2.43), (i) the extension reduces to the $c$ scheme when $\alpha = 0$; and (ii) even if $\alpha > 0$, the extension behaves very much like the $c$ scheme in any smooth solution region (where the condition $(u_{\bar{x}-})_j^n = (u_{\bar{x}+})_j^n$ more or less prevails) or at a solution extremum (where the condition $(u_{\bar{x}-})_j^n = -(u_{\bar{x}+})_j^n$ more or less prevails). *As such, the wiggle-suppressing power of the extension takes effect only if $\alpha > 0$ and only in a solution region where $|(u_{\bar{x}-})_j^n|$ and $|(u_{\bar{x}+})_j^n|$ differ substantially.*

This section is concluded with the following discussion of CESE non-reflecting boundary conditions.

## 2.4. Non-reflecting boundary conditions

An unique feature of the CESE method is its simple and robust non-reflecting boundary conditions [12,50]. Let $(j, n)$ be a mesh point on the right spatial boundary of the computational domain. Then (i) $(j - 1/2, n - 1/2)$ represents an interior mesh point at the $(n - 1/2)$th time level; and (ii)

$$u_j^n = u_{j-1/2}^{n-1/2} \quad \text{and} \quad (u_{\bar{x}})_j^n = (u_{\bar{x}})_{j-1/2}^{n-1/2} \qquad (2.44)$$

form a set of non-reflecting boundary conditions. On the other hand, for a mesh point $(j, n)$ on the left boundary, (i) $(j + 1/2, n - 1/2)$ represents an interior mesh point at the $(n - 1/2)$th time level; and (ii)

$$u_j^n = u_{j+1/2}^{n-1/2} \quad \text{and} \quad (u_{\bar{x}})_j^n = (u_{\bar{x}})_{j+1/2}^{n-1/2} \qquad (2.45)$$

form a set of non-reflecting boundary conditions. The above non-reflecting conditions are applicable to all the second-order schemes described in Sec. 2-4. A rigorous justification of these and other simple CESE non-reflecting boundary conditions is given in [50].

## 3. The $c$-$\tau$ and $c$-$\tau^*$ schemes

The ideal solvers of Eq. (2.1) defined at the end of Sec. 2.2 will be developed here. As a preliminary, we shall show that $(u^a_{\bar{x}})^n_j$ can be cast into a central-difference form when $\nu = 0$.

According to Figs. 2(d), $|\overline{EF}| = |\overline{AD}| = 0$ when $\Delta t = 0$. Moreover, because of the assumption $a \neq 0$, $\nu = 0 \Leftrightarrow \Delta t = 0$. Thus, in the case $\nu = 0$, Eq. (2.7) reduce to the condition that the sum of the fluxes leaving $CE_+(j, n)$ through its top face $\overline{AF}$ and bottom face $\overline{ED}$ is zero. Because (i) $\vec{n} = (0, 1)$ at $\overline{AF}$ and $\vec{n} = (0, -1)$ at $\overline{ED}$ ($\vec{n}$ denotes the outward unit normal vector at any boundary of $CE_+(j, n)$), (ii) $|\overline{AF}| = |\overline{ED}|$, and (iii) $\vec{h}^* = (au^*, u^*)$, by using the flux evaluation rule referred to in Remark (a) given at the end of Sec. 2.1, one concludes that Eq. (2.7) reduces to

$$(u + u_{\bar{x}})^n_j = (u - u_{\bar{x}})^{n-1/2}_{j+1/2} \qquad (\nu = 0) \qquad (3.1)$$

if $\nu = 0$. Similarly, Eq. (2.8) reduces to

$$(u - u_{\bar{x}})^n_j = (u + u_{\bar{x}})^{n-1/2}_{j-1/2} \qquad (\nu = 0) \qquad (3.2)$$

if $\nu = 0$. Because Eqs. (2.10) and (2.11) (which form the $a$ scheme) are equivalent to Eqs. (2.7) and (2.8) if $|\nu| \neq 1$, they must be equivalent to Eqs. (3.1) and (3.2) when $\nu = 0$. In fact, by subtracting Eq. (3.2) from Eq. (3.1), one obtains Eq. (2.11) with $(u^a_{\bar{x}})^n_j$ being the reduced form of Eq. (2.12) when $\nu = 0$, i.e.,

$$(u^a_{\bar{x}})^n_j = \frac{1}{2} \left[ (u - u_{\bar{x}})^{n-1/2}_{j+1/2} - (u + u_{\bar{x}})^{n-1/2}_{j-1/2} \right] \qquad (\nu = 0) \qquad (3.3)$$

On the other hand, by summing over Eqs. (3.1) and (3.2), one has the reduced form of Eq. (2.10) for the case $\nu = 0$.

With the aid of Eq. (3.3) and the facts that: (i) $(u - u_{\bar{x}})^{n-1/2}_{j+1/2}$ represents the approximation of $u$ at the midpoint of $\overline{ED}$ while $(u + u_{\bar{x}})^{n-1/2}_{j-1/2}$ represents that at the midpoint of $\overline{CD}$ (see Fig. 2(c,d)); and (ii) the distance between the two midpoints is $\Delta x/2$, one concludes that, for the special case $\nu = 0$, $(u^a_{\bar{x}})^n_j$ is indeed a central-difference approximation of $\partial u/\partial x$ at $(j, n - 1/2)$ (which coincides with $(j, n)$ when $\nu = 0$), normalized by the factor $\Delta x/4$. QED.

According to the above discussions, construction of the ideal solvers is hinged on finding central-difference approximations for $(u_{\bar{x}})^n_j$ such that each approximation (i) becomes $(u^c_{\bar{x}})^n_j$ in the limit of $|\nu| \to 1$, and (ii) reduces to the expression on the right side of Eq. (3.3) when $\nu = 0$. As a result, these new solvers will be constructed as the subschemes of the $c$-$\tau$ scheme, a new class of CESE solvers for Eq. (2.1) to be described immediately.

### 3.1. The $c$-$\tau$ scheme

To proceed, refer to Fig. 3. Here $M^+$ and $M^-$ denote the midpoints of $\overline{AF}$ and $\overline{AB}$, respectively. Also $P^+$ and $P^-$ are two points on $\overline{BF}$ that satisfy the following conditions: (i) $P^+$ is to the right (left) of $M^+ \Leftrightarrow P^-$ is to the left (right) of $M^-$; and (ii) $|\overline{M^+P^+}| = |\overline{M^-P^-}|$. In addition, the parameter $\tau$ is defined by: (i) $\tau = 0$ if $P^+$ coincides with $M^+$; (ii) $\tau \Delta x/4 = |\overline{M^+P^+}| > 0$ if $P^+$ is to the right of $M^+$; and (iii) $\tau \Delta x/4 = -|\overline{M^+P^+}| < 0$ if $P^+$ is to the left of $M^+$. Obviously, it follows from the above definitions that (i) $\tau = 0$ if $P^-$ coincides with $M^-$; (ii) $\tau \Delta x/4 = |\overline{M^-P^-}| > 0$ if $P^-$ is to the left of $M^-$; and (iii) $\tau \Delta x/4 = -|\overline{M^-P^-}| < 0$ if $P^-$ is to the right of $M^-$.

Next let

$$u(P^+) \overset{\text{def}}{=} u^* \left( x(P^+), t^n; j + 1/2, n - 1/2 \right) \qquad (3.4)$$

and

$$u(P^-) \overset{\text{def}}{=} u^* \left( x(P^-), t^n; j - 1/2, n - 1/2 \right) \qquad (3.5)$$

where $x(P^+)$ and $x(P^-)$ denote the x-coordinates of $P^+$ and $P^-$, respectively. Then, according to Eqs. (2.5) and (2.6), $u(P^+)$ is the first-order Taylor's approximation of $u$ at $P^+$ evaluated using the mesh variables

at $(j + 1/2, n - 1/2)$ (i.e., point $E$) while $u(P^-)$ is that at $P^-$ evaluated using the mesh variables at $(j - 1/2, n - 1/2)$ (i.e., point $C$). Moreover, with the aid of Eqs. (2.6) and (2.9) along with the relations

$$x(P^+) = x_j + (1 + \tau)\frac{\Delta x}{4} = x_{j+1/2} - (1 - \tau)\frac{\Delta x}{4}, \qquad -\infty < \tau < +\infty \qquad (3.6)$$

and

$$x(P^+) = x_j - (1 + \tau)\frac{\Delta x}{4} = x_{j-1/2} + (1 - \tau)\frac{\Delta x}{4}, \qquad -\infty < \tau < +\infty \qquad (3.7)$$

Eqs. (3.4) and (3.5), respectively, can be simplified as

$$u(P^+) = [u - (1 + 2\nu - \tau)u_{\bar{x}}]_{j+1/2}^{n-1/2} \qquad (3.8)$$

and

$$u(P^-) = [u + (1 - 2\nu - \tau)u_{\bar{x}}]_{j-1/2}^{n-1/2} \qquad (3.9)$$

To proceed, note that: (i) the mesh point $(j, n)$ (i.e., point $A$ depicted in Fig. 3) is the midpoint of $\overline{P^- P^+}$, and (ii) $|\overline{P^- P^+}| = (1 + \tau)\Delta x/2$. Thus

$$(\hat{u}_{\bar{x}})_j^n \overset{\text{def}}{=} \frac{\Delta x}{4}\left[\frac{u(P^+) - u(P^-)}{(1 + \tau)\Delta x/2}\right] \equiv \frac{u(P^+) - u(P^-)}{2(1 + \tau)} \qquad (\tau \neq -1) \qquad (3.10)$$

represents a central-difference approximation of $\partial u/\partial x$ at the mesh point $(j, n)$, normalized by the factor $\Delta x/4$. Thus the new scheme defined by Eq. (2.10) and $(u_{\bar{x}})_j^n = (\hat{u}_{\bar{x}})_j^n$ represents a solver for Eq. (2.1). Moreover, by using Eqs. (3.8)–(3.10), the second component of the new scheme can be explicitly expressed as

$$(u_{\bar{x}})_j^n = (\hat{u}_{\bar{x}})_j^n \equiv \frac{1}{2(1 + \tau)}\left\{[u - (1 + 2\nu - \tau)u_{\bar{x}}]_{j+1/2}^{n-1/2} - [u + (1 - 2\nu - \tau)u_{\bar{x}}]_{j-1/2}^{n-1/2}\right\} \qquad (\tau \neq -1) \quad (3.11)$$

In view of the facts that (i) $(\hat{u}_{\bar{x}})_j^n$ represents a central-difference approximation of $(u_{\bar{x}})_j^n$, and (ii) the approximation is associated with the parameter $\tau$, hereafter the new scheme will be referred to as the $c$-$\tau$ scheme.

Moreover, by using Eqs. (2.12), (2.18) and (3.11), one has

$$(\hat{u}_{\bar{x}})_j^n = (u_{\bar{x}}^a)_j^n + \frac{2\tau}{1 + \tau}(u_{\bar{x}}^c - u_{\bar{x}}^a)_j^n - \frac{\nu(1 - \tau)}{2(1 + \tau)}\left[(u_{\bar{x}})_{j+1/2}^{n-1/2} - (u_{\bar{x}})_{j-1/2}^{n-1/2}\right] \qquad (\tau \neq -1) \qquad (3.12)$$

By comparing Eq. (3.12) with (2.19), *one concludes that the $c$-$\tau$ scheme generally is different from the $a$-$\epsilon$ scheme*. In fact, the $c$-$\tau$ scheme becomes the $a$-$\epsilon$ scheme $\Leftrightarrow$ either (i) $\tau = 1$ or (ii) $\nu = 0$. For the case $\tau = 1$, Eq. (3.12) implies that $(\hat{u}_{\bar{x}})_j^n = (u_{\bar{x}}^c)_j^n$. In other words, the $c$ scheme is the special case of the $c$-$\tau$ scheme with $\tau = 1$, a fact that can also be deduced from the observation that the points $P^+$ and $P^-$ depicted in Fig. 3, respectively, coincide with points F and B (i.e., the mesh points $(j + 1/2, n)$ and $(j - 1/2, n)$) if $\tau = 1$. On the other hand, when $\nu = 0$, the $c$-$\tau$ scheme become the $a$-$\epsilon$ scheme with $\epsilon = 2\tau/(1 + \tau)$. In fact one can further deduce that $c$-$\tau$ scheme reduces to the $a$ scheme $\Leftrightarrow \nu = \tau = 0$.

Because the $c$-$\tau$ scheme is formed by two rather complicated equations (i.e., Eqs. (2.10) and (3.11)) involving two parameters $\nu$ and $\tau$, it were not expected that its von Neumann stability conditions could be cast into an *explicit analytical form*. However, by using the Jordan canonical form theorem [66], it is shown rigorously in [46] that the $c$-$\tau$ scheme is von Neumann stable $\Leftrightarrow$

$$\nu^2 \leq 1, \quad \tau \geq \tau_o(\nu^2), \quad \text{and} \quad (\nu^2, \tau) \neq (1, 1) \qquad (3.13)$$

13

where

$$\tau_o(s) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } s = 0 \\ \dfrac{4 - s - 2\sqrt{2(2 - s - s^2)}}{s} & \text{if } 0 < s \leq \tfrac{3}{11} \\ \dfrac{s - 1 + \sqrt{1 - 2s + 5s^2}}{2s} & \text{if } \tfrac{3}{11} \leq s \leq 1 \end{cases} \tag{3.14}$$

Note that:

(a) The current stability conditions agree completely with those generated through numerical experiments [31] (Note: the reader is warned that the condition $\tau \geq \tau_o(\nu^2)$ that appears in Eq. (3.13) here is expressed as $\tau \geq \tau_o(|\nu|)$ in Eq. (3.12) of [31], i.e., the function $\tau_o$ defined here is *different* from the function $\tau_o$ introduced in [31]).

(b) It can be shown that [46]: (i) $\tau_o(s)$ is continuous at $s = 0$, i.e.,

$$\lim_{s \to 0^+} \tau_o(s) = \tau_o(0) = 0$$

(ii) $\tau_o(s)$ is consistently defined at $s = 3/11$, i.e., both the second and third expressions on the right side of Eq. (3.14) implies that $\tau_o(3/11) = 1/3$; (iii) $\tau_o(s)$ is continuously differentiable at $s = 3/11$, i.e.,

$$\lim_{s \to \frac{3}{11}^-} \tau_o'(s) = \lim_{s \to \frac{3}{11}^+} \tau_o'(s) = 121/90$$

where $\tau_o'(s) \overset{\text{def}}{=} d\tau_o(s)/ds$; (iv) $\tau_o(s)$ is strictly monotonically increasing in the interval $0 < s < 1$; (v) $\tau_o(1) = 1$; and (vi)

$$s < \tau_o(s) < \sqrt{s}, \qquad 0 < s < 1 \tag{3.15}$$

(c) For any given fixed value of $|\nu| < 1$, the $c$-$\tau$ scheme tends to become more dissipative as the value of $\tau$ increases from its minimum stability bound $\tau_o(\nu^2)$.

With the above preliminaries, the ideal solvers of Eq. (2.1) will be constructed in Sec. 3.2.

**3.2. The $c$-$\tau^*$ schemes**

The value of $\tau$ used in the $c$-$\tau$ scheme generally can be chosen independent of $\nu$. Here we will introduce a subset of the $c$-$\tau$ scheme in which $\tau$ is a function of $\nu^2$ for each member of this subset. As a preliminary, note that, by using the properties of $\tau_o(s)$ presented earlier, it can be shown that there exist infinitely many choices of *a strictly monotonically increasing smooth function $h(s)$, $0 \leq s < 1$*, which possesses the following properties: (i)

$$h(0) = 0; \quad \text{and} \quad \lim_{s \to 1^-} h(s) = 1 \tag{3.16}$$

and (ii)

$$h(s) \geq \tau_o(s) \quad \text{if} \quad 0 < s < 1 \tag{3.17}$$

By definition, the subset contains every special $c$-$\tau$ scheme with

$$\tau = h(\nu^2) \qquad (|\nu| < 1) \tag{3.18}$$

where $h(s)$ is one of the special functions defined above. By using Eqs. (3.16)–(3.18) and the fact that $h(s)$ is a strictly monotonically increasing function in the domain $0 \leq s < 0$, one concludes that, for each member of the subset, (i)

$$\tau = 0 \quad \text{if} \quad \nu = 0 \tag{3.19}$$

(ii)

$$\lim_{\nu^2 \to 1^-} \tau = 1 \tag{3.20}$$

14

(iii) $\tau$ increases with $|\nu|$ in the domain $0 \le |\nu| < 1$; and (iv)

$$\tau \ge \tau_o(\nu^2) \qquad (|\nu| < 1) \tag{3.21}$$

By using the results presented in the above items (i)–(iii), one can easily infer from Fig. 3 a simple relation between the value of $|\nu|$ and the locations of $P^+$ and $P^-$, i.e., as the value of $|\nu|$ increases from 0 to 1, $P^+$ will move away from $M^+$ and edge toward the mesh point $(j + 1/2, n)$ while $P^-$ will move away from point $M^-$ and edge toward the mesh point $(j - 1/2, n)$.

Recall that (i) $(\hat{u}_{\bar{x}})_j^n = (u_{\bar{x}}^a)_j^n$ if $\tau = \nu = 0$; and (ii) $(\hat{u}_{\bar{x}})_j^n = (u_{\bar{x}}^c)_j^n$ if $\tau = 1$. As such, Eqs. (3.19) and (3.20) imply that, for each member in the subset, (i) $(\hat{u}_{\bar{x}})_j^n = (u_{\bar{x}}^a)_j^n$ if $\nu = 0$; and (ii) $(\hat{u}_{\bar{x}})_j^n = (u_{\bar{x}}^c)_j^n$ in the limit of $|\nu| \to 1^-$. In other words, in the domain $|\nu| < 1$, all members in the subset are the ideal solvers defined at the end of Sec. 2.2. Moreover, by using Eq. (3.13) and (3.21), one can also show that *these ideal solvers are also stable in the same domain*. Hereafter, each of these ideal solvers will be referred to as a $c\text{-}\tau^*$ scheme.

Corresponding to infinitely many choices of $h$, there are infinitely many different $c\text{-}\tau^*$ schemes. In particular, with the aid of Eq. (3.15), one can see easily that the special choice

$$h(s) = \sqrt{s} \qquad (0 \le s < 1) \tag{3.22}$$

represents a strictly monotonically increasing smooth function which also satisfies Eqs. (3.16) and (3.17). Thus Eq. (3.18) implies that a $c\text{-}\tau^*$ scheme can be defined using the relation

$$\tau = |\nu| \qquad (|\nu| < 1) \tag{3.23}$$

Any $c\text{-}\tau^*$ scheme described above meets all the requirements of an ideal solver. However, because of stability problem, a $c\text{-}\tau^*$ scheme and its multidimensional extensions may not be robust enough for some complicated real-world applications. To overcome this difficulty, we introduce a special $c\text{-}\tau$ scheme which is defined by Eqs. (2.10) and (3.11) with

$$\tau = \beta\,|\nu| \qquad (\beta \ge 1;\ |\nu| < 1) \tag{3.24}$$

where $\beta \ge 1$ is a constant parameter. For this scheme, Eq. (3.19) is also valid. Thus it becomes the nondissipative $a$ scheme when $\nu = 0$ and, therefore, it will not become overly dissipative as $\nu \to 0$ (i.e., the scheme's numerical dissipation is $CFL$ number insensitive). Moreover, because $\beta \ge 1$, Eqs. (3.13) and (3.15) along with Comment (c) given following Eq. (3.15) imply that the scheme is stable if $|\nu| < 1$ and it becomes more dissipative (i.e., more stable) as the value of $\beta$ increases. However, also because $\beta \ge 1$,

$$\lim_{|\nu| \to 1^-} \tau = \beta \ge 1 \tag{3.25}$$

As such, except the special case $\beta = 1$, the scheme does not become the $c$ scheme in the limit of $|\nu| \to 1^-$, i.e., it does not meet all the requirement of an ideal solver.

Note that, **unless specified otherwise, in the remainder of this paper we consider only the $c\text{-}\tau^*$ schemes or the special $c\text{-}\tau$ scheme defined by Eq. (3.25).**

Next we will introduce weighted-averaging extensions of the $c\text{-}\tau$ schemes along with two classes of advanced weighted-averaging techniques.

## 4. Other $c\text{-}\tau^*$ extensions and new weighted-averaging techniques

To proceed, let

$$(\hat{u}_{\bar{x}-})_j^n \overset{\text{def}}{=} \frac{\Delta x}{4}\left(\frac{u_j^n - u(P^-)}{(1+\tau)\Delta x/4}\right) \equiv \frac{u_j^n - u(P^-)}{(1+\tau)} \tag{4.1}$$

and

$$(\hat{u}_{\bar{x}+})_j^n \overset{\text{def}}{=} \frac{\Delta x}{4}\left(\frac{u(P^+) - u_j^n}{(1+\tau)\Delta x/4}\right) \equiv \frac{u(P^+) - u_j^n}{(1+\tau)} \tag{4.2}$$

Because $|\overline{AP_-}| = |\overline{AP_+}| = (1+\tau)\Delta x/4$ (see Fig. 3), it is easy to see that $(\hat{u}_{\bar{x}-})_j^n$ and $(\hat{u}_{\bar{x}+})_j^n$ are two normalized one-sided difference approximations of $\partial u/\partial x$ at the mesh point $(j, n)$ with one being evaluated from the left and another from the right. Also, it follows immediately from Eqs. (3.10), (4.1) and (4.2) that

$$(\hat{u}_{\bar{x}})_j^n = \frac{1}{2}\left[(\hat{u}_{\bar{x}-})_j^n + (\hat{u}_{\bar{x}+})_j^n\right] \tag{4.3}$$

Moreover, by (i) substituting Eqs. (2.10), (3.8) and (3.9) into Eqs.(4.1) and (4.2), and (ii) using Eqs. (3.11) and (3.19), one arrives at the conclusion that

$$(\hat{u}_{\bar{x}-})_j^n = (\hat{u}_{\bar{x}+})_j^n = (\hat{u}_{\bar{x}})_j^n \qquad (\nu = 0) \tag{4.4}$$

when $\nu = 0$.

With the above preliminaries, several extensions of the $c$-$\tau^*$ scheme will be constructed in the following subsections.

### 4.1. Scheme w-1

A comparison of Eqs. (4.1)–(4.3) with Eqs. (2.34)–(2.36) reveals that an obvious extension of the $c$-$\tau^*$ scheme can be obtained by replacing $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$ in Eqs. (2.37) and (2.42) with $(\hat{u}_{\bar{x}-})_j^n$ and $(\hat{u}_{\bar{x}+})_j^n$, respectively. In other words, the new extension is formed by Eq. (2.10) and

$$(u_{\bar{x}})_j^n = (w_-)_j^n(\hat{u}_{\bar{x}-})_j^n + (w_+)_j^n(\hat{u}_{\bar{x}+})_j^n \tag{4.5}$$

with

$$(w_\pm)_j^n = W_\pm((\hat{u}_{\bar{x}-})_j^n, (\hat{u}_{\bar{x}+})_j^n, \alpha) \tag{4.6}$$

Because the scheme is the first extension of the $c$-$\tau^*$ scheme in which $(u_{\bar{x}})_j^n$ is expressed as an weighted average of $(\hat{u}_{\bar{x}-})_j^n$ and $(\hat{u}_{\bar{x}+})_j^n$, for simplicity, hereafter it will be referred to as Scheme w-1. It has been shown numerically that Scheme w-1 is stable if $|\nu| < 1$ and $\alpha \geq 0$. However, as will be shown, Scheme w-1 is deficient in one key aspect.

Eqs. (4.4) and (4.6) along with the definition of $W_\pm$ imply that, for any given $\alpha \geq 0$, $(w_-)_j^n \to 1/2$ and $(w_+)_j^n \to 1/2$ as $\nu \to 0$. In other words, for Scheme w-1, the "weighted" average on the right side of Eq. (4.5) will approach the simple average as $\nu \to 0$. According to an explanation given in the last paragraph of Sec. 2.3, this implies that Scheme w-1 will lose its capability to suppress wiggles or overshoots when $\nu$ becomes small. For this reason, even though the Euler version of Scheme w-1 performs much better than that of the $w$-$\alpha$ scheme in its ability to resolve shocks and contact discontinuities crisply in a wide range (from 1 to less than 0.001) of the global $CFL$ number (i.e., the maximal value of local $CFL$ numbers), it is handicapped by the fact that wiggles or overshoots can appear near a discontinuity in a generated solution when the local $CFL$ number there becomes extremely small. In the following, it will be shown that this weakness can be overcome by simple modifications of Eq. (4.6).

### 4.2. Scheme w-2

A new scheme, referred to as Scheme w-2 is formed by Eqs. (2.10) and (4.5) with $(w_\pm)_j^n$ being given by Eq. (2.42). In other words, although $(u_{\bar{x}})_j^n$ is still constructed as an weighted average of $(\hat{u}_{\bar{x}-})_j^n$ and $(\hat{u}_{\bar{x}+})_j^n$, the associated weight factors $(w_\pm)_j^n$ are evaluated using $(u_{\bar{x}-})_j^n$ and $(u_{\bar{x}+})_j^n$ (which are defined in Eqs. (2.34) and (2.35), respectively). Because the last two parameters, respectively, are identical to the special cases of $(\hat{u}_{\bar{x}-})_j^n$ and $(\hat{u}_{\bar{x}+})_j^n$ with $\tau \equiv 1$ (see Eqs. (2.15), (2.34), (2.35), (3.8), (3.9), (4.1), and (4.2)), their values *do not* vary with $\nu$. As such, $(w_\pm)_j^n \neq 1/2$ and therefore the weighted average on the right side of Eq. (4.5) will not turn into a simple average when $\nu = 0$. In other words, Scheme w-2 is still capable of annihilating the

numerical wiggles near a discontinuity even if $\nu$ becomes small. It has been shown numerically that Scheme w-2 again is stable if $|\nu| < 1$ and $\alpha \geq 0$.

Note that a possible drawback of Scheme w-2 is that the relation $|(u_{\bar{x}-})^n_j| < |(u_{\bar{x}+})^n_j|$ ($|(u_{\bar{x}-})^n_j| > |(u_{\bar{x}+})^n_j|$) does not automatically follow from $|(\hat{u}_{\bar{x}-})^n_j| < |(\hat{u}_{\bar{x}+})^n_j|$ ($|(\hat{u}_{\bar{x}-})^n_j| > |(\hat{u}_{\bar{x}+})^n_j|$) and vice versa. As a result, at some local mesh points, it may happen that, of $(\hat{u}_{\bar{x}-})^n_j$ and $(\hat{u}_{\bar{x}+})^n_j$, the one with smaller absolute value may not be associated with a weight factor $> 1/2$. According to a discussion given in the last paragraph of Sec. 2.3, this implies that there is no guarantee that, at all localities, the weighted-averaging induced numerical dissipation will be available to suppress wiggles or overshoots. Despite this possible failing, fortunately it has been demonstrated numerically that, not only are they capable of suppressing wiggles or overshoots robustly, Scheme w-2 and its Euler extensions are also highly accurate.

In the following, schemes that overcome the weakness of Scheme w-1 and also avoid the theoretically possible failing associated with Scheme w-2 will be constructed using two new classes of weighted-averaging techniques more advanced than that given in Eqs. (2.40) and (2.41).

## 4.3. New classes of weighted-averaging techniques

To pave the way, first we shall discuss a limitation of Eqs. (2.40) and (2.41) as a generator of weight factors.

Let $x_\pm \neq 0$. Then, for a given $\alpha > 0$, obviously $W_- \to 1/2$ and $W_+ \to 1/2$ as $|x_+/x_-| \to 1$. As such, when $|x_+/x_-|$ is very close to 1, then both $W_-$ and $W_+$ will be very close to $1/2$ *unless $\alpha \gg 1$*. As a result, in case that (i) $(\hat{u}_{\bar{x}\pm})^n_j \neq 0$, (ii) $|(\hat{u}_{\bar{x}+})^n_j/(\hat{u}_{\bar{x}-})^n_j|$ is very close to 1 (e.g., when $\nu$ is very close to 0); and (iii) Eqs. (4.6) is assumed, then the only way to prevent the weighted average that appears on the right side of Eq. (4.5) from becoming almost a simple average (and thus causing the overshoot problem discussed in Sec. 4.1) is to increase the value of $\alpha$ used. However, *this approach may be impracticable because numerical evaluation of a quantity such as $x^\alpha$ for any real number $x$ generally is hampered by round-off errors and thus becomes highly inaccurate if the value of $\alpha$ becomes too large, say 100.*

To overcome the limitation discussed above, two classes of weighted-averaging techniques much more potent and flexible than that discussed in Sec. 2.3 were presented in Sec. 4.3 of [31]. These new techniques can be used to constructed weighted averages of any set of $N$ ($> 1$) parameters. By applying these new techniques for the special case $N = 2$ [62], one can construct two more extensions of the $c$-$\tau^*$ scheme. These extensions, referred to as Schemes w-3 and w-4, are presented in Sec. 4.4 and 4.5, respectively.

## 4.4. Scheme w-3

Let

$$\sigma = \frac{\sigma_o}{|\nu|} \qquad (\sigma_o > 0 \text{ and } \nu \neq 0) \tag{4.7}$$

$$\xi(j, n; \alpha) \overset{\text{def}}{=} \frac{|(\hat{u}_{\bar{x}+})^n_j|^\alpha - |(\hat{u}_{\bar{x}-})^n_j|^\alpha}{|(\hat{u}_{\bar{x}+})^n_j|^\alpha + |(\hat{u}_{\bar{x}-})^n_j|^\alpha} \qquad (\alpha > 0 \text{ and } |(\hat{u}_{\bar{x}-})^n_j| + |(\hat{u}_{\bar{x}+})^n_j| > 0) \tag{4.8}$$

$$w^n_j \overset{\text{def}}{=} \frac{1}{2}\left[1 + \sigma\,\xi(j, n; \alpha)\right] \tag{4.9}$$

$$(w'_-)^n_j \overset{\text{def}}{=} \begin{cases} w^n_j & \text{if } 0 \leq w^n_j \leq 1 \\ 1 & \text{if } w^n_j > 1 \\ 0 & \text{if } w^n_j < 0 \end{cases} \tag{4.10}$$

and

$$(w'_+)^n_j \overset{\text{def}}{=} 1 - (w'_-)^n_j \tag{4.11}$$

where $\sigma_o > 0$ and $\alpha > 0$ are preset parameters in the order of 1. Note that: (i) to avoid dividing by zero, in practice a small positive number such as $10^{-20}$ should be added to each of the denominators of the fractions that appear in Eqs. (4.7) and (4.8); and (ii) Eqs. (4.10) and (4.11) imply that

$$(w'_-)^n_j + (w'_+)^n_j = 1 \tag{4.12}$$

17

and

$$(w'_-)^n_j \geq 0 \quad \text{and} \quad (w'_+)^n_j \geq 0 \tag{4.13}$$

i.e., $(w'_-)^n_j$ and $(w'_+)^n_j$ represent a pair of interpolating weight factors. With the aid of these weight factors, Scheme w-3 is formed by Eq. (2.10) and

$$(u_{\bar{x}})^n_j = (w'_-)^n_j (\hat{u}_{\bar{x}-})^n_j + (w'_+)^n_j (\hat{u}_{\bar{x}+})^n_j \tag{4.14}$$

To explore Scheme w-3, note that an immediate result of Eqs. (4.7)–(4.9) is

$$w^n_j \begin{cases} > 1/2 & \text{if } |(\hat{u}_{\bar{x}+})^n_j| > |(\hat{u}_{\bar{x}-})^n_j| \\ = 1/2 & \text{if } |(\hat{u}_{\bar{x}+})^n_j| = |(\hat{u}_{\bar{x}-})^n_j| \\ < 1/2 & \text{if } |(\hat{u}_{\bar{x}+})^n_j| < |(\hat{u}_{\bar{x}-})^n_j| \end{cases} \qquad (\nu \neq 0) \tag{4.15}$$

In turn, by using Eq. (4.15), Eqs. (4.10) and (4.11) imply that

$$\begin{cases} 0 \leq (w'_+)^n_j < 1/2 < (w'_-)^n_j \leq 1 & \text{if } |(\hat{u}_{\bar{x}+})^n_j| > |(\hat{u}_{\bar{x}-})^n_j| \\ (w'_+)^n_j = (w'_-)^n_j = 1/2 & \text{if } |(\hat{u}_{\bar{x}+})^n_j| = |(\hat{u}_{\bar{x}-})^n_j| \\ 0 \leq (w'_-)^n_j < 1/2 < (w'_+)^n_j \leq 1 & \text{if } |(\hat{u}_{\bar{x}+})^n_j| < |(\hat{u}_{\bar{x}-})^n_j| \end{cases} \qquad (\nu \neq 0) \tag{4.16}$$

As such, of $(\hat{u}_{\bar{x}-})^n_j$ and $(\hat{u}_{\bar{x}+})^n_j$, the one with smaller absolute value is assigned an weight factor $> 1/2$ in Eq. (4.14). As explained at the end of Sec. 2.3, it follows that Scheme w-3 also has the capability to suppress wiggles or overshoots. In fact, as will be shown, it would retain this capability even when $|\nu| \ll 1$.

According to Eqs. (4.4) and (4.8), (i) $\xi(j, n; \alpha)$ implicitly is a function of $\nu$, and (ii) $\xi(j, n; \alpha) \to 0$ as $\nu \to 0$. Let $\xi(j, n; \alpha)$ varies linearly with $|\nu|$ in the immediate neighborhood of $\nu = 0$, i.e.,

$$\lim_{\nu \to 0^+} (\xi(j, n; \alpha)/|\nu|) = \xi^+_o(j, n; \alpha) \quad \text{and} \quad \lim_{\nu \to 0^-} (\xi(j, n; \alpha)/|\nu|) = \xi^-_o(j, n; \alpha) \tag{4.17}$$

where $\xi^+_o(j, n; \alpha)$ and $\xi^-_o(j, n; \alpha)$ are *parameters independent of $\nu$*. Then, by substituting Eq. (4.7) into Eq. (4.9) and using Eq. (4.17), one has

$$\lim_{\nu \to 0^+} w^n_j = \frac{1}{2} \left[ 1 + \sigma_o \xi^+_o(j, n; \alpha) \right] \quad \text{and} \quad \lim_{\nu \to 0^-} w^n_j = \frac{1}{2} \left[ 1 + \sigma_o \xi^-_o(j, n; \alpha) \right] \tag{4.18}$$

By assumption, $\sigma_o > 0$. Thus, one concludes that

$$\lim_{|\nu| \to 0^+} w^n_j \neq \frac{1}{2} \tag{4.19}$$

if

$$\xi^+_o(j, n; \alpha) \neq 0 \quad \text{and} \quad \xi^-_o(j, n; \alpha) \neq 0 \tag{4.20}$$

In turn, with the aid of Eq. (4.19), Eqs. (4.10) and (4.11) imply that

$$\lim_{|\nu| \to 0^+} (w'_+)^n_j \neq \frac{1}{2} \quad \text{and} \quad \lim_{|\nu| \to 0^+} (w'_-)^n_j \neq \frac{1}{2} \tag{4.21}$$

if Eq. (4.20) is assumed. In other words, *the expression on the right side of Eq. (4.14) generally would not reduce to the simple average of $(\hat{u}_{\bar{x}-})^n_j$ and $(\hat{u}_{\bar{x}+})^n_j$, and as such, Scheme w-3 would retain its capability to suppress wiggles and overshoots even when $|\nu| \ll 1$.* This conclusion was supported by the numerical results of Euler solvers [31] which were constructed using the weighted-average technique similar to that presented here.

### 4.4. Scheme w-4

18

Scheme w-4 is originally presented in [25]. As will be shown in [62], the weight factors used in this scheme can be constructed using the second class of weighted-averaging techniques reported in Sec. 4.3 of [31]. Let

$$(\eta_\pm)_j^n \stackrel{\text{def}}{=} \frac{|(\hat{u}_{\bar{x}\pm})_j^n|^\alpha}{\min\{|(\hat{u}_{\bar{x}-})_j^n|^\alpha, |(\hat{u}_{\bar{x}+})_j^n|^\alpha\}} - 1 \qquad (\alpha > 0 \text{ and } |(\hat{u}_{\bar{x}-})_j^n| + |(\hat{u}_{\bar{x}+})_j^n| > 0) \qquad (4.22)$$

$$(\tilde{w}_+)_j^n \stackrel{\text{def}}{=} \frac{1 + \sigma(\eta_-)_j^n}{2 + \sigma[(\eta_+)_j^n + (\eta_-)_j^n]} \qquad (\nu \neq 0) \qquad (4.23)$$

and

$$(\tilde{w}_-)_j^n \stackrel{\text{def}}{=} \frac{1 + \sigma(\eta_+)_j^n}{2 + \sigma[(\eta_+)_j^n + (\eta_-)_j^n]} \qquad (\nu \neq 0) \qquad (4.24)$$

where $\alpha > 0$ is a preset parameter in the order of 1 while $\sigma$ is that defined in Eq. (4.7). Note that: (i) to avoid dividing by zero, in practice a small positive number such as $10^{-20}$ should be added to the denominator of the fraction that appears in Eq. (4.22); and (ii) Eq. (4.22) implies that

$$(\eta_+)_j^n \begin{cases} > 0 & \text{if } |(\hat{u}_{\bar{x}+})_j^n| > |(\hat{u}_{\bar{x}-})_j^n| \\ = 0 & \text{if } |(\hat{u}_{\bar{x}+})_j^n| \leq |(\hat{u}_{\bar{x}-})_j^n| \end{cases} \qquad (4.25)$$

and

$$(\eta_-)_j^n \begin{cases} > 0 & \text{if } |(\hat{u}_{\bar{x}-})_j^n| > |(\hat{u}_{\bar{x}+})_j^n| \\ = 0 & \text{if } |(\hat{u}_{\bar{x}-})_j^n| \leq |(\hat{u}_{\bar{x}+})_j^n| \end{cases} \qquad (4.26)$$

Eqs. (4.23)–(4.26) imply that: (i)

$$(\tilde{w}_+)_j^n + (\tilde{w}_-)_j^n = 1 \qquad (4.27)$$

and

$$(\tilde{w}_+)_j^n > 0 \quad \text{and} \quad (\tilde{w}_-)_j^n > 0 \qquad (4.28)$$

i.e., $(\tilde{w}_+)_j^n$ and $(\tilde{w}_-)_j^n$ represent a pair of interpolating weight factors; and (ii)

$$\begin{cases} 0 < (\tilde{w}_+)_j^n < 1/2 < (\tilde{w}_-)_j^n < 1 & \text{if } |(\hat{u}_{\bar{x}+})_j^n| > |(\hat{u}_{\bar{x}-})_j^n| \\ (\tilde{w}_+)_j^n = (\tilde{w}_-)_j^n = 1/2 & \text{if } |(\hat{u}_{\bar{x}+})_j^n| = |(\hat{u}_{\bar{x}-})_j^n| \\ 0 < (\tilde{w}_-)_j^n < 1/2 < (\tilde{w}_+)_j^n < 1 & \text{if } |(\hat{u}_{\bar{x}+})_j^n| < |(\hat{u}_{\bar{x}-})_j^n| \end{cases} \qquad (\nu \neq 0) \qquad (4.29)$$

Scheme w-4 is formed by Eq. (2.10) and

$$(u_{\bar{x}})_j^n = (\tilde{w}_-)_j^n (\hat{u}_{\bar{x}-})_j^n + (\tilde{w}_+)_j^n (\hat{u}_{\bar{x}+})_j^n \qquad (4.30)$$

According to Eq. (4.29), of $(\hat{u}_{\bar{x}-})_j^n$ and $(\hat{u}_{\bar{x}+})_j^n$ in Eq. (4.30), the one with the smaller absolute value is also assigned an weight factor $> 1/2$. As such, Scheme w-4 also has the capability to suppress wiggles or overshoots. In fact, if $(\eta_\pm)_j^n$ varies linearly with $|\nu|$ in the immediate neighborhood of $\nu = 0$, one would conclude from the argument that lead to Eq. (4.21) that in general

$$\lim_{|\nu| \to 0^+} (\tilde{w}_+)_j^n \neq \frac{1}{2} \quad \text{and} \quad \lim_{|\nu| \to 0^+} (\tilde{w}_-)_j^n \neq \frac{1}{2} \qquad (4.31)$$

As such *the expression on the right side of Eq. (4.30) generally would not reduce to the simple average of $(\hat{u}_{\bar{x}-})_j^n$ and $(\hat{u}_{\bar{x}+})_j^n$, i.e., Scheme w-4 would also retain its capability to suppress wiggles and overshoots even when $|\nu| \ll 1$.* This conclusion was supported by the numerical results of Euler solvers [31] which were constructed using the weighted-average technique similar to that presented here.

A rather thorough review of several 2nd-order 1D CESE schemes has been provided in Sec. 2–4. In the following, it will be shown that each of these second-order scheme can be "naturally" extended to become its

4th-order version. In Sec. 5, first we will construct the $a$-4 and $a$-$\epsilon$-4 schemes, i.e., the 4th-order extensions of the $a$ and $a$-$\epsilon$ schemes, respectively.

## 5. The $a$-4 and $a$-$\epsilon$-4 schemes

To construct the $a$-4 and $a$-$\epsilon$-4 schemes, again we consider Eq. (2.1). Also we will use (i) the same mesh structure and the same CEs and SEs defined in Sec. 2 and (ii) the same definitions of $\vec{h}^*(x,t\,;j,n)$, $\nu$, and $(u_{\bar{x}})_j^n$ defined in Eqs. (2.4) and (2.9). However, Eq. (2.3) is now replaced by

$$
\begin{aligned}
u^*(x,t\,;j,n) \stackrel{\text{def}}{=} &\; u_j^n + (u_x)_j^n(x - x_j) + (u_t)_j^n(t - t^n) + \frac{1}{2}(u_{xx})_j^n(x - x_j)^2 \\
&+ (u_{xt})_j^n(x - x_j)(t - t^n) + \frac{1}{2}(u_{tt})_j^n(t - t^n)^2 + \frac{1}{6}(u_{xxx})_j^n(x - x_j)^3 \\
&+ \frac{1}{2}(u_{xxt})_j^n(x - x_j)^2(t - t^n) + \frac{1}{2}(u_{xtt})_j^n(x - x_j)(t - t^n)^2 + \frac{1}{6}(u_{ttt})_j^n(t - t^n)^3
\end{aligned}
\tag{5.1}
$$

Here $u_j^n$, $(u_x)_j^n$, $(u_t)_j^n$, $(u_{xx})_j^n$, $(u_{xt})_j^n$, $(u_{tt})_j^n$ $(u_{xxx})_j^n$, $(u_{xxt})_j^n$, $(u_{xtt})_j^n$, and $(u_{ttt})_j^n$ are constants in SE$(j,n)$. They can be considered as the numerical analogues of the values of $u$, $\partial u/\partial x$, $\partial u/\partial t$, $\partial^2 u/\partial x^2$, $\partial^2 u/\partial x \partial t$, $\partial^2 u/\partial t^2$, $\partial^3 u/\partial x^3$, $\partial^3 u/\partial x^2 \partial t$, $\partial^3 u/\partial x \partial t^2$, and $\partial^3 u/\partial t^3$ at the mesh point $(j,n)$, respectively. As such $u^*(x,t\,;j,n)$ represents a 3rd-order Taylor's approximation of $u$.

For any $(j,n) \in \Omega$, let $u = u^*(x,t\,;j,n)$ be a solution to Eq. (2.1) in SE$(j,n)$, i.e.,

$$
\frac{\partial u^*(x,t\,;j,n)}{\partial t} + a \frac{\partial u^*(x,t\,;j,n)}{\partial x} \equiv 0 \qquad (x,t) \in \text{SE}(j,n)
\tag{5.2}
$$

Then one has

$$
\begin{array}{ll}
(u_t)_j^n = -a(u_x)_j^n, \quad (u_{xt})_j^n = -a(u_{xx})_j^n, \quad (u_{tt})_j^n = a^2(u_{xx})_j^n, & \\
(u_{xxt})_j^n = -a(u_{xxx})_j^n, \quad (u_{xtt})_j^n = a^2(u_{xxx})_j^n, \quad (u_{ttt})_j^n = -a^3(u_{xxx})_j^n & 
\end{array}
\qquad (j,n) \in \Omega
\tag{5.3}
$$

Substituting Eq. (5.3) into Eq. (5.1), one obtains

$$
\begin{aligned}
u^*(x,t\,;j,n) = &\; u_j^n + (u_x)_j^n \left[(x - x_j) - a(t - t^n)\right] \\
&+ \frac{1}{2}(u_{xx})_j^n \left[(x - x_j) - a(t - t^n)\right]^2 + \frac{1}{6}(u_{xxx})_j^n \left[(x - x_j) - a(t - t^n)\right]^3
\end{aligned}
\tag{5.4}
$$

Thus, (i) $u^*(x,t\,;j,n)$ is a 3rd-order polynomial function of $\left[(x - x_j) - a(t - t^n)\right]$; and (ii) $u_j^n$, $(u_x)_j^n$, $(u_{xx})_j^n$, and $(u_{xxx})_j^n$ are the only independent mesh variables associated with $(j,n)$. Note that, for the same reason presented immediately following Eq. (2.6), *except for flux evaluation, Eq. (5.4) and Eq. (5.7) (see below) are applicable even for* $(x,t) \notin SE(j,n)$.

Let

$$
u_{xx}^*(x,t;j,n) \stackrel{\text{def}}{=} \frac{\partial^2 u^*(x,t\,;j,n)}{\partial x^2}
\tag{5.5}
$$

Then Eqs. (5.2) and (5.4), respectively, imply that

$$
\frac{\partial u_{xx}^*(x,t;j,n)}{\partial t} + a \frac{\partial u_{xx}^*(x,t;j,n)}{\partial x} \equiv 0 \qquad (x,t) \in \text{SE}(j,n)
\tag{5.6}
$$

and

$$
u_{xx}^*(x,t;j,n) = (u_{xx})_j^n + (u_{xxx})_j^n \left[(x - x_j) - a(t - t^n)\right] \qquad (x,t) \in \text{SE}(j,n)
\tag{5.7}
$$

i.e., (i) $u = u_{xx}^*(x,t;j,n)$ is also a solution to Eq. (2.1) in SE$(j,n)$; and (ii) $u_{xx}^*(x,t;j,n)$ is a first-order polynomial function of $\left[(x - x_j) - a(t - t^n)\right]$. Note that the fact that Eq. (5.2) implies Eq. (5.6) is but the special case of the more general result that Eq. (2.1) implies

$$
\frac{\partial}{\partial t}\left(\frac{\partial^2 u}{\partial x^2}\right) + a \frac{\partial}{\partial x}\left(\frac{\partial^2 u}{\partial x^2}\right) = 0
\tag{5.8}
$$

20

In fact, for a smooth function $u(x,t)$, the conservation condition Eq. (2.2) implies the conservation condition

$$\oint_{S(V)} \vec{h}_{xx} \cdot d\vec{s} = 0 \tag{5.9}$$

where

$$\vec{h}_{xx} \stackrel{\text{def}}{=} \frac{\partial^2 \vec{h}}{\partial x^2} \equiv \left( a \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial x^2} \right) \tag{5.10}$$

As a result of the above observations on the "equivalence" relation between $u$ and $\partial^2 u/\partial x^2$, the reader should no longer be surprised by the fact that, by replacing the symbols $u_j^n$, $(u_x)_j^n$, $u^*(x,t;j,n)$, and $\vec{h}^*(x,t;j,n)$ with $(u_{xx})_j^n$, $(u_{xxx})_j^n$, $u_{xx}^*(x,t;j,n)$, and $\vec{h}_{xx}^*(x,t;j,n)$, respectively, Eqs. (2.4) and (2.6) would turn into

$$\vec{h}_{xx}^*(x,t;j,n) \stackrel{\text{def}}{=} (au_{xx}^*(x,t;j,n), u_{xx}^*(x,t;j,n)) \qquad (x,t) \in \text{ SE}(j,n) \tag{5.11}$$

(i.e., the numerical analogue of Eq. (5.10)) and Eq. (5.7), respectively. *This structural similarity strongly suggests that the mesh variables $(u_{xx})_j^n$ and $(u_{xxx})_j^n$ can be advanced in time by using one of the 2nd-order CESE schemes described earlier with the understanding that the role played by $u_j^n$, $(u_x)_j^n$, $u^*(x,t;j,n)$, and $\vec{h}^*(x,t;j,n)$ in a 2nd-order CESE scheme be replaced by $(u_{xx})_j^n$, $(u_{xxx})_j^n$, $u_{xx}^*(x,t;j,n)$, and $\vec{h}_{xx}^*(x,t;j,n)$, respectively. Note that this simplified approach is justified by the observation that $(u_{xx})_j^n$ and $(u_{xxx})_j^n$, respectively, appear in the 2nd-order and 3rd-order terms on the right side of Eq. (5.4) and therefore the 4th-order accuracy requirement of $u^*(x,t;j,n)$ demands only that $(u_{xx})_j^n$ and $(u_{xxx})_j^n$, respectively, be 2nd-order and 1st-order in accuracy, i.e., the same orders of accuracy achieved by $u_j^n$ and $(u_x)_j^n$ in a 2nd-order CESE scheme.*

For the reason presented above, equations such as Eqs. (5.11) and (5.7) hereafter will be designated as the substitution forms of Eqs. (2.4) and (2.6), respectively. With the aid of this designation, in the following $(u_{xx})_j^n$ and $(u_{xxx})_j^n$ will be advanced in time using the $a$ and $a$-$\epsilon$ schemes.

## 5.1. Time marching of $(u_{xx})_j^n$ and $(u_{xxx})_j^n$

By imposing the substitution forms of Eqs. (2.7) and (2.8), i.e.,

$$\oint_{S(\text{CE}_+(j,n))} \vec{h}_{xx}^* \cdot d\vec{s} = 0 \qquad (j,n) \in \Omega \tag{5.12}$$

and

$$\oint_{S(\text{CE}_-(j,n))} \vec{h}_{xx}^* \cdot d\vec{s} = 0 \qquad (j,n) \in \Omega \tag{5.13}$$

and using Eqs. (5.7) and (5.11), one can show that (i)

$$(u_{\bar{x}\bar{x}})_j^n = \frac{1}{2} \left\{ (1-\nu) \left[ u_{\bar{x}\bar{x}} - (1+\nu)u_{\bar{x}\bar{x}\bar{x}} \right]_{j+1/2}^{n-1/2} + (1+\nu) \left[ u_{\bar{x}\bar{x}} + (1-\nu)u_{\bar{x}\bar{x}\bar{x}} \right]_{j-1/2}^{n-1/2} \right\} \qquad (j,n) \in \Omega \tag{5.14}$$

and, assuming $|\nu| \neq 1$, (ii)

$$(u_{\bar{x}\bar{x}\bar{x}})_j^n = (u_{\bar{x}\bar{x}\bar{x}}^a)_j^n \qquad (j,n) \in \Omega \tag{5.15}$$

with

$$(u_{\bar{x}\bar{x}\bar{x}}^a)_j^n \stackrel{\text{def}}{=} \frac{1}{2} \left\{ \left[ u_{\bar{x}\bar{x}} - (1+\nu)u_{\bar{x}\bar{x}\bar{x}} \right]_{j+1/2}^{n-1/2} - \left[ u_{\bar{x}\bar{x}} + (1-\nu)u_{\bar{x}\bar{x}\bar{x}} \right]_{j-1/2}^{n-1/2} \right\} \tag{5.16}$$

where $\nu$, $(u_{\bar{x}})_j^n$, $(u_{\bar{x}\bar{x}})_j^n$, and $(u_{\bar{x}\bar{x}\bar{x}})_j^n$ are defined by Eq. (2.9) and

$$(u_{\bar{x}\bar{x}})_j^n \stackrel{\text{def}}{=} \left( \frac{\Delta x}{4} \right)^2 (u_{xx})_j^n \quad \text{and} \quad (u_{\bar{x}\bar{x}\bar{x}})_j^n \stackrel{\text{def}}{=} \left( \frac{\Delta x}{4} \right)^3 (u_{xxx})_j^n \qquad (j,n) \in \Omega \tag{5.17}$$

21

It can be shown that Eqs. (5.14)–(5.15) are equivalent to the substitution form of Eqs. (2.10) and (2.11). As such, the time marching scheme formed by Eqs. (5.14) and (5.15) literally is the same $a$ scheme defined earlier. Also, one can shown easily that *Eq. (5.14) is equivalent to the substitution form of Eq. (2.13), i.e.,*

$$\oint_{S(CE_{(j,n)})} \vec{h}^*_{xx} \cdot d\vec{s} = 0 \qquad\qquad (j,n) \in \Omega \qquad\qquad (5.18)$$

Note that, as explained in Comment (d) given at the end of Sec. 2.1, Eq. (5.18) is a direct result of Eqs. (5.12) and (5.13).

Next let

$$u^*_{\bar{x}\bar{x}}(x,t;j,n) \stackrel{\text{def}}{=} \left(\frac{\Delta x}{4}\right)^2 u^*_{xx}(x,t;j,n) \quad \text{and} \quad (u_{\bar{x}\bar{x}t})(j,n) \stackrel{\text{def}}{=} \left(\frac{\Delta x}{4}\right)^2 (u_{xxt})^n_j \qquad (j,n) \in \Omega \qquad (5.19)$$

Then the substitution forms of Eqs. (2.14)–(2.18) can be expressed as

$$(u'_{\bar{x}\bar{x}})^n_{j\pm 1/2} \stackrel{\text{def}}{=} u^*_{\bar{x}\bar{x}}\left(x_{j\pm 1/2}, t^n; j\pm 1/2, n-1/2\right) = [u_{\bar{x}\bar{x}} + (\Delta t/2)u_{\bar{x}\bar{x}t}]^{n-1/2}_{j\pm 1/2} \qquad (5.20)$$

$$(u'_{\bar{x}\bar{x}})^n_{j\pm 1/2} = (u_{\bar{x}\bar{x}} - 2\nu\, u_{\bar{x}\bar{x}\bar{x}})^{n-1/2}_{j\pm 1/2} \qquad (5.21)$$

$$(u^c_{\bar{x}\bar{x}\bar{x}})^n_j \stackrel{\text{def}}{=} \frac{\Delta x}{4}\left(\frac{(u'_{\bar{x}\bar{x}})^n_{j+1/2} - (u'_{\bar{x}\bar{x}})^n_{j-1/2}}{\Delta x}\right) \equiv \frac{(u'_{\bar{x}\bar{x}})^n_{j+1/2} - (u'_{\bar{x}\bar{x}})^n_{j-1/2}}{4} \qquad (5.22)$$

$$(u^c_{\bar{x}\bar{x}\bar{x}})^n_j = \frac{1}{4}\left[(u_{\bar{x}\bar{x}} - 2\nu u_{\bar{x}\bar{x}\bar{x}})^{n-1/2}_{j+1/2} - (u_{\bar{x}\bar{x}} - 2\nu u_{\bar{x}\bar{x}\bar{x}})^{n-1/2}_{j-1/2}\right] \qquad (5.23)$$

and

$$(u^c_{\bar{x}\bar{x}\bar{x}} - u^a_{\bar{x}\bar{x}\bar{x}})^n_j = \frac{1}{4}\left[(u_{\bar{x}\bar{x}} + 2u_{\bar{x}\bar{x}\bar{x}})^{n-1/2}_{j-1/2} - (u_{\bar{x}\bar{x}} - 2u_{\bar{x}\bar{x}\bar{x}})^{n-1/2}_{j+1/2}\right] \qquad (5.24)$$

respectively. Note that the validity of the last equality sign in Eq. (5.20) can be proved using Eqs. (5.7), (5.17) and (5.19) along with the relation $(u_{xxt})^n_j = -a(u_{xxx})^n_j$ which is part of Eq. (5.3).

Moreover, let the symbol $\epsilon$ be replaced by $\epsilon_1$. Then the substitution form of Eq. (2.19) can be expressed as

$$(u_{\bar{x}\bar{x}\bar{x}})^n_j = (u^a_{\bar{x}\bar{x}\bar{x}})^n_j + 2\epsilon_1(u^c_{\bar{x}\bar{x}\bar{x}} - u^a_{\bar{x}\bar{x}\bar{x}})^n_j \qquad (5.25)$$

Obviously Eq. (5.25) reduces to (i) $(u_{\bar{x}\bar{x}\bar{x}})^n_j = (u^a_{\bar{x}\bar{x}\bar{x}})^n_j$ if $\epsilon_1 = 0$; and (ii) $(u_{\bar{x}\bar{x}\bar{x}})^n_j = (u^c_{\bar{x}\bar{x}\bar{x}})^n_j$ if $\epsilon_1 = 1/2$. With the aid of Eqs. (5.16) and (5.24), Eq. (5.25) can be rewritten as

$$(u_{\bar{x}\bar{x}\bar{x}})^n_j = \frac{1}{2}\left\{[(1-\epsilon_1)u_{\bar{x}\bar{x}} + (2\epsilon_1 - 1 - \nu)u_{\bar{x}\bar{x}\bar{x}}]^{n-1/2}_{j+1/2} - [(1-\epsilon_1)u_{\bar{x}\bar{x}} + (1 - \nu - 2\epsilon_1)u_{\bar{x}\bar{x}\bar{x}}]^{n-1/2}_{j-1/2}\right\} \qquad (5.26)$$

Obviously Eq. (5.26) is the substitution form of Eq. (2.19a) if the symbol $\epsilon$ in the latter equation is replaced by $\epsilon_1$.

Because (i) $(u_{\bar{x}\bar{x}})^n_j$ and $(u_{\bar{x}\bar{x}\bar{x}})^n_j$ can be explicitly evaluated in terms of the mesh variables at the $(n-1/2)$th time level using Eqs. (5.14) and (5.26), and (ii) the latter equations are equivalent to the substitution forms of Eqs. (2.10) and (2.19a), it has been shown that indeed $(u_{\bar{x}\bar{x}})^n_j$ and $(u_{\bar{x}\bar{x}\bar{x}})^n_j$ can be advanced in time using the $a$-$\epsilon$ scheme.

This completes the description of the first step in the construction of the $a$-$\epsilon$-4 scheme. In the following, we describe the second step, i.e., the details of how $u^n_j$ and $(u_{\bar{x}})^n_j$ can be advanced in time. As a preliminary, we conclude this subsection with the following remarks:

(a) As a result of its construction, the scheme formed by Eqs. (5.14) and (5.26) observes both the conservation relations Eqs. (5.12) and (5.13) if and only if $\epsilon_1 = 0$.

22

(b) Because Eq. (5.14) is equivalent to Eq. (5.18), the scheme formed by Eq. (5.14) and (5.26) still observes the less stringent conservation relation Eq. (5.18) if $\epsilon_1 \neq 0$.

(c) Because of how they are derived, by replacing the symbols $\epsilon_1$, $u_{\bar{x}\bar{x}}$, and $u_{\bar{x}\bar{x}\bar{x}}$ with $\epsilon$, $u$, and $u_{\bar{x}}$, respectively, Eqs. (5.14) and (5.26), respectively, will become Eq. (2.10) and (2.19a), the defining equations of the $a$-$\epsilon$ scheme.

## 5.2. Time marching of $u_j^n$ and $(u_x)_j^n$

Let

$$(s_+)_j^n \stackrel{\text{def}}{=} \left[ u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}} - \frac{(1+\nu)(1+\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_j^n \qquad (j,n) \in \Omega \qquad (5.27)$$

$$(s_-)_j^n \stackrel{\text{def}}{=} \left[ u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}} + \frac{(1-\nu)(1+\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_j^n \qquad (j,n) \in \Omega \qquad (5.28)$$

Then, assuming Eqs. (2.4), (2.9), (5.4), and (5.17), it can be shown that the conservation relations Eqs. (2.7) and (2.8) are equivalent to

$$(1-\nu)\left\{ \left[ u + (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}} + \frac{(1+\nu)(1+\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_j^n - (s_+)_{j+1/2}^{n-1/2} \right\} = 0 \qquad (5.29)$$

and

$$(1+\nu)\left\{ \left[ u - (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}} - \frac{(1-\nu)(1+\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_j^n - (s_-)_{j-1/2}^{n-1/2} \right\} = 0 \qquad (5.30)$$

respectively.

By adding Eqs. (5.29) and (5.30) together, one has

$$u_j^n = \frac{1}{2}\left[ (1-\nu)(s_+)_{j+1/2}^{n-1/2} + (1+\nu)(s_-)_{j-1/2}^{n-1/2} \right] - \frac{2}{3}(u_{\bar{x}\bar{x}})_j^n \qquad (j,n) \in \Omega \qquad (5.31)$$

Let $1 - \nu^2 \neq 0$, i.e., $1 - \nu \neq 0$ and $1 + \nu \neq 0$. Then Eqs. (5.29) and (5.30) can be divided by $(1-\nu)$ and $(1+\nu)$, respectively. By subtracting the resulting equations from each other, one has

$$(u_{\bar{x}})_j^n = (u_{\bar{x}}^{a\text{-}4})_j^n \qquad (j,n) \in \Omega \qquad (5.32)$$

where

$$(u_{\bar{x}}^{a\text{-}4})_j^n \stackrel{\text{def}}{=} \frac{1}{2}\left[ (s_+)_{j+1/2}^{n-1/2} - (s_-)_{j-1/2}^{n-1/2} \right] - \frac{2\nu}{3}(u_{xx})_j^n - \frac{1}{3}\left(1+\nu^2\right)(u_{\bar{x}\bar{x}\bar{x}})_j^n \qquad (j,n) \in \Omega \qquad (5.33)$$

Hereafter the superscript "4" is used to distinguish the parameters introduced in the current construction of the 4th-order schemes from their counterparts that appear in the 2nd-order schemes.

At this juncture, note that:

(a) The zero-order and first-order terms in the third-order expansion on the right side of Eq. (5.4) are identical to those in the first-order expansion on the right side of Eq. (2.6), and (ii) Eqs. (5.31) and (5.32) were obtained by imposing the conservation relations Eqs. (2.7) and (2.8) and assuming Eqs. (2.4) and (5.4) while Eqs. (2.10) and (2.11) were obtained by imposing the same conservation relations and assuming Eqs. (2.4) and (2.6). As such, one would expect that Eqs. (5.31) and (5.32) differ from Eqs. (2.10) and (2.11) only in the appearance of the extra 2nd-order and 3rd-order terms in the former equations. With the aid of Eqs. (2.12), (5.27), (5.28), and (5.33), the reader can prove for himself that this is indeed true.

23

(b) Assuming Eqs. (2.4), (2.9), (5.4), (5.17), (5.27), and (5.28), it has been shown that Eqs. (2.7) and (2.8) imply (i) Eq. (5.31) for all $\nu$; and (ii) Eq. (5.32) if $|\nu| \neq 1$. On the other hand, it can be shown that Eqs. (5.31) and (5.32) are equivalent to

$$\left[ u + (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}} + \frac{(1+\nu)(1+\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_j^n - (s_+)_{j+1/2}^{n-1/2} = 0 \qquad (5.29a)$$

and

$$\left[ u - (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}} - \frac{(1-\nu)(1+\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_j^n - (s_-)_{j-1/2}^{n-1/2} = 0 \qquad (5.30a)$$

for all $\nu$. Because (i) Eqs. (5.29a) and (5.30a) imply Eqs. (5.29) and (5.30), respectively; and (ii) Eqs. (5.29) and (5.30) are equivalent to Eqs. (2.7) and (2.8), respectively, one concludes that, for all $\nu$, the last two conservation conditions are enforced by Eqs. (5.31) and (5.32).

(c) Under the current assumptions, it can be shown that Eq. (5.31) is equivalent to the conservation relation Eq. (2.13).

For Eqs. (5.31) and (5.32) to be considered as a complete marching scheme for $u_j^n$ and $(u_{\bar{x}})_j^n$, the mesh variables $(u_{\bar{x}\bar{x}})_j^n$ and $(u_{\bar{x}\bar{x}\bar{x}})_j^n$ which appear on right sides of Eqs. (5.31) and (5.33) must be expressed explicitly in terms of the mesh variables at the $(n-1/2)$th time level. In the current construction, this is done by demanding that the former mesh variables be evaluated in terms of the latter mesh variables using Eqs. (5.14) and (5.26). By using this assumption and Eqs. (5.27) and (5.28), Eqs. (5.31)–(5.33) can now be cast into the forms by which $u_j^n$ and $(u_{\bar{x}})_j^n$ can be evaluated in terms of the mesh variables at the $(n-1/2)$th time level, i.e.,

$$
\begin{aligned}
u_j^n = {}& \frac{1-\nu}{2}\left[ u - (1+\nu)u_{\bar{x}} + \frac{2\nu(1+\nu)}{3}u_{\bar{x}\bar{x}} + \frac{(1+\nu)(1-\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_{j+1/2}^{n-1/2} \\
&+ \frac{1+\nu}{2}\left[ u + (1-\nu)u_{\bar{x}} - \frac{2\nu(1-\nu)}{3}u_{\bar{x}\bar{x}} - \frac{(1-\nu)(1-\nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_{j-1/2}^{n-1/2}
\end{aligned}
\qquad (j,n)\in\Omega \qquad (5.34)
$$

and

$$(u_{\bar{x}})_j^n = \left[ u_{\bar{x}}^{a-4}(\epsilon_1) \right]_j^n \qquad (j,n)\in\Omega \qquad (5.35)$$

where

$$
\begin{aligned}
\left[ u_{\bar{x}}^{a-4}(\epsilon_1) \right]_j^n \overset{\text{def}}{=} {}& \frac{1}{2}\left\{ u - (1+\nu)u_{\bar{x}} + \frac{1+3\nu^2+\epsilon_1(1+\nu^2)}{3}u_{\bar{x}\bar{x}} + \frac{2\left[\nu(1-\nu^2) - \epsilon_1(1+\nu^2)\right]}{3}u_{\bar{x}\bar{x}\bar{x}} \right\}_{j+1/2}^{n-1/2} \\
&- \frac{1}{2}\left\{ u + (1-\nu)u_{\bar{x}} + \frac{1+3\nu^2+\epsilon_1(1+\nu^2)}{3}u_{\bar{x}\bar{x}} + \frac{2\left[\nu(1-\nu^2) + \epsilon_1(1+\nu^2)\right]}{3}u_{\bar{x}\bar{x}\bar{x}} \right\}_{j-1/2}^{n-1/2} \qquad (j,n)\in\Omega
\end{aligned}
$$

$$(5.36)$$

At this juncture, note that:

(a) As a result of how Eqs. (5.34)–(5.36) are derived, one concludes that: (i) Eqs. (5.14) and (5.34) satisfy Eq. (5.31) automatically; and (ii) Eqs. (5.14), (5.26), and (5.35) satisfy Eqs. (5.32) automatically. Because the conservation relations Eq. (2.7) and (2.8), in turn, are enforced by Eqs. (5.31) and (5.32) for all $\nu$ (see Comment (b) given following Eq. (5.33)), the two conservation relations are observed by the scheme formed by Eqs. (5.14), (5.26), (5.34), and (5.35) for all $\nu$ and $\epsilon_1$.

(b) It is pointed out earlier (see comments (a) and (b) given at the end of Sec. 5.1) that the scheme formed by Eqs. (5.14) and (5.26) observes (i) the conservation relations Eqs. (5.12) and (5.13) if $\epsilon_1 = 0$; and (ii) only the conservation relation Eq. (5.18) if $\epsilon_1 \neq 0$.

(c) As a result of the comments made in the above items (a) and (b), one concludes that *the special scheme formed by Eqs. (5.14), (5.26), (5.34), and (5.35) with $\epsilon_1 = 0$ observes the four conservation relations Eqs. (2.7), (2.8), (5.12), and (5.13). In fact, for this special scheme, the time evolution of the four independent mesh variables $u_j^n$, $(u_{\bar{x}})_j^n$, $(u_{\bar{x}\bar{x}})_j^n$, and $(u_{\bar{x}\bar{x}\bar{x}})_j^n$ is completely determined by these four conservation relations. As such, this special scheme is referred to as the a-4 scheme.*

(d) On the other hand, for the case $\epsilon_1 \neq 0$, the scheme defined in the above item (a) satisfies only three conservation relations, i.e., Eqs. (2.7), (2.8), and (5.18). In the following, this scheme will be extended so that the extension satisfies only the two least stringent conservation relations Eq. (2.13) and (2.18).

To proceed, first we will construct the current version of $(u_{\bar{x}}^c)_j^n$ which was defined in Eq. (2.16). As a preliminary, Taylor's expansion is invoked to obtain

$$
u\big(x_j \pm [(1+\tau)\Delta x/4], t^n\big) = \Big\{ 1 \pm [(1+\tau)\Delta x/4]\partial/\partial x + (1/2)[(1+\tau)\Delta x/4]^2\partial^2/\partial x^2
$$
$$
\pm (1/6)[(1+\tau)\Delta x/4]^3\partial^3/\partial x^3 + (1/24)[(1+\tau)\Delta x/4]^4\partial^4/\partial x^4 \Big\} u(x_j, t^n) + O\big[(\Delta x)^5\big]
$$
$$(\tau > -1) \qquad (5.37)$$

Here (i) $\tau > -1$ is an adjustable parameter, and (ii) $(x_j + [(1+\tau)\Delta x/4], t^n)$ and $(x_j - [(1+\tau)\Delta x/4], t^n)$, are the coordinates of points $P^+$ and $P^-$ depicted in Fig. 3, respectively. By subtracting the two equations in Eq. (5.37) from each other and then rearranging the resulting equation, one has

$$
\frac{\partial u}{\partial x}(x_j, t^n) = \frac{u\,(x_j + [(1+\tau)\Delta x/4], t^n) - u\,(x_j - [(1+\tau)\Delta x/4], t^n)}{(1+\tau)\Delta x/2} - \frac{[(1+\tau)\Delta x]^2}{96}\frac{\partial^3 u}{\partial x^3}(x_j, t^n) + O\big[(\Delta x)^4\big]
$$
$$(5.38)$$

Next, by using Eqs. (3.6) and (3.7), one has

$$
u(x_j \pm [(1+\tau)\Delta x/4], t^n) = u(x_{j\pm1/2} \mp [(1-\tau)\Delta x/4], t^{n-1/2} + \Delta t/2) \qquad (5.39)
$$

The expression on the right side of Eq. (5.39) can be expressed as a Taylor's expansion about the point $(x_{j\pm1/2}, t^{n-1/2})$. Substituting the expansion into Eq. (5.39), one has

$$
u\,(x_j \pm [(1+\tau)\Delta x/4], t^n) = u^4(P^\pm) + \text{4th-order terms involving } \Delta t \text{ and } \Delta x \qquad (5.40)
$$

where

$$
u^4(P^\pm) \overset{\text{def}}{=} \Big\{ 1 + \big\{(\Delta t/2)\partial/\partial t \mp [(1-\tau)\Delta x/4]\partial/\partial x\big\} + (1/2)\big\{(\Delta t/2)\partial/\partial t \mp [(1-\tau)\Delta x/4]\partial/\partial x\big\}^2
$$
$$
+ (1/6)\big\{(\Delta t/2)\partial/\partial t \mp [(1-\tau)\Delta x/4]\partial/\partial x\big\}^3 \Big\} u(x_{j\pm1/2}, t^{n-1/2})
$$
$$(5.41)$$

Hereafter, we consider only a mesh refinement procedure in which the ratio $\Delta t/\Delta x$ and the parameter $\tau$ are held constant as $\Delta t \to 0$ and $\Delta x \to 0$. Because the advection speed $a$ is a constant, in such a mesh refinement procedure, $\nu$ is also held constant as $\Delta t \to 0$ and $\Delta x \to 0$. Moreover, with this assumption, the 4th-order terms which appear on the right side of Eq. (5.40) can be expressed as $O\big[(\Delta x)^4\big]$, i.e., Eq. (5.40) can be expressed as

$$
u\,(x_j \pm [(1+\tau)\Delta x/4], t^n) = u^4(P^\pm) + O\big[(\Delta x)^4\big] \qquad (5.42)
$$

In turn, by substituting Eqs. (5.42) into Eq. (5.38), one has

$$
\frac{\partial u}{\partial x}(x_j, t^n) = \frac{u^4(P^+) - u^4(P^-)}{(1+\tau)\Delta x/2} - \frac{[(1+\tau)\Delta x]^2}{96}\frac{\partial^3 u}{\partial x^3}(x_j, t^n) + O\big[(\Delta x)^3\big] \qquad (5.43)
$$

25

For any $(j, n) \in \Omega$, $u^4(P^+)$ and $u^4(P^-)$ will be approximated by

$$u^{4*}(P^+) \stackrel{\text{def}}{=} u^*(x_j + [(1+\tau)\Delta x/4], t^n; j+1/2, n-1/2)$$
$$\equiv \left[ u - (1 - \tau + 2\nu)u_{\bar{x}} + (1/2)(1 - \tau + 2\nu)^2\, u_{\bar{x}\bar{x}} - (1/6)(1 - \tau + 2\nu)^3\, u_{\bar{x}\bar{x}\bar{x}} \right]_{j+1/2}^{n-1/2} \tag{5.44}$$

and

$$u^{4*}(P^-) \stackrel{\text{def}}{=} u^*(x_j - [(1+\tau)\Delta x/4], t^n; j-1/2, n-1/2)$$
$$\equiv \left[ u + (1 - \tau - 2\nu)u_{\bar{x}} + (1/2)(1 - \tau - 2\nu)^2\, u_{\bar{x}\bar{x}} + (1/6)(1 - \tau - 2\nu)^3\, u_{\bar{x}\bar{x}\bar{x}} \right]_{j-1/2}^{n-1/2} \tag{5.45}$$

respectively. According to Eqs. (3.6), (3.7), and (5.1)–(5.4), $u^{4*}(P^+)$ is the 3rd-order Taylor's approximation (i.e., the error is $O\left[(\Delta x)^4\right]$) of $u$ at $P^+$ evaluated using the mesh variables at $(j+1/2, n-1/2)$ while $u^{4*}(P^-)$ is that at $P^-$ evaluated using the mesh variables at $(j-1/2, n-1/2)$. Note that the validity of the last equality signs on the right sides of Eqs. (5.44) and (5.45) follows from Eqs. (2.9), (3.6), (3.7), (5.4) and (5.17). Moreover, according to Eq. (5.43), the parameter

$$\left[ u_x^4(\tau) \right]_j^n \stackrel{\text{def}}{=} \frac{u^{4*}(P^+) - u^{4*}(P^-)}{(1+\tau)\Delta x/2} - \frac{[(1+\tau)\Delta x]^2}{96}(u_{xxx})_j^n \tag{5.46}$$

represents a numerical analogue of the value of $\partial u/\partial x$ at $(j, n)$ with 3rd-order accuracy. This is consistent with the 4th-order accuracy requirement of $u^*(x, t; j, n)$ because $(u_x)_j^n$ (the numerical analogue of $\partial u/\partial x$) appears in the first-order term on the right side of Eq. (5.4). By using Eq. (5.17) and

$$\left[ u_{\bar{x}}^4(\tau) \right]_j^n \stackrel{\text{def}}{=} \frac{\Delta x}{4} \left[ u_x^4(\tau) \right]_j^n \tag{5.47}$$

Eq. (5.46) can be cast into its normalized form, i.e.,

$$\left[ u_{\bar{x}}^4(\tau) \right]_j^n \stackrel{\text{def}}{=} \frac{u^{4*}(P^+) - u^{4*}(P^-)}{2(1+\tau)} - \frac{(1+\tau)^2}{6}(u_{\bar{x}\bar{x}\bar{x}})_j^n \tag{5.48}$$

Because points $P^+$ and $P^-$ coincide with the mesh points $(j+1/2, n)$ and $(j-1/2, n)$, respectively, when $\tau = 1$, the current version of $(u_{\bar{x}}^c)_j^n$ is defined by

$$(u_{\bar{x}}^{c\text{-}4})_j^n \stackrel{\text{def}}{=} \left[ u_{\bar{x}}^4(1) \right]_j^n = \frac{1}{4}\left[ u - 2\nu u_{\bar{x}} + 2\nu^2 u_{\bar{x}\bar{x}} - \frac{4\nu^3}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_{j+1/2}^{n-1/2}$$
$$- \frac{1}{4}\left[ u - 2\nu u_{\bar{x}} + 2\nu^2 u_{\bar{x}\bar{x}} - \frac{4\nu^3}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_{j-1/2}^{n-1/2} - \frac{2}{3}(u_{\bar{x}\bar{x}\bar{x}})_j^n \qquad (j, n) \in \Omega \tag{5.49}$$

The validity of the last equality sign in Eq. (5.49) can be proved by using Eqs. (5.44), (5.45), and (5.48) with $\tau = 1$. Substituting Eq. (5.26) into the last term on the right side of Eq. (5.49) and denoting the resulting expression as $\left[ u_{\bar{x}}^{c\text{-}4}(\epsilon_1) \right]_j^n$, one concludes that

$$\left[ u_{\bar{x}}^{c\text{-}4}(\epsilon_1) \right]_j^n = \frac{1}{4}\left\{ u - 2\nu\, u_{\bar{x}} + \left[ 2\nu^2 - \frac{4(1-\epsilon_1)}{3} \right]u_{\bar{x}\bar{x}} + \frac{4(1 + \nu - \nu^3 - 2\epsilon_1)}{3}u_{\bar{x}\bar{x}\bar{x}} \right\}_{j+1/2}^{n-1/2}$$
$$- \frac{1}{4}\left\{ u - 2\nu\, u_{\bar{x}} + \left[ 2\nu^2 - \frac{4(1-\epsilon_1)}{3} \right]u_{\bar{x}\bar{x}} - \frac{4(1 - \nu + \nu^3 - 2\epsilon_1)}{3}u_{\bar{x}\bar{x}\bar{x}} \right\}_{j-1/2}^{n-1/2} \qquad (j, n) \in \Omega \tag{5.50}$$

26

By definition, the $a$-$\epsilon$-4 scheme is formed by Eqs. (5.14), (5.26), (5.34), and

$$(u_{\bar{x}})^n_j = \left[u^{a\text{-}4}_{\bar{x}}(\epsilon_1)\right]^n_j + 2\epsilon_2 \left[u^{c\text{-}4}_{\bar{x}}(\epsilon_1) - u^{a\text{-}4}_{\bar{x}}(\epsilon_1)\right]^n_j \qquad (j,n) \in \Omega \qquad (5.51)$$

where $\epsilon_2$ is another adjustable parameter. With the aid of Eqs. (5.36) and (5.50), Eq. (5.51) can be expressed as

$$
\begin{aligned}
(u_{\bar{x}})^n_j = \frac{1}{6}\Big\{ &3(1-\epsilon_2)u - 3(1+\nu-2\epsilon_2)u_{\bar{x}} + \left[1 + 3\nu^2 - 6\epsilon_2 + (1+\nu^2)\epsilon_1 + 2(1-\nu^2)\epsilon_1\epsilon_2\right]u_{\bar{x}\bar{x}} \\
&+ 2\left[\nu(1-\nu^2) + 2\epsilon_2 - (1+\nu^2)\epsilon_1 - 2(1-\nu^2)\epsilon_1\epsilon_2\right]u_{\bar{x}\bar{x}\bar{x}}\Big\}^{n-1/2}_{j+1/2} \\
-\frac{1}{6}\Big\{ &3(1-\epsilon_2)u + 3(1-\nu-2\epsilon_2)u_{\bar{x}} + \left[1 + 3\nu^2 - 6\epsilon_2 + (1+\nu^2)\epsilon_1 + 2(1-\nu^2)\epsilon_1\epsilon_2\right]u_{\bar{x}\bar{x}} \\
&+ 2\left[\nu(1-\nu^2) - 2\epsilon_2 + (1+\nu^2)\epsilon_1 + 2(1-\nu^2)\epsilon_1\epsilon_2\right]u_{\bar{x}\bar{x}\bar{x}}\Big\}^{n-1/2}_{j-1/2} \qquad (j,n) \in \Omega
\end{aligned}
$$

$$(5.52)$$

We conclude this subsection with the following remarks:

(a) Eqs. (5.51) reduces to (i) $(u_{\bar{x}})^n_j = \left[u^{a\text{-}4}_{\bar{x}}(\epsilon_1)\right]^n_j$ if $\epsilon_2 = 0$; and (ii) $(u_{\bar{x}})^n_j = \left[u^{c\text{-}4}_{\bar{x}}(\epsilon_1)\right]^n_j$ if $\epsilon_2 = 1/2$. *Because the $a$-4 scheme is the special scheme formed by Eqs. (5.14), (5.26), (5.34) and (5.35) with $\epsilon_1 = 0$, the $a$-$\epsilon$-4 scheme reduces to the $a$-4 scheme when $\epsilon_1 = \epsilon_2 = 0$.*

(b) Because of how they are derived, by replacing the symbol $\epsilon_2$ with $\epsilon$ and dropping all the 2nd-order and 3rd-order terms, Eqs. (5.34) and (5.52), respectively, will become Eqs. (2.10) and (2.19a), the defining equations of the $a$-$\epsilon$ scheme.

In the following, we will discuss the stability of the $a$-$\epsilon$-4 scheme. As a preliminary, the scheme will be cast into a matrix form.

## 5.3. Matrix form of the $a$-$\epsilon$-4 scheme

The component equations of the $a$-$\epsilon$-4 scheme, i.e., Eqs. (5.14), (5.26), (5.34) and (5.52), can be cast into the matrix equation

$$\vec{q}(j,n) = Q_-(\epsilon_1, \epsilon_2, \nu)\,\vec{q}(j+1/2, n-1/2) + Q_+(\epsilon_1, \epsilon_2, \nu)\,\vec{q}(j-1/2, n-1/2) \qquad (j,n) \in \Omega \qquad (5.53)$$

where

$$\vec{q}(j,n) \stackrel{\text{def}}{=} \begin{pmatrix} u^n_j \\ (u_{\bar{x}})^n_j \\ (u_{\bar{x}\bar{x}})^n_j \\ (u_{\bar{x}\bar{x}\bar{x}})^n_j \end{pmatrix} \qquad (j,n) \in \Omega \qquad (5.54)$$

and

$$Q_\pm(\epsilon_1, \epsilon_2, \nu) \stackrel{\text{def}}{=} \begin{pmatrix} q^\pm_{11} & q^\pm_{12} & q^\pm_{13} & q^\pm_{14} \\ q^\pm_{21} & q^\pm_{22} & q^\pm_{23} & q^\pm_{24} \\ 0 & 0 & q^\pm_{33} & q^\pm_{34} \\ 0 & 0 & q^\pm_{43} & q^\pm_{44} \end{pmatrix} \qquad (5.55)$$

with

$$q^\pm_{11} \stackrel{\text{def}}{=} \frac{1\pm\nu}{2}, \qquad q^\pm_{12} \stackrel{\text{def}}{=} \pm\frac{1-\nu^2}{2}, \qquad q^\pm_{13} \stackrel{\text{def}}{=} \mp\frac{\nu(1-\nu^2)}{3}, \qquad q^\pm_{14} \stackrel{\text{def}}{=} \mp\frac{(1-\nu^2)^2}{6} \qquad (5.56)$$

27

$$q_{21}^\pm \overset{\text{def}}{=} \mp \frac{1 - \epsilon_2}{2}, \quad q_{22}^\pm \overset{\text{def}}{=} \frac{2\epsilon_2 - 1 \pm \nu}{2}, \quad q_{23}^\pm \overset{\text{def}}{=} \mp \frac{1 + 3\nu^2 - 6\epsilon_2 + (1 + \nu^2)\epsilon_1 + 2(1 - \nu^2)\epsilon_1\epsilon_2}{6}$$

$$q_{24}^\pm \overset{\text{def}}{=} \mp \frac{\nu(1 - \nu^2) \mp 2\epsilon_2 \pm (1 + \nu^2)\epsilon_1 \pm 2(1 - \nu^2)\epsilon_1\epsilon_2}{3} \tag{5.57}$$

$$q_{33}^\pm \overset{\text{def}}{=} \frac{1 \pm \nu}{2}, \quad q_{34}^\pm \overset{\text{def}}{=} \pm \frac{1 - \nu^2}{2}, \quad q_{43}^\pm \overset{\text{def}}{=} \mp \frac{1 - \epsilon_1}{2}, \quad q_{44}^\pm \overset{\text{def}}{=} \frac{2\epsilon_1 - 1 \pm \nu}{2} \tag{5.58}$$

For simplicity, hereafter $Q_+(\epsilon_1, \epsilon_2, \nu)$ and $Q_-(\epsilon_1, \epsilon_2, \nu)$ may be abbreviated as $Q_+$ and $Q_-$, respectively.

Eq. (5.53) represents the form of the $a$-$\epsilon$-4 scheme by which mesh variables can be advanced over $\Delta t/2$. By applying Eq. (5.53) successively over two consecutive marching steps, one obtain a compounded form by which mesh variables can be advanced over $\Delta t$, i.e.,

$$\vec{q}(j, n+1) = (Q_+)^2 \vec{q}(j-1, n) + (Q_+ Q_- + Q_- Q_+) \vec{q}(j, n) + (Q_-)^2 \vec{q}(j+1, n) \qquad (j, n) \in \Omega \tag{5.59}$$

Eq. (5.59) has the advantage that the mesh variables involved are associated with mesh points which are not staggered in space-time.

To further explore the relations between the $a$-$\epsilon$-4 and $a$-$\epsilon$ schemes, note that the component equations of the $a$-$\epsilon$ scheme, i.e., Eqs. (2.10) and (2.19a), can be cast into the matrix form

$$\underline{\vec{q}}(j, n) = M_-(\epsilon, \nu) \, \underline{\vec{q}}(j + 1/2, n - 1/2) + M_+(\epsilon, \nu) \, \underline{\vec{q}}(j - 1/2, n - 1/2) \qquad (j, n) \in \Omega \tag{5.60}$$

where

$$\underline{\vec{q}}(j, n) \overset{\text{def}}{=} \begin{pmatrix} u_j^n \\ (u_{\bar{x}})_j^n \end{pmatrix} \qquad (j, n) \in \Omega \tag{5.61}$$

and

$$M_\pm(\epsilon, \nu) \overset{\text{def}}{=} \frac{1}{2} \begin{pmatrix} 1 \pm \nu & \pm(1 - \nu^2) \\ \mp(1 - \epsilon) & 2\epsilon - 1 \pm \nu \end{pmatrix} \qquad -\infty < \epsilon < +\infty \tag{5.62}$$

With the aid of Eqs. (5.56)–(5.58) and (5.62), it can be shown that

$$\begin{pmatrix} q_{11}^\pm & q_{12}^\pm \\ q_{21}^\pm & q_{22}^\pm \end{pmatrix} = M_\pm(\epsilon_2, \nu) \tag{5.63}$$

and

$$\begin{pmatrix} q_{33}^\pm & q_{34}^\pm \\ q_{43}^\pm & q_{44}^\pm \end{pmatrix} = M_\pm(\epsilon_1, \nu) \tag{5.64}$$

By comparing Eqs. (5.63) and (5.64) with Eq. (5.55), one now arrives at the conclusion that the $2 \times 2$ submatrices in the top-left and bottom right corners of the $4 \times 4$ matrix $Q_\pm(\epsilon_1, \epsilon_2, \nu)$ are $M_\pm(\epsilon_2, \nu)$ and $M_\pm(\epsilon_1, \nu)$, respectively.

Because of the relations Eqs. (5.63) and (5.64), as will be shown in the next subsection, the $a$-$\epsilon$-4 and $a$-$\epsilon$ schemes share essentially the same stability conditions.

### 5.4. Stability conditions of the $a$-$\epsilon$-4 scheme

Note that a scheme such as Eq. (5.53) would become computationally unstable if the round-off errors introduced initially at any time level are amplified without bound as they propagate down the subsequent time levels. To be specific, consider the effect of contamination of the initial values by round-off errors. Let $\vec{q}(j, n)$ denote the solution associated with the given exact initial values while $\vec{q}(j, n) + \delta\vec{q}(j, n)$ denote the solution associated with the contaminated initial values. Then $\vec{q}(j, n)$ must satisfy Eq. (5.53). Also, because

28

the coefficient matrices $Q_+$ and $Q_-$ in the linear scheme Eq. (5.53) have constant elements independent of mesh variables, the contaminated solution satisfies the equation

$$\vec{q}(j,n) + \delta\vec{q}(j,n) = Q_-\left[\vec{q}(j+1/2, n-1/2) + \delta\vec{q}(j+1/2, n-1/2)\right]$$
$$+ Q_+\left[\vec{q}(j-1/2, n-1/2) + \delta\vec{q}(j-1/2, n-1/2)\right] \qquad (j,n) \in \Omega \qquad (5.65)$$

By subtracting Eq. (5.53) from (5.65), one has

$$\delta\vec{q}(j,n) = Q_-\,\delta\vec{q}(j+1/2, n-1/2) + Q_+\,\delta\vec{q}(j-1/2, n-1/2) \qquad (j,n) \in \Omega \qquad (5.66)$$

Thus the time evolution of $\delta\vec{q}(j,n)$, which represents the solution errors induced by the initial-value contamination, is also governed by the scheme Eq. (5.53). For this reason, stability of the linear scheme Eq. (5.53) can be determined by the time-evolution behavior of $\vec{q}(j,n)$ itself.

The stability of the $a$-$\epsilon$-4 scheme will be studied using the von Neumann method. In this method (see Sec. 4 in [1] for a rigorous treatment based on the discrete Fourier analysis), the time-evolution behavior of each Fourier component (characterized by a phase angle $\theta$) of $\vec{q}(j,n)$ is studied separately. Because these components propagate in time independently for a linear scheme such as Eq. (5.53), the scheme is stable if only if each component is bounded as it evolves in time.

Let

$$\vec{q}(j,n) = \vec{q}^*(n,\theta)e^{ij\theta} \qquad (j,n) \in \Omega;\ -\infty < \theta < +\infty;\ i \overset{\text{def}}{=} \sqrt{-1} \qquad (5.67)$$

Here (i) $\theta$ is the phase angle variation of a Fourier component over the length $\Delta x$, and (ii) $\vec{q}^*(n,\theta)$ is a $2 \times 1$ column matrix. Substituting Eq. (5.67) into Eq. (5.59), one has

$$\vec{q}^*(n+1,\theta) = [Q(\epsilon_1, \epsilon_2, \nu, \theta)]^2\,\vec{q}^*(n,\theta) \qquad n = 0, \pm 1/2, \pm 1, \pm 3/2, \ldots \qquad (5.68)$$

where

$$Q(\epsilon_1, \epsilon_2, \nu, \theta) = e^{i\theta/2}Q_-(\epsilon_1, \epsilon_2, \nu) + e^{-i\theta/2}Q_+(\epsilon_1, \epsilon_2, \nu)$$

$$= \begin{pmatrix} e^{i\theta/2}q_{11}^- + e^{-i\theta/2}q_{11}^+ & e^{i\theta/2}q_{12}^- + e^{-i\theta/2}q_{12}^+ & e^{i\theta/2}q_{13}^- + e^{-i\theta/2}q_{13}^+ & e^{i\theta/2}q_{14}^- + e^{-i\theta/2}q_{14}^+ \\ e^{i\theta/2}q_{21}^- + e^{-i\theta/2}q_{21}^+ & e^{i\theta/2}q_{22}^- + e^{-i\theta/2}q_{22}^+ & e^{i\theta/2}q_{23}^- + e^{-i\theta/2}q_{23}^+ & e^{i\theta/2}q_{24}^- + e^{-i\theta/2}q_{24}^+ \\ 0 & 0 & e^{i\theta/2}q_{33}^- + e^{-i\theta/2}q_{33}^+ & e^{i\theta/2}q_{34}^- + e^{-i\theta/2}q_{34}^+ \\ 0 & 0 & e^{i\theta/2}q_{43}^- + e^{-i\theta/2}q_{43}^+ & e^{i\theta/2}q_{44}^- + e^{-i\theta/2}q_{44}^+ \end{pmatrix} \qquad (5.69)$$

Because of Eq. (5.68), $[Q(\epsilon_1, \epsilon_2, \nu, \theta)]^2$ is referred to as the amplification matrix of the $a$-$\epsilon$-4 scheme per marching cycle (i.e., per $\Delta t$) while $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ is referred to as the amplification factor per marching step (i.e., per $\Delta t/2$). Also by using Eq. (5.68), one has

$$\vec{q}^*(n+m,\theta) = [Q(\epsilon_1, \epsilon_2, \nu, \theta)]^{2m}\vec{q}^*(n,\theta) \qquad m = 1,2,3\ldots;\ n = 0, \pm 1/2, \pm 1, \pm = 3/2, \ldots \qquad (5.70)$$

There is an important relation between the matrix $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ and the amplification matrix per marching step of the $a$-$\epsilon$ scheme, i.e.,

$$M(\epsilon, \nu, \theta) \overset{\text{def}}{=} e^{i\theta/2}M_-(\epsilon, \nu) + e^{-i\theta/2}M_+(\epsilon, \nu) \equiv$$

$$\begin{pmatrix} \cos(\theta/2) - i\nu\sin(\theta/2) & -i(1-\nu^2)\sin(\theta/2) \\ i(1-\epsilon)\sin(\theta/2) & (2\epsilon - 1)\cos(\theta/2) - i\nu\sin(\theta/2) \end{pmatrix} \qquad -\infty < \epsilon < +\infty \qquad (5.71)$$

In fact, it follow from Eqs. (5.63), (5.64), (5.69), and (5.71) that

$$\begin{pmatrix} e^{i\theta/2}q_{11}^- + e^{-i\theta/2}q_{11}^+ & e^{i\theta/2}q_{12}^- + e^{-i\theta/2}q_{12}^+ \\ e^{i\theta/2}q_{21}^- + e^{-i\theta/2}q_{21}^+ & e^{i\theta/2}q_{22}^- + e^{-i\theta/2}q_{22}^+ \end{pmatrix} = M(\epsilon_2, \nu, \theta) \tag{5.72}$$

and

$$\begin{pmatrix} e^{i\theta/2}q_{33}^- + e^{-i\theta/2}q_{33}^+ & e^{i\theta/2}q_{34}^- + e^{-i\theta/2}q_{34}^+ \\ e^{i\theta/2}q_{43}^- + e^{-i\theta/2}q_{43}^+ & e^{i\theta/2}q_{44}^- + e^{-i\theta/2}q_{44}^+ \end{pmatrix} = M(\epsilon_1, \nu, \theta) \tag{5.73}$$

i.e., the $2 \times 2$ submatrices in the top-left and bottom-right corners of the matrix $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ are $M(\epsilon_2, \nu, \theta)$ and $M(\epsilon_1, \nu, \theta)$, respectively. As such, *the amplification matrix of the a-$\epsilon$ scheme is identical to each of the above $2 \times 2$ submatrices of $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ except that the parameter $\epsilon$ in the former matrix is replaced by $\epsilon_2$ or $\epsilon_1$ in the latter two submatrices.* By using this property, it will be shown that, *for each set of $\epsilon_1$, $\epsilon_2$, $\nu$, and $\theta$, the eigenvalues of $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ can be expressed as $\lambda_\pm(\epsilon_1, \nu, \theta)$ and $\lambda_\pm(\epsilon_2, \nu, \theta)$ where the function $\lambda_\pm(\epsilon, \nu, \theta)$ was defined in Eq. (2.21).*

To prove the above proposition, note that, for each set of $\epsilon_1$, $\epsilon_2$, $\nu$, and $\theta$, the eigenvalues $\lambda$ of $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ are the four roots of the characteristic equation

$$\det\left[Q(\epsilon_1, \epsilon_2, \nu, \theta) - \lambda I_4\right] = 0 \tag{5.74}$$

Hereafter $I_n$ denote the $n \times n$ identity matrix. Let $0_2$ denote the $2 \times 2$ zero matrix. Then according to Eqs. (5.69), (5.72), and (5.73), the $2 \times 2$ submatrices in the top-left, bottom-left, and bottom-right corners of the $4 \times 4$ matrix $[Q(\epsilon_1, \epsilon_2, \nu, \theta) - \lambda I_4]$ are $[M(\epsilon_2, \nu, \theta) - \lambda I_2]$, $0_2$, and $[M(\epsilon_1, \nu, \theta) - \lambda I_2]$, respectively. As such, by using elementary theory of determinant, we have

$$\det\left[Q(\epsilon_1, \epsilon_2, \nu, \theta) - \lambda I_4\right] \equiv \det\left[M(\epsilon_1, \nu, \theta) - \lambda I_2\right] \cdot \det\left[M(\epsilon_2, \nu, \theta) - \lambda I_2\right] \tag{5.75}$$

It now follows from Eqs. (5.74) and (5.75) that, for each set of $\epsilon_1$, $\epsilon_2$, and $\nu$, the eigenvalues of $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ are the four roots of the characteristic equations

$$\det\left[M(\epsilon_1, \nu, \theta) - \lambda I_2\right] = 0 \tag{5.76}$$

and

$$\det\left[M(\epsilon_2, \nu, \theta) - \lambda I_2\right] = 0 \tag{(5.77)}$$

Because the roots of Eq. (5.76) and ((5.77) are $\lambda_\pm(\epsilon_1, \nu, \theta)$ and $\lambda_\pm(\epsilon_2, \nu, \theta)$, respectively, the proof is completed. QED.

Because of the relation Eq. (5.70), *for a given set of $\epsilon_1$, $\epsilon_2$, and $\nu$, the a-$\epsilon$-4 scheme is said to be stable relative to this set if and only if, for all $\theta$, $-\infty < \theta < +\infty$, all the elements of the matrix $[Q(\epsilon_1, \epsilon_2, \nu, \theta)]^m$ remain bounded as $m \to +\infty$.* By using the Jordan canonical form theorem [66] and the proposition we have just proved, it can be shown that, for each given set of $\epsilon_1$, $\epsilon_2$, $\nu$, and $\theta$, all the elements of $[Q(\epsilon_1, \epsilon_2, \nu, \theta)]^m$ remain bounded as $m \to +\infty$ if and only if

$$\max\left\{|\lambda_+(\epsilon_1, \nu, \theta)|, |\lambda_-(\epsilon_1, \nu, \theta)|, |\lambda_+(\epsilon_2, \nu, \theta)|, |\lambda_-(\epsilon_2, \nu, \theta)|\right\} \leq 1 \quad \text{if } Q(\epsilon_1, \epsilon_2, \nu, \theta) \text{ is nondefective} \tag{5.78}$$

and

$$\max\left\{|\lambda_+(\epsilon_1, \nu, \theta)|, |\lambda_-(\epsilon_1, \nu, \theta)|, |\lambda_+(\epsilon_2, \nu, \theta)|, |\lambda_-(\epsilon_2, \nu, \theta)|\right\} < 1 \quad \text{if } Q(\epsilon_1, \epsilon_2, \nu, \theta) \text{ is defective} \tag{5.79}$$

Note that an $n \times n$ matrix is said to be defective if and only if the number of its linearly independent eigenvectors is less than $n$ [66].

30

From the above discussions, it is seen that, for a given set of $\epsilon_1$, $\epsilon_2$, and $\nu$, a necessary condition for the stability of $a$-$\epsilon$-4 scheme is

$$\max_{-\infty<\theta<+\infty} \{|\lambda_+(\epsilon_1,\nu,\theta)|, |\lambda_-(\epsilon_1,\nu,\theta)|, |\lambda_+(\epsilon_2,\nu,\theta)|, |\lambda_-(\epsilon_2,\nu,\theta)|\} \le 1 \tag{5.80}$$

By using Eq. (2.21) and a line of arguments which was used in establishing the stability condition Eq. (3.14) of [5] (Note: In [5], the $a$-$\epsilon$ scheme is not defined for the case $|\nu| = 1$), one can show that Eq. (5.80) $\Leftrightarrow$

$$0 \le \epsilon_1 \le 1, \qquad 0 \le \epsilon_2 \le 1 \qquad \text{and} \quad |\nu| \le 1 \tag{5.81}$$

Thus Eq. (5.81) represents a set of necessary stability conditions for the $a$-$\epsilon$-4 scheme.

Because $Q(\epsilon_1, \epsilon_2, \nu, \theta)$ does become defective for some special values of $\epsilon_1$ $\epsilon_2$, $\nu$ and $\theta$, according to Eq. (5.79), Eq. (5.80) and therefore Eq. (5.81) generally is not sufficient to insure the stability of the $a$-$\epsilon$-4 scheme. However, it is expected that the necessary and sufficient stability conditions (not yet established) for the $a$-$\epsilon$-4 scheme would be formed by Eq. (5.81) and some exclusion conditions involving some boundaries of the stability domain defined in Eq. (5.81).

From the above discussions, one concludes that *the 2nd-order and 3rd-order terms which appear on the right sides of Eqs. (5.34) and (5.52) (which are associated with the $2 \times 2$ submatrices in the top-right corners of the matrices $Q_+$ and $Q_-$) play only a marginal role in the stability of the $a$-$\epsilon$-4 scheme even though they play a decisive role in the 4th-order accuracy of the scheme*

Next we will present a set of non-reflecting boundary conditions with 4th-order accuracy.

## 5.5. Non-reflecting boundary conditions

Let $(j, n)$ be a mesh point on the right spatial boundary. It has been shown numerically the current extension of Eq. (2.44) i.e.,

$$u_j^n = u_{j-1/2}^{n-1/2}, \quad (u_{\bar{x}})_j^n = (u_{\bar{x}})_{j-1/2}^{n-1/2}, \quad (u_{\bar{x}\bar{x}})_j^n = (u_{\bar{x}\bar{x}})_{j-1/2}^{n-1/2}, \quad \text{and} \quad (u_{\bar{x}\bar{x}\bar{x}})_j^n = (u_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2} \tag{5.82}$$

and a similar extension of Eq. (2.45) also form sets of non-reflecting boundary conditions. However, these simple extensions are not 4th order in accuracy. In the following, we present non-reflecting boundary conditions with 4th-order accuracy.

Let $(j, n)$ be a mesh point on the right spatial boundary. Again we assume

$$(u_{\bar{x}\bar{x}\bar{x}})_j^n = (u_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2} \tag{5.83}$$

However, $(u_{\bar{x}\bar{x}})_j^n$ will be determined in terms of the mesh variables at $(j - 1/2, n - 1/2)$ by imposing the conservation condition Eq. (5.13) which links the mesh variables at the two diagonally opposite mesh points $(j, n)$ and $(j - 1/2, n - 1/2)$. Because Eq. (5.13) $\Leftrightarrow$

$$(1 + \nu)\left[u_{\bar{x}\bar{x}} - (1 - \nu)u_{\bar{x}\bar{x}\bar{x}}\right]_j^n = (1 + \nu)\left[u_{\bar{x}\bar{x}} + (1 - \nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2} \tag{5.84}$$

we assume

$$\left[u_{\bar{x}\bar{x}} - (1 - \nu)u_{\bar{x}\bar{x}\bar{x}}\right]_j^n = \left[u_{\bar{x}\bar{x}} + (1 - \nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2} \tag{5.85}$$

(Eq. (5.85) is a condition stronger than Eq. (5.84)). By combining Eqs. (5.83) and (5.85), one has

$$(u_{\bar{x}\bar{x}})_j^n = \left[u_{\bar{x}\bar{x}} + 2(1 - \nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2} \tag{5.86}$$

Next $(u_{\bar{x}})_j^n$ will be evaluated using the Taylor's expansion

$$(u_x)_j^n = \left[u_x + \frac{\Delta x}{2}u_{xx} + \frac{\Delta t}{2}u_{xt} + \frac{1}{2}\left(\frac{\Delta x}{2}\right)^2 u_{xxx} + \frac{\Delta x}{2}\frac{\Delta t}{2}u_{xxt} + \frac{1}{2}\left(\frac{\Delta t}{2}\right)^2 u_{xtt}\right]_{j-1/2}^{n-1/2} \tag{5.87}$$

31

With the aid of Eqs. (2.9), (5.3), and (5.17), Eq. (5.87) can be rewritten as

$$(u_{\bar{x}})_j^n = \left[ u_{\bar{x}} + 2(1 - \nu)u_{\bar{x}\bar{x}} + 2(1 - \nu)^2 u_{\bar{x}\bar{x}\bar{x}} \right]_{j-1/2}^{n-1/2} \tag{5.88}$$

The mesh variable $u_j^n$ will be evaluated in terms of the mesh variables at $(j - 1/2, n - 1/2)$ by imposing the conservation condition Eq. (2.8). Because Eq. (2.8) $\Leftrightarrow$ Eq. (5.30), we will assume Eq. (5.30a) (a stronger condition than Eq. (5.30)). As such one has

$$u_j^n = \left[ (1 - \nu)u_{\bar{x}} - \frac{2(1 - \nu + \nu^2)}{3}u_{\bar{x}\bar{x}} + \frac{(1 - \nu)(1 + \nu^2)}{3}u_{\bar{x}\bar{x}\bar{x}} \right]_j^n + (s_-)_{j-1/2}^{n-1/2} \tag{5.89}$$

where $(s_-)_{j-1/2}^{n-1/2}$ is defined in Eq. (5.28). With the aid of Eqs. (5.83), (5.86), and (5.88), $u_j^n$ can now be evaluated in terms of the mesh variables at $(j - 1/2, n - 1/2)$.

A similar treatment can also be applied at the left spatial boundary. As will be shown, these new non-reflecting boundary conditions are 4th-order in accuracy.

## 6. The $c$-$\tau$-4 and $c$-$\tau^*$-4 schemes

The $c$-$\tau$-4 and $c$-$\tau^*$-4 scheme will be built from the $c$-$\tau$ scheme in a procedure similar to that was used to built the $a$-$\epsilon$-4 scheme from $a$-$\epsilon$ scheme. Thus the basic equations including Eqs. (2.4), (2.9), (5.1)–(5.7), (5.11), and (5.17) will be assumed here. In addition, both the conservation conditions Eq. (2.13) and (5.18) will also be assumed here.

### 6.1. The $c$-$\tau$-4 scheme

Because Eq. (5.18) is equivalent to Eq. (5.14) assuming Eqs. (5.7), (5.11), and (5.17), in the current development, *the mesh variable $(u_{\bar{x}\bar{x}})_j^n$ will be evaluated using Eq. (5.14)*. Moreover, because (i) Eq. (5.31) is equivalent to Eq. (2.13) assuming Eqs. (2.4), (2.9), (5.4), and (5.17); and (ii) Eq. (5.34) can be obtained from Eqs. (5.31) and (5.14), *the mesh variable $u_j^n$ will be evaluated using Eq. (5.34)*. As such, the current schemes share with the $a$-$\epsilon$-4 scheme two component equations, i.e., Eqs. (5.14) and (5.34).

In the $a$-$\epsilon$-4 scheme, $(u_{\bar{x}\bar{x}\bar{x}})_j^n$ is evaluated using Eq. (5.26) which is the substitution form of Eq. (2.19a) (by which $(u_{\bar{x}})_j^n$ is evaluated in the $a$-$\epsilon$ scheme) except that the symbol $\epsilon$ in Eq. (2.19a) is replaced by $\epsilon_1$ in Eq. (5.26). Similarly, in the $c$-$\tau$-4 scheme, $(u_{\bar{x}\bar{x}\bar{x}})_j^n$ is evaluated using the substitution form of Eq. (3.11) (by which $(u_{\bar{x}})_j^n$ is evaluated in the $c$-$\tau$ scheme) with the understanding that the symbol $\tau$ in Eq. (3.11) is replaced by $\tau_1$ in the substitution form. Thus we have

$$(u_{\bar{x}\bar{x}\bar{x}})_j^n = \frac{1}{2(1 + \tau_1)} \left\{ \left[ u_{\bar{x}\bar{x}} - (1 + 2\nu - \tau_1)u_{\bar{x}\bar{x}\bar{x}} \right]_{j+1/2}^{n-1/2} - \left[ u_{\bar{x}\bar{x}} + (1 - 2\nu - \tau_1)u_{\bar{x}\bar{x}\bar{x}} \right]_{j-1/2}^{n-1/2} \right\}$$
$$(j, n) \in \Omega; \ \tau_1 \neq -1 \tag{6.1}$$

In the $c$-$\tau$-4 scheme, $(u_{\bar{x}})_j^n$ will be evaluated using an equation which is the result of the following assumptions: (i) $(u_{\bar{x}})_j^n = \left[ u_{\bar{x}}^4(\tau_2) \right]_j^n$ (see Eq. (5.48)); (ii) the mesh variable $(u_{\bar{x}\bar{x}\bar{x}})_j^n$ which appears on the right side of Eq. (5.48) is evaluated using Eq. (6.1); and (iii) the parameter $\tau$ which appears on the right sides of Eqs. (5.44) and (5.45) is replaced by $\tau_2$. Then, by using Eqs. (5.44), (5.46), (5.48), and (6.1), one

32

has

$$(u_{\bar x})_j^n = \frac{1}{2} \left\{ \frac{u}{1+\tau_2} - \frac{1-\tau_2+2\nu}{1+\tau_2} u_{\bar x} + \frac{1}{2}\left[ \frac{(1-\tau_2+2\nu)^2}{1+\tau_2} - \frac{(1+\tau_2)^2}{3(1+\tau_1)} \right] u_{\bar x \bar x} \right.$$

$$\left. + \frac{1}{6}\left[ \frac{(1+\tau_2)^2(1-\tau_1+2\nu)}{1+\tau_1} - \frac{(1-\tau_2+2\nu)^3}{1+\tau_2} \right] u_{\bar x \bar x \bar x} \right\}_{j+1/2}^{n-1/2}$$

$$- \frac{1}{2}\left\{ \frac{u}{1+\tau_2} + \frac{1-\tau_2-2\nu}{1+\tau_2} u_{\bar x} + \frac{1}{2}\left[ \frac{(1-\tau_2-2\nu)^2}{1+\tau_2} - \frac{(1+\tau_2)^2}{3(1+\tau_1)} \right] u_{\bar x \bar x} \right. \tag{6.2}$$

$$\left. - \frac{1}{6}\left[ \frac{(1+\tau_2)^2(1-\tau_1-2\nu)}{1+\tau_1} - \frac{(1-\tau_2-2\nu)^3}{1+\tau_2} \right] u_{\bar x \bar x \bar x} \right\}_{j-1/2}^{n-1/2}$$

$$(j,n) \in \Omega; \ \tau_1 \neq -1; \ \tau_2 \neq -1$$

By definition, the $c$-$\tau$-4 scheme is formed by Eqs. (5.14), (5.34), (6.1), and (6.2). It enforces the local conservation relations Eqs. (2.13) and (5.18). According to an argument given in Comment (d) presented at the end of Sec. 2.1, it will also enforce a global conservation relation for each of the fluxes of $\vec{h}^*$ and $\vec{h}^*_{xx}$ over any space-time region that is the union of any combination of CE$(j,n)$, $(j,n) \in \Omega$. Moreover, the scheme is completely explicit, i.e., the four independent mesh variables at each $(j,n)$ can be evaluated directly in terms of those at the $(n-1/2)$th time level without resorting to solving a system of linear equations involving the independent mesh variables at $(j,n)$ such as Eqs. (2.7a) and (2.8a).

The $c$-$\tau$ and $a$ schemes become the same scheme when $\nu = \tau = 0$ (see Eq. (3.12)). It happens that the $c$-$\tau$-4 and $a$-4 schemes also become the same scheme when $\nu = \tau_1 = \tau_2 = 0$. In fact, for the case $\nu = \tau_1 = \tau_2 = 0$, Eqs. (6.1) and (6.2) reduce to

$$(u_{\bar x \bar x \bar x})_j^n = \frac{1}{2}\left[ (u_{\bar x \bar x} - u_{\bar x \bar x \bar x})_{j+1/2}^{n-1/2} - (u_{\bar x \bar x} + u_{\bar x \bar x \bar x})_{j-1/2}^{n-1/2} \right] \tag{6.3}$$

and

$$(u_{\bar x})_j^n = \frac{1}{2}\left[ \left( u - u_{\bar x} + \frac{1}{3}u_{\bar x \bar x} \right)_{j+1/2}^{n-1/2} - \left( u + u_{\bar x} + \frac{1}{3}u_{\bar x \bar x} \right)_{j-1/2}^{n-1/2} \right] \tag{6.4}$$

respectively. It can be shown that Eqs. (5.26) and (5.52) (the counterparts of Eqs. (6.1) and (6.2) in the $a$-$\epsilon$-4 scheme) also reduce to Eqs. (6.3) and (6.4), respectively, when $\nu = \epsilon_1 = \epsilon_2 = 0$. Because (i) the $a$-4 scheme is a special case of $a$-$\epsilon$-4 scheme with $\epsilon_1 = \epsilon_2 = 0$ and (ii) Eqs. (5.14) and (5.34) are shared by the $c$-$\tau$-4 and $a$-$\epsilon$-4 schemes, the proof is completed. QED.

Because the conservation relations Eqs. (2.7), (2.8), (5.12), and (5.13) are satisfied by the $a$-4 scheme for all $\nu$, it now follows that these conservation relations are satisfied by the $c$-$\tau$-4 scheme when $\nu = \tau_1 = \tau_2 = 0$.

The $c$-$\tau$ and $c$ schemes (the special case of the $a$-$\epsilon$ scheme with $\epsilon = 1/2$) become the same scheme when $\tau = 1$ (see Eq. 3.12). Let the $c$-4 scheme be the special case of the $a$-$\epsilon$-4 scheme with $\epsilon_1 = \epsilon_2 = 1/2$. Then one can also show that the $c$-$\tau$-4 and $c$-4 schemes become the same scheme when $\tau_1 = \tau_2 = 1$. In fact, for the case $\tau_1 = \tau_2 = 1$, Eqs. (6.1) and (6.2) reduce to

$$(u_{\bar x \bar x \bar x})_j^n = \frac{1}{4}\left[ (u_{\bar x \bar x} - 2\nu\, u_{\bar x \bar x \bar x})_{j+1/2}^{n-1/2} - (u_{\bar x \bar x} - 2\nu\, u_{\bar x \bar x \bar x})_{j-1/2}^{n-1/2} \right] \tag{6.5}$$

and

$$(u_{\bar x})_j^n = \frac{1}{12}\left\{ \left[ 3u - 6\nu\, u_{\bar x} + 2(3\nu^2 - 1)u_{\bar x \bar x} + 4\nu(1-\nu^2)u_{\bar x \bar x \bar x} \right]_{j+1/2}^{n-1/2} \right.$$

$$\left. - \left[ 3u - 6\nu\, u_{\bar x} + 2(3\nu^2 - 1)u_{\bar x \bar x} + 4\nu(1-\nu^2)u_{\bar x \bar x \bar x} \right]_{j-1/2}^{n-1/2} \right\} \tag{6.6}$$

33

respectively. It can be shown that Eqs. (5.26) and (5.52) also reduce to Eqs. (6.5) and (6.6), respectively, when $\epsilon_1 = \epsilon_2 = 1/2$ (i.e., when the $a$-$\epsilon$-4 scheme becomes the $c$-4 scheme). QED.

## 6.2. Matrix form of the $c$-$\tau$-4 scheme

The component equations of the $c$-$\tau$-4 scheme, i.e., Eqs. (5.14), (5.34), (6.1) and (6.2), can be cast into the matrix equation

$$\vec{q}(j,n) = \widehat{Q}_-(\tau_1, \tau_2, \nu)\, \vec{q}(j+1/2, n-1/2) + \widehat{Q}_+(\tau_1, \tau_2, \nu)\, \vec{q}(j-1/2, n-1/2) \qquad (j,n) \in \Omega \qquad (6.7)$$

where $\vec{q}(j,n)$ is defined in Eq. (5.54), and

$$\widehat{Q}_\pm(\tau_1, \tau_2, \nu) \overset{\text{def}}{=} \begin{pmatrix} \hat{q}_{11}^\pm & \hat{q}_{12}^\pm & \hat{q}_{13}^\pm & \hat{q}_{14}^\pm \\[2mm] \hat{q}_{21}^\pm & \hat{q}_{22}^\pm & \hat{q}_{23}^\pm & \hat{q}_{24}^\pm \\[2mm] 0 & 0 & \hat{q}_{33}^\pm & \hat{q}_{34}^\pm \\[2mm] 0 & 0 & \hat{q}_{43}^\pm & \hat{q}_{44}^\pm \end{pmatrix} \qquad (6.8)$$

with

$$\hat{q}_{11}^\pm \overset{\text{def}}{=} \frac{1 \pm \nu}{2}, \quad \hat{q}_{12}^\pm \overset{\text{def}}{=} \pm \frac{(1-\nu^2)}{2}, \quad \hat{q}_{13}^\pm \overset{\text{def}}{=} \mp \frac{\nu(1-\nu^2)}{3}, \quad \hat{q}_{14}^\pm \overset{\text{def}}{=} \mp \frac{(1-\nu^2)^2}{6} \qquad (6.9)$$

$$\hat{q}_{21}^\pm \overset{\text{def}}{=} \frac{\mp 1}{2(1+\tau_2)}, \quad \hat{q}_{22}^\pm \overset{\text{def}}{=} -\frac{(1-\tau_2 \mp 2\nu)}{2(1+\tau_2)}, \quad \hat{q}_{23}^\pm \overset{\text{def}}{=} \pm \frac{(1+\tau_2)^2}{12(1+\tau_1)} \mp \frac{(1-\tau_2 \mp 2\nu)^2}{4(1+\tau_2)}$$

$$\hat{q}_{24}^\pm \overset{\text{def}}{=} \frac{(1+\tau_2)^2(1-\tau_1 \mp 2\nu)}{12(1+\tau_1)} - \frac{(1-\tau_2 \mp 2\nu)^3}{12(1+\tau_2)} \qquad (6.10)$$

and

$$\hat{q}_{33}^\pm \overset{\text{def}}{=} \frac{1 \pm \nu}{2}, \quad \hat{q}_{34}^\pm \overset{\text{def}}{=} \pm \frac{(1-\nu^2)}{2}, \quad \hat{q}_{43}^\pm \overset{\text{def}}{=} \frac{\mp 1}{2(1+\tau_1)}, \quad \hat{q}_{44}^\pm \overset{\text{def}}{=} -\frac{(1-\tau_1 \mp 2\nu)}{2(1+\tau_1)} \qquad (6.11)$$

To further explore the relations between the $c$-$\tau$-4 and $c$-$\tau$ schemes, note that the component equations of the $c$-$\tau$ scheme, i.e., Eqs. (2.10) and (3.11), can be cast into the matrix form

$$\underline{\vec{q}}(j,n) = \widehat{M}_-(\tau, \nu)\, \underline{\vec{q}}(j+1/2, n-1/2) + \widehat{M}_+(\tau, \nu)\, \underline{\vec{q}}(j-1/2, n-1/2) \qquad (j,n) \in \Omega \qquad (6.12)$$

where $\underline{\vec{q}}(j,n)$ is defined in Eq. (5.61), and

$$\widehat{M}_\pm(\tau, \nu) \overset{\text{def}}{=} \frac{1}{2} \begin{pmatrix} 1 \pm \nu & \pm(1-\nu^2) \\[2mm] \dfrac{\mp 1}{1+\tau} & -\dfrac{1-\tau \mp 2\nu}{1+\tau} \end{pmatrix} \qquad -\infty < \tau < +\infty \qquad (6.13)$$

With the aid of Eqs. (6.9)–(6.11) and (6.13), it can be shown that

$$\begin{pmatrix} \hat{q}_{11}^\pm & \hat{q}_{12}^\pm \\[2mm] \hat{q}_{21}^\pm & \hat{q}_{22}^\pm \end{pmatrix} = \widehat{M}_\pm(\tau_2, \nu) \qquad (6.14)$$

and

$$\begin{pmatrix} \hat{q}_{33}^\pm & \hat{q}_{34}^\pm \\[2mm] \hat{q}_{43}^\pm & \hat{q}_{44}^\pm \end{pmatrix} = \widehat{M}_\pm(\tau_1, \nu) \qquad (6.15)$$

34

By comparing Eqs. (6.14) and (6.15) with Eq. (6.8), one now arrives at the conclusion that the $2 \times 2$ submatrices in the top-left and bottom right corners of the $4 \times 4$ matrix $\widehat{Q}_{\pm}(\tau_1, \tau_2, \nu)$ are $\widehat{M}_{\pm}(\tau_2, \nu)$ and $\widehat{M}_{\pm}(\tau_1, \nu)$, respectively.

### 6.3. Stability conditions of the $c$-$\tau$-4 scheme

By using similar arguments presented in Sec. 5.4, one concludes that the amplification matrices per marching step of the $c$-$\tau$-4 and $c$-$\tau$ schemes are

$$\widehat{Q}(\tau_1, \tau_2, \nu, \theta) = e^{i\theta/2}\widehat{Q}_-(\tau_1, \tau_2, \nu) + e^{-i\theta/2}\widehat{Q}_+(\tau_1, \tau_2, \nu)$$

$$= \begin{pmatrix} e^{i\theta/2}\hat{q}_{11}^- + e^{-i\theta/2}\hat{q}_{11}^+ & e^{i\theta/2}\hat{q}_{12}^- + e^{-i\theta/2}\hat{q}_{12}^+ & e^{i\theta/2}\hat{q}_{13}^- + e^{-i\theta/2}\hat{q}_{13}^+ & e^{i\theta/2}\hat{q}_{14}^- + e^{-i\theta/2}\hat{q}_{14}^+ \\ e^{i\theta/2}\hat{q}_{21}^- + e^{-i\theta/2}\hat{q}_{21}^+ & e^{i\theta/2}\hat{q}_{22}^- + e^{-i\theta/2}\hat{q}_{22}^+ & e^{i\theta/2}\hat{q}_{23}^- + e^{-i\theta/2}\hat{q}_{23}^+ & e^{i\theta/2}\hat{q}_{24}^- + e^{-i\theta/2}\hat{q}_{24}^+ \\ 0 & 0 & e^{i\theta/2}\hat{q}_{33}^- + e^{-i\theta/2}\hat{q}_{33}^+ & e^{i\theta/2}\hat{q}_{34}^- + e^{-i\theta/2}\hat{q}_{34}^+ \\ 0 & 0 & e^{i\theta/2}\hat{q}_{43}^- + e^{-i\theta/2}\hat{q}_{43}^+ & e^{i\theta/2}\hat{q}_{44}^- + e^{-i\theta/2}\hat{q}_{44}^+ \end{pmatrix} \tag{6.16}$$

and

$$\widehat{M}(\tau, \nu, \theta) \overset{\text{def}}{=} e^{i\theta/2}\widehat{M}_-(\tau, \nu) + e^{-i\theta/2}\widehat{M}_+(\tau, \nu) \equiv$$

$$= \begin{pmatrix} \cos(\theta/2) - i\nu\sin(\theta/2) & -i(1-\nu^2)\sin(\theta/2) \\ \dfrac{i\sin(\theta/2)}{1+\tau} & -\left[\dfrac{(1-\tau)\cos(\theta/2) + 2i\nu\sin(\theta/2)}{1+\tau}\right] \end{pmatrix} \tag{6.17}$$

respectively. Moreover, by using Eqs. (6.14)–(6.17), one concludes that

$$\begin{pmatrix} e^{i\theta/2}\hat{q}_{11}^- + e^{-i\theta/2}\hat{q}_{11}^+ & e^{i\theta/2}\hat{q}_{12}^- + e^{-i\theta/2}\hat{q}_{12}^+ \\ e^{i\theta/2}\hat{q}_{21}^- + e^{-i\theta/2}\hat{q}_{21}^+ & e^{i\theta/2}\hat{q}_{22}^- + e^{-i\theta/2}\hat{q}_{22}^+ \end{pmatrix} = \widehat{M}(\tau_2, \nu, \theta) \tag{6.18}$$

and

$$\begin{pmatrix} e^{i\theta/2}\hat{q}_{33}^- + e^{-i\theta/2}\hat{q}_{33}^+ & e^{i\theta/2}\hat{q}_{34}^- + e^{-i\theta/2}\hat{q}_{34}^+ \\ e^{i\theta/2}\hat{q}_{43}^- + e^{-i\theta/2}\hat{q}_{43}^+ & e^{i\theta/2}\hat{q}_{44}^- + e^{-i\theta/2}\hat{q}_{44}^+ \end{pmatrix} = \widehat{M}(\tau_1, \nu, \theta) \tag{6.19}$$

i.e., the $2 \times 2$ submatrices in the top-left and bottom-right corners of the matrix $\widehat{Q}(\tau_1, \tau_2, \nu, \theta)$ are $\widehat{M}(\tau_2, \nu, \theta)$ and $\widehat{M}(\tau_1, \nu, \theta)$, respectively. As such, *the amplification matrix of the $c$-$\tau$ scheme is identical to each of the above $2 \times 2$ submatrices of $\widehat{Q}(\tau_1, \tau_2, \nu, \theta)$ except that the parameter $\tau$ in the former matrix is replaced by $\tau_2$ and $\tau_1$ in the latter two submatrices.* By using this property and using a line of arguments similar to that presented in Sec. 5.4, one can show that, *for each set of $\tau_1$, $\tau_2$, $\nu$, and $\theta$, the eigenvalues of $\widehat{Q}(\tau_1, \tau_2, \nu, \theta)$ are the union of those of $\widehat{M}(\tau_1, \nu, \theta)$ and $\widehat{M}(\tau_2, \nu, \theta)$.* As a result, by using a line of arguments presented in the last part of Sec. 5.4, one can infer from the stability conditions Eqs. (3.13) of the $c$-$\tau$ scheme that

$$\nu^2 \le 1, \quad \tau_1 \ge \tau_o(\nu^2), \quad \tau_2 \ge \tau_o(\nu^2), \quad (\nu^2, \tau_1) \ne (1, 1), \quad \text{and} \quad (\nu^2, \tau_2) \ne (1, 1) \tag{6.20}$$

form a set of necessary stability conditions for the $c$-$\tau$-4 scheme.

### 6.4. The $c$-$\tau^*$-4 schemes

Let $h_1(s)$ and $h_2(s)$ be any two functions that meet the same requirements imposed on the function $h(s)$ introduced in Sec. 3.2. Then, because of the close relations between the $c$-$\tau$ and $c$-$\tau$-4 schemes established above, here a $c$-$\tau^*$-4 scheme is defined to be a special $c$-$\tau$-4 scheme with

$$\tau_1 = h_1(\nu^2) \quad \text{and} \quad \tau_2 = h_2(\nu^2) \qquad (|\nu| < 1) \tag{6.21}$$

35

According to Eq. (3.22), one can assume $h_1(s) = h_2(s) = \sqrt{s}$. As such a special $c$-$\tau^*$-4 scheme can be defined by the relation

$$\tau_1 = \tau_2 = |\nu| \qquad (|\nu| < 1) \tag{6.22}$$

For robust real-world applications, we may also introduce a special $c$-$\tau$-4 scheme which is defined by an extension of Eq. (3.24), i.e.,

$$\tau_1 = \beta_1 |\nu| \quad \text{and} \quad \tau_2 = \beta_2 |\nu| \qquad (\beta_1 \geq 1; \ \beta_2 \geq 1; \ |\nu| < 1) \tag{6.23}$$

where $\beta_1 \geq 1$ and $\beta_2 \geq 1$ are two adjustable parameters.

## 7. Fourth-order solvers for the inviscid Burgers equation

Consider the inviscid Burgers equation

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{7.1}$$

where

$$f(u) \overset{\text{def}}{=} \frac{u^2}{2} \tag{7.2}$$

Because, in $E_2$, Eq. (7.1) can be expressed as $\nabla \cdot \vec{h} = 0$ with

$$\vec{h} \overset{\text{def}}{=} (f, u) \tag{7.3}$$

Gauss' divergence theorem in the space-time $E_2$ implies that Eq. (7.1) is the differential form of the integral conservation law Eq. (2.2) if $\vec{h}$ is defined by Eq. (7.3). Moreover, for a smooth function $u(x,t)$, Eq. (2.2) implies the conservation condition Eq. (5.9) if

$$\vec{h}_{xx} \overset{\text{def}}{=} \frac{\partial^2 \vec{h}}{\partial x^2} \equiv \left( \frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 u}{\partial x^2} \right) \tag{7.4}$$

### 7.1. The inviscid Burgers $c$-$\tau$-4 scheme

In this section, the $c$-$\tau$-4 scheme will be extended to become a solver for the inviscid Burgers equation. Here again Eq. (5.1) is assumed. However, for any $(x,t) \in \text{SE}(j,n)$, the flux function $f(u)$ will be approximated by

$$f^*(x, t, j, n) \overset{\text{def}}{=} f_j^n + (f_x)_j^n (x - x_j) + (f_t)_j^n (t - t^n) + \frac{1}{2} (f_{xx})_j^n (x - x_j)^2$$
$$+ (f_{xt})_j^n (x - x_j)(t - t^n) + \frac{1}{2} (f_{tt})_j^n (t - t^n)^2 + \frac{1}{6} (f_{xxx})_j^n (x - x_j)^3 \qquad (x, t) \in \ \text{SE}(j,n) \tag{7.5}$$
$$+ \frac{1}{2} (f_{xxt})_j^n (x - x_j)^2 (t - t^n) + \frac{1}{2} (f_{xtt})_j^n (x - x_j)(t - t^n)^2 + \frac{1}{6} (f_{ttt})_j^n (t - t^n)^3$$

Here $f_j^n$, $(f_x)_j^n$, $(f_t)_j^n$, $(f_{xx})_j^n$, $(f_{xt})_j^n$, $(f_{tt})_j^n$ $(f_{xxx})_j^n$, $(f_{xxt})_j^n$, $(f_{xtt})_j^n$, and $(f_{ttt})_j^n$ are constants in $\text{SE}(j,n)$. They can be considered as the numerical analogues of the values of $f$, $\partial f/\partial x$, $\partial f/\partial t$, $\partial^2 f/\partial x^2$, $\partial^2 f/\partial x \partial t$, $\partial^2 f/\partial t^2$, $\partial^3 f/\partial x^3$, $\partial^3 f/\partial x^2 \partial t$, $\partial^3 f/\partial x \partial t^2$, and $\partial^3 f/\partial t^3$ at the mesh point $(j,n)$, respectively. Thus $f^*(x, t, j, n)$ represents a 3rd-order Taylor's approximation of $f$. Moreover, as will be shown, by imposing proper assumptions, the coefficients on the right sides of Eqs. (5.1) and (7.5) can be explicity expressed as the functions of the independent unknowns $u_j^n$, $(u_x)_j^n$, $(u_{xx})_j^n$ and $(u_{xxx})_j^n$.

For any $(j, n) \in \Omega$, let $u = u^*(x, t; j, n)$ and $f(u) = f^*(x, t, j, n)$ satisfy Eq. (7.1) uniformly in $\text{SE}(j, n)$, i.e.,

$$\frac{\partial u^*(x, t; j, n)}{\partial t} + \frac{\partial f^*(x, t, j, n)}{\partial x} \equiv 0 \qquad (j, n) \in \Omega \tag{7.6}$$

36

It can be shown that Eq. (7.6) $\Leftrightarrow$, for all $(j, n) \in \Omega$,

$$(u_t)_j^n = -(f_x)_j^n \tag{7.7}$$

$$(u_{xt})_j^n = -(f_{xx})_j^n \tag{7.8}$$

$$(u_{tt})_j^n = -(f_{xt})_j^n \tag{7.9}$$

$$(u_{xxt})_j^n = -(f_{xxx})_j^n \tag{7.10}$$

$$(u_{xtt})_j^n = -(f_{xxt})_j^n \tag{7.11}$$

and

$$(u_{ttt})_j^n = -(f_{xtt})_j^n \tag{7.12}$$

Moreover, by taking the numerical analogues of all the differential forms of Eq. (7.2) up to the third order, one has, for all $(j, n) \in \Omega$,

$$f_j^n = \frac{1}{2}(u^2)_j^n \tag{7.13}$$

$$(f_x)_j^n = (u u_x)_j^n \tag{7.14}$$

$$(f_t)_j^n = (u u_t)_j^n \tag{7.15}$$

$$(f_{xx})_j^n = \left[(u u_{xx} + (u_x)^2\right]_j^n \tag{7.16}$$

$$(f_{xt})_j^n = (u u_{xt} + u_x u_t)_j^n \tag{7.17}$$

$$(f_{tt})_j^n = \left[(u u_{tt} + (u_t)^2\right]_j^n \tag{7.18}$$

$$(f_{xxx})_j^n = (u u_{xxx} + 3 u_x u_{xx})_j^n \tag{7.19}$$

$$(f_{xxt})_j^n = (u u_{xxt} + u_t u_{xx} + 2 u_x u_{xt})_j^n \tag{7.20}$$

$$(f_{xtt})_j^n = (u u_{xtt} + u_x u_{tt} + 2 u_t u_{xt})_j^n \tag{7.21}$$

$$(f_{ttt})_j^n = (u u_{ttt} + 3 u_t u_{tt})_j^n \tag{7.22}$$

By substituting Eqs. (7.14), (7.16), and (7.19) into Eqs. (7.7), (7.8), and (7.10), respectively, one has

$$(u_t)_j^n = -(u u_x)_j^n \tag{7.23}$$

$$(u_{xt})_j^n = -\left[u u_{xx} + (u_x)^2\right]_j^n \tag{7.24}$$

and

$$(u_{xxt})_j^n = -(u u_{xxx} + 3 u_x u_{xx})_j^n \tag{7.25}$$

respectively. In turn, by substituting Eqs. (7.23)–(7.25) into Eqs. (7.15), (7.17), and (7.20), one has

$$(f_t)_j^n = -(u^2 u_x)_j^n \tag{7.26}$$

$$(f_{xt})_j^n = -\left[2u(u_x)^2 + u^2 u_{xx}\right]_j^n \tag{7.27}$$

and

$$(f_{xxt})_j^n = -\left[2(u_x)^3 + 6u u_x u_{xx} + u^2 u_{xxx}\right]_j^n \tag{7.28}$$

Moreover, by substituting Eqs. (7.27) and (7.28) into Eqs. (7.9) and (7.11), respectively, one has

$$(u_{tt})_j^n = \left[2u(u_x)^2 + u^2 u_{xx}\right]_j^n \tag{7.29}$$

37

and

$$(u_{xtt})_j^n = \left[ 2(u_x)^3 + 6uu_x u_{xx} + u^2 u_{xxx} \right]_j^n \tag{7.30}$$

respectively.

Next, by substituting Eqs. (7.23), (7.24), (7.29), and (7.30) into Eqs. (7.18) and (7.21), one obtains

$$(f_{tt})_j^n = \left[ 3u^2 (u_x)^2 + u^3 u_{xx} \right]_j^n \tag{7.31}$$

and

$$(f_{xtt})_j^n = \left[ 6u(u_x)^3 + 9u^2 u_x u_{xx} + u^3 u_{xxx} \right]_j^n \tag{7.32}$$

In turn, by substituting Eq. (7.32) into Eq. (7.12), one has

$$(u_{ttt})_j^n = - \left[ 6u(u_x)^3 + 9u^2 u_x u_{xx} + u^3 u_{xxx} \right]_j^n \tag{7.33}$$

Finally, by substituting Eqs. (7.23), (7.29), and (7.33) into Eq. (7.22), one has

$$(f_{ttt})_j^n = - \left[ 12u^2 (u_x)^3 + 12u^3 u_x u_{xx} + u^4 u_{xxx} \right]_j^n \tag{7.34}$$

According to the set of sixteen equations formed by Eqs. (7.13), (7.14), (7.16), (7.19), and (7.23)–(7.34), all the twenty coefficients on the right sides of Eqs. (5.1) and (7.5) other than $u_j^n$, $(u_x)_j^n$, $(u_{xx})_j^n$, and $(u_{xxx})_j^n$ are explicit functions of the latter coefficients. Thus there are only four independent mesh variables associated with each $(j, n) \in \Omega$. In fact, by substituting these sixteen equations into Eqs. (5.1) and (7.5), one has

$$\begin{aligned}
u^*(x, t; j, n) = {}& u_j^n + (u_x)_j^n \left\{ 1 - (u_x)_j^n (t - t^n) + \left[ (u_x)_j^n \right]^2 (t - t^n)^2 \right\} \left[ x - x_j - u_j^n (t - t^n) \right] \\
& + \frac{(u_{xx})_j^n}{2} \left[ 1 - 3(u_x)_j^n (t - t^n) \right] \left[ x - x_j - u_j^n (t - t^n) \right]^2 + \frac{(u_{xxx})_j^n}{6} \left[ x - x_j - u_j^n (t - t^n) \right]^3
\end{aligned} \tag{7.35}$$

and

$$\begin{aligned}
f^*(x, t, j, n) = {}& \frac{(u_j^n)^2}{2} + (u_x)_j^n \left\{ u_j^n + \frac{(u_x)_j^n}{2} \left[ x - x_j - 3u_j^n (t - t^n) \right] \right. \\
& \left. - \left[ (u_x)_j^n \right]^2 (t - t^n) \left[ x - x_j - 2u_j^n (t - t^n) \right] \right\} \left[ x - x_j - u_j^n (t - t^n) \right] \\
& + \frac{(u_{xx})_j^n}{2} \left\{ u_j^n + (u_x)_j^n \left[ x - x_j - 4u_j^n (t - t^n) \right] \right\} \left[ x - x_j - u_j^n (t - t^n) \right]^2 \\
& + \frac{u_j^n (u_{xxx})_j^n}{6} \left[ x - x_j - u_j^n (t - t^n) \right]^3
\end{aligned} \tag{7.36}$$

In turn, by substituting Eqs. (7.35) and (7.36) into Eq. (5.5) and

$$f_{xx}^*(x, t; j, n) \stackrel{\text{def}}{=} \frac{\partial^2 f^*(x, t, j, n)}{\partial x^2} \tag{7.37}$$

one has

$$u_{xx}^*(x, t; j, n) = (u_{xx})_j^n + (u_{xxx})_j^n (x - x_j) - (uu_{xxx} + 3u_x u_{xx})_j^n (t - t^n) \tag{7.38}$$

and

$$f_{xx}^*(x, t; j, n) = \left[ uu_{xx} + (u_x)^2 \right]_j^n + (uu_{xxx} + 3u_x u_{xx})_j^n (x - x_j) - \left[ 2(u_x)^3 + 6uu_x u_{xx} + u^2 u_{xxx} \right]_j^n (t - t^n) \tag{7.39}$$

38

Note that, for the same reason presented immediately following Eq. (2.6), *except for flux evaluation,*
*Eqs. (7.35) and (7.38) are applicable even for* $(x,t) \notin SE(j,n)$.

Next let the numerical analogues of Eqs. (7.3) and (7.4) be

$$\vec{h}^*(x,t\,;j,n) \stackrel{\text{def}}{=} (f^*(x,t,j,n),\, u^*(x,t\,;j,n)) \qquad (x,t) \in \ SE(j,n) \qquad (7.40)$$

and

$$\vec{h}^*_{xx}(x,t;j,n) \stackrel{\text{def}}{=} (f^*_{xx}(x,t;j,n),\, u^*_{xx}(x,t;j,n)) \qquad (x,t) \in \ SE(j,n) \qquad (7.41)$$

respectively. Also, let

$$(u_{\bar{x}})^n_j \stackrel{\text{def}}{=} \frac{\Delta x}{4}(u_x)^n_j, \quad (u_{\bar{x}\bar{x}})^n_j \stackrel{\text{def}}{=} \left(\frac{\Delta x}{4}\right)^2 (u_{xx})^n_j, \quad \text{and} \quad (u_{\bar{x}\bar{x}\bar{x}})^n_j \stackrel{\text{def}}{=} \left(\frac{\Delta x}{4}\right)^3 (u_{xxx})^n_j \quad (j,n) \in \Omega \qquad (7.42)$$

and

$$\nu^n_j \stackrel{\text{def}}{=} \frac{u^n_j \Delta t}{\Delta x}, \quad (\nu_{\bar{x}})^n_j \stackrel{\text{def}}{=} \frac{(u_{\bar{x}})^n_j \Delta t}{\Delta x}, \quad (\nu_{\bar{x}\bar{x}})^n_j \stackrel{\text{def}}{=} \frac{(u_{\bar{x}\bar{x}})^n_j \Delta t}{\Delta x}, \quad \text{and} \quad (\nu_{\bar{x}\bar{x}\bar{x}})^n_j \stackrel{\text{def}}{=} \frac{(u_{\bar{x}\bar{x}\bar{x}})^n_j \Delta t}{\Delta x} \quad (j,n) \in \Omega \quad (7.43)$$

Obviously, $(u_{\bar{x}})^n_j$, $(u_{\bar{x}\bar{x}})^n_j$, and $(u_{\bar{x}\bar{x}\bar{x}})^n_j$ defined here are identical to those defined in Eqs. (2.9) and (5.17). However, the "Courant number" $\nu^n_j$ and its "derivatives" $(\nu_{\bar{x}})^n_j$, $(\nu_{\bar{x}\bar{x}})^n_j$, and $(\nu_{\bar{x}\bar{x}\bar{x}})^n_j$ are mesh-point dependent in the current development.

Like the $c$-$\tau$-4 scheme, the current extension will be construct to enforce the least stringent conservation conditions Eqs. (2.13) and (5.18). In fact, by using Eqs. (7.38), (7.39), (7.41), (7.42), and (7.43), it can be shown that Eq. (5.18) $\Leftrightarrow$

$$
\begin{aligned}
(u_{\bar{x}\bar{x}})^n_j = \frac{1}{2}\Big\{ &\big[\nu_{\bar{x}}(2\nu_{\bar{x}}-1)u_{\bar{x}} + [1-\nu(1-6\nu_{\bar{x}})]u_{\bar{x}\bar{x}} - (1-\nu^2)u_{\bar{x}\bar{x}\bar{x}}\big]^{n-1/2}_{j+1/2} \\
&+ \big[\nu_{\bar{x}}(1-2\nu_{\bar{x}})u_{\bar{x}} + [1+\nu(1-6\nu_{\bar{x}})]u_{\bar{x}\bar{x}} + (1-\nu^2)u_{\bar{x}\bar{x}\bar{x}}\big]^{n-1/2}_{j-1/2} \Big\}
\end{aligned}
\qquad (7.44)
$$

Moreover, by using Eqs. (7.35), (7.36), (7.40), (7.42), and (7.43), it can be shown that Eq. (2.13) $\Leftrightarrow$

$$
\begin{aligned}
2\Big[&u + \frac{2u_{\bar{x}\bar{x}}}{3}\Big]^n_j \\
&= \Big\{ \Big(1-\frac{\nu}{2}\Big)u - \big[1-\nu^2+2\nu_{\bar{x}}\nu^2-4(\nu_{\bar{x}})^2\nu^2\big]u_{\bar{x}} + \Big[\frac{2(1-\nu^3)}{3}+4\nu_{\bar{x}}\nu^3\Big]u_{\bar{x}\bar{x}} - \frac{(1-\nu^4)}{3}u_{\bar{x}\bar{x}\bar{x}}\Big\}^{n-1/2}_{j+1/2} \\
&\quad + \Big\{ \Big(1+\frac{\nu}{2}\Big)u + \big[1-\nu^2+2\nu_{\bar{x}}\nu^2-4(\nu_{\bar{x}})^2\nu^2\big]u_{\bar{x}} + \Big[\frac{2(1+\nu^3)}{3}-4\nu_{\bar{x}}\nu^3\Big]u_{\bar{x}\bar{x}} + \frac{(1-\nu^4)}{3}u_{\bar{x}\bar{x}\bar{x}}\Big\}^{n-1/2}_{j-1/2}
\end{aligned}
\qquad (7.45)
$$

Expressing the mesh variable $(u_{\bar{x}\bar{x}})^n_j$ on the left side of Eq. (7.45) in terms of the mesh variables at the $(n-1/2)$th time level by using Eq. (7.44), one obtains

$$
\begin{aligned}
u^n_j = \Big\{ &\frac{(2-\nu)}{4}u - \frac{[3(1-\nu^2)+2(3\nu^2-1)(1-2\nu_{\bar{x}})\nu_{\bar{x}}]}{6}u_{\bar{x}} + \frac{\nu(1-\nu^2)(1-6\nu_{\bar{x}})}{3}u_{\bar{x}\bar{x}} + \frac{(1-\nu^2)^2}{6}u_{\bar{x}\bar{x}\bar{x}} \Big\}^{n-1/2}_{j+1/2} \\
&+ \Big\{ \frac{(2+\nu)}{4}u + \frac{[3(1-\nu^2)+2(3\nu^2-1)(1-2\nu_{\bar{x}})\nu_{\bar{x}}]}{6}u_{\bar{x}} - \frac{\nu(1-\nu^2)(1-6\nu_{\bar{x}})}{3}u_{\bar{x}\bar{x}} - \frac{(1-\nu^2)^2}{6}u_{\bar{x}\bar{x}\bar{x}} \Big\}^{n-1/2}_{j-1/2}
\end{aligned}
$$
$$(7.46)$$

It has been shown that, by imposing the conservation conditions Eqs. (2.13) and (5.18), one can obtain Eqs. (7.44) and (7.46) by which both $u_j^n$ and $(u_{\bar{x}\bar{x}})_j^n$ can be explicitly evaluated in terms of the mesh variables at the $(n-1/2)$th time level. In the following, by using techniques developed earlier, similar explicit expressions will be developed for $(u_{\bar{x}})_j^n$ and $(u_{\bar{x}\bar{x}\bar{x}})_j^n$.

To proceed, let $P^+$ and $P^-$ be the points depicted in Fig. 3, *except that the parameter $\tau$ shown in the figure be replaced by the mesh-point dependent parameter* $(\tau_1)_j^n$. Then the x-coordinates of $P^+$ and $P^-$ are

$$x(P^+) = x_j + \left[1 + (\tau_1)_j^n\right]\frac{\Delta x}{4} = x_{j+1/2} - \left[1 - (\tau_1)_j^n\right]\frac{\Delta x}{4}, \qquad -\infty < (\tau_1)_j^n < +\infty \qquad (7.47)$$

and

$$x(P^-) = x_j - \left[1 + (\tau_1)_j^n\right]\frac{\Delta x}{4} = x_{j-1/2} + \left[1 - (\tau_1)_j^n\right]\frac{\Delta x}{4}, \qquad -\infty < (\tau_1)_j^n < +\infty \qquad (7.48)$$

respectively. In the current development, we assume

$$(u_{xxx})_j^n = \frac{u_{xx}(P^+) - u_{xx}(P^-)}{\left[1 + (\tau_1)_j^n\right]\Delta x/2} \qquad (7.49)$$

where

$$u_{xx}(P^+) \overset{\text{def}}{=} u_{xx}^*\left(x(P^+), t^n; j+1/2, n-1/2\right) \qquad (7.50)$$

and

$$u_{xx}(P^-) \overset{\text{def}}{=} u_{xx}^*\left(x(P^-), t^n; j-1/2, n-1/2\right) \qquad (7.51)$$

Because $|\overline{P^-P^+}| = \left[1 + (\tau_1)_j^n\right]\Delta x/2$, the expression on the right side of Eq. (7.49) represents a central-difference approximation of $(u_{xxx})_j^n$. With the aid of Eqs. (7.38), (7.42), (7.43), (7.47), (7.48), (7.50), and (7.51), it can be shown that Eq. (7.49) $\Leftrightarrow$

$$(u_{\bar{x}\bar{x}\bar{x}})_j^n = \frac{\left[(1-6\nu_{\bar{x}})u_{\bar{x}\bar{x}} - (1-\tau_1+2\nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j+1/2}^{n-1/2} - \left[(1-6\nu_{\bar{x}})u_{\bar{x}\bar{x}} + (1-\tau_1-2\nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2}}{2\left[1 + (\tau_1)_j^n\right]} \qquad (7.52)$$

To simplify notation, in Eq. (7.52) and hereafter in this section we adopt a convention that can be explained using two expression on the right side of Eq. (7.52) as examples, i.e.,

$$\left[(1-6\nu_{\bar{x}})u_{\bar{x}\bar{x}} - (1-\tau_1+2\nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j+1/2}^{n-1/2} = \left[(1-6\nu_{\bar{x}})u_{\bar{x}\bar{x}}\right]_{j+1/2}^{n-1/2} - \left[1 - (\tau_1)_j^n + 2\nu_{j+1/2}^{n-1/2}\right](u_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2}$$

and

$$\left[(1-6\nu_{\bar{x}})u_{\bar{x}\bar{x}} + (1-\tau_1-2\nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2} = \left[(1-6\nu_{\bar{x}})u_{\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2} + \left[1 - (\tau_1)_j^n - 2\nu_{j-1/2}^{n-1/2}\right](u_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2}$$

In other words, **the parameter $\tau_1$ which appears on the right side of an equation such as Eq. (7.52) and the mesh variable which appears on the left side are associated with the same mesh point** $(j, n)$.

Next, let $P^+$ and $P^-$ be the points depicted in Fig. 3, *except that the parameter $\tau$ shown in the figure be replaced by the mesh-point dependent parameter* $(\tau_2)_j^n$. Then the x-coordinates of $P^+$ and $P^-$ are

$$x(P^+) = x_j + \left[1 + (\tau_2)_j^n\right]\frac{\Delta x}{4} = x_{j+1/2} - \left[1 - (\tau_2)_j^n\right]\frac{\Delta x}{4}, \qquad -\infty < (\tau_2)_j^n < +\infty \qquad (7.53)$$

and

$$x(P^-) = x_j - \left[1 + (\tau_2)_j^n\right]\frac{\Delta x}{4} = x_{j-1/2} + \left[1 - (\tau_2)_j^n\right]\frac{\Delta x}{4}, \qquad -\infty < (\tau_2)_j^n < +\infty \qquad (7.54)$$

40

respectively. In the current development, $(u_x)_j^n$ is approximated with the current version of the expression on the right side of Eq. (5.46), i.e.,

$$(u_x)_j^n = \frac{u^{4*}(P^+) - u^{4*}(P^-)}{\left[1 + (\tau_2)_j^n\right]\Delta x/2} - \frac{\left\{\left[1 + (\tau_2)_j^n\right]\Delta x\right\}^2}{96}(u_{xxx})_j^n \tag{7.55}$$

where

$$u^{4*}(P^\pm) \stackrel{\text{def}}{=} u^* \left(x(P^\pm), t^n; j \pm 1/2, n - 1/2\right) \tag{7.56}$$

According to Eq. (7.35), $u^{4*}(P^+)$ is the 3rd-order Taylor's approximation (i.e., the error is $O\left[(\Delta x)^4\right]$) of $u$ at $P^+$ evaluated using the mesh variables at $(j + 1/2, n - 1/2)$ while $u^{4*}(P^-)$ is that at $P^-$ evaluated using the mesh variables at $(j - 1/2, n - 1/2)$. As such, Eq. (5.43) implies that Eq. (7.55) is an approximation of 3rd-order accuracy for the value of $\partial u/\partial x$ at the mesh point $(j, n)$.

Next, with the aid Eqs. (7.35), (7.42), (7.43), (7.53), and (7.54), it can be shown that

$$u^{4*}(P^\pm) = \Big\{u \mp (1 - \tau_2 \pm 2\nu)[1 - 2\nu_{\bar{x}} + 4(\nu_{\bar{x}})^2]u_{\bar{x}} + \frac{1}{2}(1 - \tau_2 \pm 2\nu)^2(1 - 6\nu_{\bar{x}})u_{\bar{x}\bar{x}}$$
$$\mp \frac{1}{6}(1 - \tau_2 \pm 2\nu)^3 u_{\bar{x}\bar{x}\bar{x}}\Big\}_{j\pm 1/2}^{n-1/2} \tag{7.57}$$

In turn, by using Eqs. (7.42),(7.52) and (7.57), it can be shown that Eq. (7.55) $\Leftrightarrow$

$$(u_{\bar{x}})_j^n = \Big\{\frac{u}{2(1 + \tau_2)} - \frac{(1 - \tau_2 + 2\nu)\left[1 - 2\nu_{\bar{x}} + 4(\nu_{\bar{x}})^2\right]u_{\bar{x}}}{2(1 + \tau_2)} + (1 - 6\nu_{\bar{x}})\left[\frac{(1 - \tau_2 + 2\nu)^2}{4(1 + \tau_2)} - \frac{(1 + \tau_2)^2}{12(1 + \tau_1)}\right]u_{\bar{x}\bar{x}}$$
$$+ \left[\frac{(1 + \tau_2)^2(1 - \tau_1 + 2\nu)}{12(1 + \tau_1)} - \frac{(1 - \tau_2 + 2\nu)^3}{12(1 + \tau_2)}\right]u_{\bar{x}\bar{x}\bar{x}}\Big\}_{j+1/2}^{n-1/2}$$
$$- \Big\{\frac{u}{2(1 + \tau_2)} + \frac{(1 - \tau_2 - 2\nu)\left[1 - 2\nu_{\bar{x}} + 4(\nu_{\bar{x}})^2\right]u_{\bar{x}}}{2(1 + \tau_2)} + (1 - 6\nu_{\bar{x}})\left[\frac{(1 - \tau_2 - 2\nu)^2}{4(1 + \tau_2)} - \frac{(1 + \tau_2)^2}{12(1 + \tau_1)}\right]u_{\bar{x}\bar{x}}$$
$$- \left[\frac{(1 + \tau_2)^2(1 - \tau_1 - 2\nu)}{12(1 + \tau_1)} - \frac{(1 - \tau_2 - 2\nu)^3}{12(1 + \tau_2)}\right]u_{\bar{x}\bar{x}\bar{x}}\Big\}_{j-1/2}^{n-1/2} \tag{7.58}$$

**Note that the simplifying convention established earlier for the parameter $\tau_1$ in Eq. (7.52) applies here for both $\tau_1$ and $\tau_2$, i.e., everywhere on the right sides of Eqs. (7.57) and (7.58), the symbols $\tau_1$ and $\tau_2$ are the shorthands for $(\tau_1)_j^n$ and $(\tau_2)_j^n$, respectively.**

By definition, the Burgers $c$-$\tau$-4 scheme is formed by Eqs. (7.44), (7.46), (7.52), and (7.58). It enforces the local conservation relations Eqs. (2.13) and (5.18). According to an argument given in Comment (d) presented at the end of Sec. 2.1, it will also enforce a global conservation relation for each of the fluxes of $\vec{h}^*$ and $\vec{h}_{xx}^*$ over any space-time region that is the union of any combination of CE$(j, n)$, $(j, n) \in \Omega$. Moreover, even though it is a solver for a nonlinear equation Eq. (7.1), *by explicitly specifying $(\tau_1)_j^n$ and $(\tau_2)_j^n$ as functions of the mesh variables at $(j - 1/2, n - 1/2)$ and $(j + 1/2, n - 1/2)$, the scheme will become completely explicit*, i.e., the four independent mesh variables at each $(j, n)$ can be evaluated in terms of those at the $(n - 1/2)$th time level without resorting to solving a system of nonlinear equations involving the independent mesh variables at $(j, n)$. Also note that, because $\nu_j^n = (\nu_{\bar{x}})_j^n = 0$ when $\Delta t = 0$, Eqs. (7.52) and (7.58) reduce to Eqs. (6.3) and (6.4) respectively when $\Delta t = \tau_1 = \tau_2 = 0$ In fact, it can be shown that the Burgers $c$-$\tau$-4 scheme satisfies Eqs. (2.7), (2.8), (5.12), and (5.13) (i.e., the conservation relations over each CE$_\pm(j, n)$) when $\Delta t = \tau_1 = \tau_2 = 0$.

To turn the current scheme into an explicit scheme, as an example, we can assume that

$$(\tau_1)_j^n = \beta_1|\nu_j^n| \quad \text{and} \quad (\tau_2)_j^n = \beta_2|\nu_j^n| \qquad (\beta_1 \geq 1;\ \beta_2 \geq 1) \qquad (7.59)$$

where $\beta_1 \geq 1$ and $\beta_2 \geq 1$ are two adjustable constant parameters. According to Eqs. (7.43) and (7.46), Eq. (7.59) effectively defines $(\tau_1)_j^n$ and $(\tau_2)_j^n$ in terms of the mesh variables at $(j + 1/2, n - 1/2)$ and $(j - 1/2, n - 1/2)$. As another example, one can assume

$$(\tau_1)_j^n = \beta_1 \max\left\{|\nu_{j-1/2}^{n-1/2}|, |\nu_{j+1/2}^{n-1/2}|\right\} \text{ and } (\tau_2)_j^n = \beta_2 \max\left\{|\nu_{j-1/2}^{n-1/2}|, |\nu_{j+1/2}^{n-1/2}|\right\} \quad (\beta_1 \geq 1;\ \beta_2 \geq 1) \quad (7.60)$$

where $\beta_1 \geq 1$ and $\beta_2 \geq 1$ are again two adjustable constant parameters. Other possible choices of $(\tau_1)_j^n$ and $(\tau_2)_j^n$ are specified in Eqs. (7.108)–(7.110).

In the following, it will be shown that the stability conditions of the Burgers $c$-$\tau$-4 scheme are essentially similar to those of the $c$-$\tau$-4 scheme.

### 7.2. Stability conditions of the inviscid Burgers $c$-$\tau$-4 scheme

As explained in Sec. 5.4, for linear solvers such as the $a$-$\epsilon$-4 or $c$-$\tau$-4 scheme, propagation of round-off errors is governed by the same linear system of equations governing the associated independent mesh variables. As will be shown, the above simple relation is not valid for a nonlinear solver such as the Burgers $c$-$\tau$-4 scheme.

To proceed, note that Eqs. (7.1) and (7.2) imply that the dependent variable $u$ has the dimension of $\Delta t/\Delta x$. Thus $\nu_j^n$, $(\nu_{\bar{x}})_j^n$, $(\nu_{\bar{x}\bar{x}})_j^n$, and $(\nu_{\bar{x}\bar{x}\bar{x}})_j^n$ defined in Eq. (7.43) are all dimensionless. As a result, by multiplying both sides of each of Eqs. (7.44), (7.46), (7.52), and (7.58) with the factor $\Delta t/\Delta x$, one can turn the latter equations into their dimensionless versions. The dimensionless versions can be obtained from the original versions by replacing the symbols $u$, $u_{\bar{x}}$, $u_{\bar{x}\bar{x}}$ and $u_{\bar{x}\bar{x}\bar{x}}$ with $\nu$, $\nu_{\bar{x}}$, $\nu_{\bar{x}\bar{x}}$, and $\nu_{\bar{x}\bar{x}\bar{x}}$, respectively. Explicitly, the dimensionless versions of Eqs. (7.46), (7.58), (7.44), and (7.52) are

$$\nu_j^n = \left\{\frac{(2-\nu)}{4}\nu - \frac{\left[3(1-\nu^2)+2(3\nu^2-1)(1-2\nu_{\bar{x}})\nu_{\bar{x}}\right]}{6}\nu_{\bar{x}} + \frac{\nu(1-\nu^2)(1-6\nu_{\bar{x}})}{3}\nu_{\bar{x}\bar{x}} + \frac{(1-\nu^2)^2}{6}\nu_{\bar{x}\bar{x}\bar{x}}\right\}_{j+1/2}^{n-1/2}$$
$$+ \left\{\frac{(2+\nu)}{4}\nu + \frac{\left[3(1-\nu^2)+2(3\nu^2-1)(1-2\nu_{\bar{x}})\nu_{\bar{x}}\right]}{6}\nu_{\bar{x}} - \frac{\nu(1-\nu^2)(1-6\nu_{\bar{x}})}{3}\nu_{\bar{x}\bar{x}} - \frac{(1-\nu^2)^2}{6}\nu_{\bar{x}\bar{x}\bar{x}}\right\}_{j-1/2}^{n-1/2}$$

$$(7.61)$$

$$(\nu_{\bar{x}})_j^n = \left\{\frac{\nu}{2(1+\tau_2)} - \frac{(1-\tau_2+2\nu)\left[1-2\nu_{\bar{x}}+4(\nu_{\bar{x}})^2\right]\nu_{\bar{x}}}{2(1+\tau_2)} + (1-6\nu_{\bar{x}})\left[\frac{(1-\tau_2+2\nu)^2}{4(1+\tau_2)} - \frac{(1+\tau_2)^2}{12(1+\tau_1)}\right]\nu_{\bar{x}\bar{x}}\right.$$
$$\left. + \left[\frac{(1+\tau_2)^2(1-\tau_1+2\nu)}{12(1+\tau_1)} - \frac{(1-\tau_2+2\nu)^3}{12(1+\tau_2)}\right]\nu_{\bar{x}\bar{x}\bar{x}}\right\}_{j+1/2}^{n-1/2}$$
$$- \left\{\frac{\nu}{2(1+\tau_2)} + \frac{(1-\tau_2-2\nu)\left[1-2\nu_{\bar{x}}+4(\nu_{\bar{x}})^2\right]\nu_{\bar{x}}}{2(1+\tau_2)} + (1-6\nu_{\bar{x}})\left[\frac{(1-\tau_2-2\nu)^2}{4(1+\tau_2)} - \frac{(1+\tau_2)^2}{12(1+\tau_1)}\right]\nu_{\bar{x}\bar{x}}\right.$$
$$\left. - \left[\frac{(1+\tau_2)^2(1-\tau_1-2\nu)}{12(1+\tau_1)} - \frac{(1-\tau_2-2\nu)^3}{12(1+\tau_2)}\right]\nu_{\bar{x}\bar{x}\bar{x}}\right\}_{j-1/2}^{n-1/2}$$

$$(7.62)$$

$$(\nu_{\bar{x}\bar{x}})_j^n = \frac{1}{2}\left\{\left[(2\nu_{\bar{x}}-1)(\nu_{\bar{x}})^2 + [1-\nu(1-6\nu_{\bar{x}})]\nu_{\bar{x}\bar{x}} - (1-\nu^2)\nu_{\bar{x}\bar{x}\bar{x}}\right]_{j+1/2}^{n-1/2}\right.$$
$$\left. + \left[(1-2\nu_{\bar{x}})(\nu_{\bar{x}})^2 + [1+\nu(1-6\nu_{\bar{x}})]\nu_{\bar{x}\bar{x}} + (1-\nu^2)\nu_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2}\right\}$$

$$(7.63)$$

and

$$(\nu_{\bar{x}\bar{x}\bar{x}})_j^n = \frac{\left[(1 - 6\nu_{\bar{x}})\nu_{\bar{x}\bar{x}} - (1 - \tau_1 + 2\nu)\nu_{\bar{x}\bar{x}\bar{x}}\right]_{j+1/2}^{n-1/2} - \left[(1 - 6\nu_{\bar{x}})\nu_{\bar{x}\bar{x}} + (1 - \tau_1 - 2\nu)\nu_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2}}{2\left[1 + (\tau_1)_j^n\right]} \tag{7.64}$$

respectively.

In the current study, we assume that

$$|(u_{\bar{x}\bar{x}\bar{x}})_j^n| \ll |(u_{\bar{x}\bar{x}})_j^n| \ll |(u_{\bar{x}})_j^n| \ll |u_j^n| \qquad (j, n) \in \Omega \tag{7.65}$$

$$|\nu_j^n| \le 1 \qquad (j, n) \in \Omega \tag{7.66}$$

and

$$|(\tau_1)_j^n| = O(|\nu_j^n|) \quad \text{and} \quad |(\tau_2)_j^n| = O(|\nu_j^n|), \qquad (j, n) \in \Omega \tag{7.67}$$

Note that, by using Eqs. (7.35) and (7.42), one has

$$u^*(x_j \pm \frac{\Delta x}{4}, t^n; j, n) = u_j^n \pm (u_{\bar{x}})_j^n + \frac{1}{2}(u_{\bar{x}\bar{x}})_j^n \pm \frac{1}{6}(u_{\bar{x}\bar{x}\bar{x}})_j^n \qquad (j, n) \in \Omega \tag{7.68}$$

Thus, Eq. (7.65) simply states that the magnitude of a term on the right side of Eq. (7.68) is much larger than that of the next higher-order term to its right. In view of Eqs. (7.42), obviously, this assumption can be made valid by choosing a small enough value of $\Delta x$. Moreover, by using Eqs. (7.43), it can be shown that Eqs. (7.65) and (7.66) $\Leftrightarrow$

$$|(\nu_{\bar{x}\bar{x}\bar{x}})_j^n| \ll |(\nu_{\bar{x}\bar{x}})_j^n| \ll |(\nu_{\bar{x}})_j^n| \ll |\nu_j^n| \le 1 \qquad (j, n) \in \Omega \tag{7.69}$$

With the aid of Eqs. (7.42) and (7.43), it is seen that the assumption $|(\nu_{\bar{x}})_j^n| \ll |\nu_j^n|$ in Eq. (7.69) $\Leftrightarrow$

$$\Delta x (u_x)_j^n \ll u_j^n \qquad (j, n) \in \Omega \tag{7.70}$$

*Within a smooth solution region*, $(u_x)_j^n$ is a good approximation of the value of $\partial u / \partial x$ at $(j, n)$. Thus Eq. (7.70) implies that

$$|u_{j+1/2}^{n-1/2} - u_{j-1/2}^{n-1/2}| \ll |u_{j\pm1/2}^{n-1/2}| \qquad (j, n) \in \Omega \tag{7.71}$$

By using Eq. (7.43), one can cast Eq. (7.71) into its dimensionless form, i.e.,

$$|\nu_{j+1/2}^{n-1/2} - \nu_{j-1/2}^{n-1/2}| \ll |\nu_{j\pm1/2}^{n-1/2}| \qquad (j, n) \in \Omega \tag{7.72}$$

In turn, Eq. (7.72) implies that

$$|\nu_{j\pm1/2}^{n-1/2} - \tilde{\nu}_j^n| \ll |\tilde{\nu}_j^n| \qquad (j, n) \in \Omega \tag{7.73}$$

with

$$\tilde{\nu}_j^n \stackrel{\text{def}}{=} \frac{1}{2}\left(\nu_{j+1/2}^{n-1/2} + \nu_{j-1/2}^{n-1/2}\right) \qquad (j, n) \in \Omega \tag{7.74}$$

Eq. (7.73) implies that

$$\nu_{j\pm1/2}^{n-1/2} \doteq \tilde{\nu}_j^n \qquad (j, n) \in \Omega \tag{7.75}$$

Hereafter the symbol "$\doteq$" denote the fact that the magnitude of the difference of the quantities on the both sides of the symbol is much less than the magnitude of either quantity itself. In other words, either is a close approximation of another.

The stability study will be carried out using Eqs. (7.61)–(7.64). In these dimensionless equations, the independent mesh variables at any mesh point $(j, n) \in \Omega$ are $\nu_j^n$, $(\nu_{\bar{x}})_j^n$, $(\nu_{\bar{x}\bar{x}})_j^n$, and $(\nu_{\bar{x}\bar{x}\bar{x}})_j^n$. Hereafter the round-off errors associated with these variables are denoted by $\delta\nu_j^n$, $\delta(\nu_{\bar{x}})_j^n$, $\delta(\nu_{\bar{x}\bar{x}})_j^n$, and $\delta(\nu_{\bar{x}\bar{x}\bar{x}})_j^n$, respectively. **Because the magnitude of the round-off error of a quantity with a smaller magnitude generally is not necessary smaller than that of a quantity with a larger magnitude**, in contrast to Eq. (7.69), we assume that

$$|\delta(\tau_1)_j^n| \approx |\delta(\tau_2)_j^n| \approx |\delta\nu_j^n| \approx |\delta(\nu_{\bar{x}})_j^n| \approx |\delta(\nu_{\bar{x}\bar{x}})_j^n| \approx |\delta(\nu_{\bar{x}\bar{x}\bar{x}})_j^n| \qquad (j, n) \in \Omega \qquad (7.76)$$

Hereafter (i) the symbol "$\approx$" implies that the quantities on both sides of this symbol have the same order of magnitude, and (ii) $\delta(\tau_1)_j^n$ and $\delta(\tau_2)_j^n$ denote the round-off errors of $(\tau_1)_j^n$ and $(\tau_2)_j^n$, respectively. Also **because round-off errors could vary erratically from one mesh point to another**, in contrast to Eq. (7.72), in general, we have

$$|\delta\nu_{j+1/2}^{n-1/2} - \delta\nu_{j-1/2}^{n-1/2}| \approx |\delta\nu_{j\pm1/2}^{n-1/2}| \approx |\delta(\tau_1)_j^n| \approx |\delta(\tau_2)_j^n| \qquad (j, n) \in \Omega \qquad (7.77)$$

Let $x_\ell$, $\ell = 1, 2, 3, 4$, be independent mesh variables at a mesh point. Let $\delta x_\ell$, $\ell = 1, 2, 3, 4$ be their round-off errors, respectively. Then because each $\delta x_\ell$ is a very small perturbation of $x_\ell$, the round-off errors of a differentiable function $f(x_1, x_2, x_3, x_4)$ can be approximated by

$$\delta f(x_1, x_2, x_3, x_4) \overset{\text{def}}{=} \sum_{\ell=1}^{4} \frac{\partial f}{\partial x_\ell} \delta x_\ell \qquad (7.78)$$

As an example, let

$$x_1 \overset{\text{def}}{=} \nu_{j+1/2}^{n-1/2}, \quad x_2 \overset{\text{def}}{=} (\nu_{\bar{x}})_{j+1/2}^{n-1/2}, \quad x_3 \overset{\text{def}}{=} (\nu_{\bar{x}\bar{x}})_{j+1/2}^{n-1/2}, \quad \text{and} \quad x_4 \overset{\text{def}}{=} (\nu_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2} \qquad (j, n) \in \Omega \qquad (7.79)$$

Then the expression on the right side of Eq. (7.61) associated with the mesh point $(j + 1/2, n - 1/2)$ can be expressed as $g(x_1, x_2, x_3, x_4)$ where

$$g(x_1, x_2, x_3, x_4) \overset{\text{def}}{=} \frac{1}{2}\left(1 - \frac{x_1}{2}\right)x_1 - \frac{1}{2}\left[1 - (x_1)^2\right]x_2 - \frac{1}{3}\left[3(x_1)^2 - 1\right](1 - 2x_2)(x_2)^2$$
$$+ \frac{x_1}{3}\left[1 - (x_1)^2\right](1 - 6x_2)x_3 + \frac{1}{6}\left[1 - (x_1)^2\right]^2 x_4 \qquad (7.80)$$

Eq. (7.80) implies that

$$\frac{\partial g}{\partial x_1} = \frac{1 - x_1}{2} + x_1 x_2 - 2x_1(1 - 2x_2)(x_2)^2 + \frac{\left[1 - 3(x_1)^2\right]}{3}(1 - 6x_2)x_3 - \frac{2x_1 x_4}{3}\left[1 - (x_1)^2\right] \qquad (7.81)$$

$$\frac{\partial g}{\partial x_2} = -\frac{\left[1 - (x_1)^2\right]}{2} - \frac{2}{3}\left[3(x_1)^2 - 1\right]x_2(1 - 3x_2) - 2x_1 x_3\left[1 - (x_1)^2\right] \qquad (7.82)$$

$$\frac{\partial g}{\partial x_3} = \frac{x_1\left[1 - (x_1)^2\right]}{3} - 2x_1 x_2\left[1 - (x_1)^2\right] \qquad (7.83)$$

and

$$\frac{\partial g}{\partial x_4} = \frac{\left[1 - (x_1)^2\right]^2}{6} \qquad (7.84)$$

Note that Eq. (7.69) implies that

$$|x_4| \ll |x_3| \ll |x_2| \ll |x_1| \leq 1 \qquad (7.85)$$

44

As such, by excluding the case

$$|1 - (x_1)^2| \ll 1 \tag{7.86}$$

one concludes that the expression on the right side of each of Eqs. (7.81)–(7.83) is dominated by the first term which involves only the variable $x_1$. Moreover, because Eq. (7.76) implies that

$$|\delta x_4| \approx |\delta x_3| \approx |\delta x_2| \approx |\delta x_1| \tag{7.87}$$

With the aid of Eqs. (7.78) and (7.79), one concludes that

$$
\begin{aligned}
\delta g(x_1, x_2, x_3, x_4) &\doteq \left(\frac{1 - x_1}{2}\right)\delta x_1 - \left[\frac{1 - (x_1)^2}{2}\right]\delta x_2 + \frac{x_1\left[1 - (x_1)^2\right]}{3}\delta x_3 + \frac{\left[1 - (x_1)^2\right]^2}{6}\delta x_4 \\
&= \left[\left(\frac{1 - \nu}{2}\right)\delta\nu - \left(\frac{1 - \nu^2}{2}\right)\delta\nu_{\bar{x}} + \frac{\nu(1 - \nu^2)}{3}\delta\nu_{\bar{x}\bar{x}} + \frac{(1 - \nu^2)^2}{6}\delta\nu_{\bar{x}\bar{x}\bar{x}}\right]_{j+1/2}^{n-1/2}
\end{aligned}
\tag{7.88}
$$

Next, by combining Eqs. (7.75) and (7.88), one arrives at the conclusion

$$
\begin{aligned}
\delta g(x_1, x_2, x_3, x_4) &\doteq \left(\frac{1 - \tilde{\nu}}{2}\right)_j^n \delta\nu_{j+1/2}^{n-1/2} - \left(\frac{1 - \tilde{\nu}^2}{2}\right)_j^n \delta(\nu_{\bar{x}})_{j+1/2}^{n-1/2} \\
&+ \left[\frac{\tilde{\nu}(1 - \tilde{\nu}^2)}{3}\right]_j^n \delta(\nu_{\bar{x}\bar{x}})_{j+1/2}^{n-1/2} + \left[\frac{(1 - \tilde{\nu}^2)^2}{6}\right]_j^n \delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2}
\end{aligned}
\tag{7.89}
$$

Note that, according to Eqs. (7.75) and (7.79), Eq. (7.86) $\Leftrightarrow$

$$\left(\tilde{\nu}_j^n\right)^2 \doteq 1 \tag{7.90}$$

As such the validity of Eq. (7.89) has been established except the case Eq. (7.90).

By excluding the case Eq. (7.90) and using similar arguments, one can also derive the counterpart of Eq. (7.89) for the expression on the right side of Eq. (7.61) associated with the mesh point $(j - 1/2, n - 1/2)$. After combining Eq. (7.89) and its counterpart, one concludes that, excluding the case Eq. (7.90),

$$
\begin{aligned}
\delta\nu_j^n &\doteq \left(\frac{1 - \tilde{\nu}}{2}\right)_j^n \delta\nu_{j+1/2}^{n-1/2} + \left(\frac{1 + \tilde{\nu}}{2}\right)_j^n \delta\nu_{j-1/2}^{n-1/2} - \left(\frac{1 - \tilde{\nu}^2}{2}\right)_j^n \left[\delta(\nu_{\bar{x}})_{j+1/2}^{n-1/2} - \delta(\nu_{\bar{x}})_{j-1/2}^{n-1/2}\right] \\
&+ \left[\frac{\tilde{\nu}(1 - \tilde{\nu}^2)}{3}\right]_j^n \left[\delta(\nu_{\bar{x}\bar{x}})_{j+1/2}^{n-1/2} - \delta(\nu_{\bar{x}\bar{x}})_{j-1/2}^{n-1/2}\right] + \left[\frac{(1 - \tilde{\nu}^2)^2}{6}\right]_j^n \left[\delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2} - \delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2}\right]
\end{aligned}
\tag{7.91}
$$

The proof of Eq. (7.91) is completed if one can show that it is valid even for the case Eq. (7.90). Note that Eq. (7.90) $\Leftrightarrow$ either

$$\tilde{\nu}_j^n \doteq 1 \tag{7.92}$$

or

$$\tilde{\nu}_j^n \doteq -1 \tag{7.93}$$

but not both. On the other hand, it can be shown easily that (i)

$$\delta\nu_j^n \doteq \delta\nu_{j-1/2}^{n-1/2} \tag{7.94}$$

for the case Eq. (7.92), and (ii)

$$\delta\nu_j^n \doteq \delta\nu_{j+1/2}^{n-1/2} \tag{7.95}$$

for the case Eq. (7.93). By comparing Eq. (7.91) with Eqs. (7.94) and (7.95), it is clear that Eq. (7.91) is also valid for the case Eq. (7.90). QED.

Except for a possible complication to be discussed later, in a similar manner one can derive from Eqs. (7.62)–(7.64) the following results:

$$
\begin{aligned}
\delta(\nu_{\bar{x}})_j^n \doteq {}& \left[\frac{1}{2(1+\tau_2)}\right]_j^n \left(\delta\nu_{j+1/2}^{n-1/2} - \delta\nu_{j-1/2}^{n-1/2}\right) - \left[\frac{1-\tau_2+2\tilde{\nu}}{2(1+\tau_2)}\right]_j^n \delta(\nu_{\bar{x}})_{j+1/2}^{n-1/2} \\
& - \left[\frac{1-\tau_2-2\tilde{\nu}}{2(1+\tau_2)}\right]_j^n \delta(\nu_{\bar{x}})_{j-1/2}^{n-1/2} + \left[\frac{(1-\tau_2+2\tilde{\nu})^2}{4(1+\tau_2)} - \frac{(1+\tau_2)^2}{12(1+\tau_1)}\right]_j^n \delta(\nu_{\bar{x}\bar{x}})_{j+1/2}^{n-1/2} \\
& - \left[\frac{(1-\tau_2-2\tilde{\nu})^2}{4(1+\tau_2)} - \frac{(1+\tau_2)^2}{12(1+\tau_1)}\right]_j^n \delta(\nu_{\bar{x}\bar{x}})_{j-1/2}^{n-1/2} \\
& + \left[\frac{(1+\tau_2)^2(1-\tau_1+2\tilde{\nu})}{12(1+\tau_1)} - \frac{(1-\tau_2+2\tilde{\nu})^3}{12(1+\tau_2)}\right]_j^n \delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2} \\
& + \left[\frac{(1+\tau_2)^2(1-\tau_1-2\tilde{\nu})}{12(1+\tau_1)} - \frac{(1-\tau_2-2\tilde{\nu})^3}{12(1+\tau_2)}\right]_j^n \delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2}
\end{aligned}
\tag{7.96}
$$

$$
\delta(\nu_{\bar{x}\bar{x}})_j^n \doteq \frac{1}{2}\left\{(1-\tilde{\nu})_j^n \delta(\nu_{\bar{x}\bar{x}})_{j+1/2}^{n-1/2} + (1+\tilde{\nu})_j^n \delta(\nu_{\bar{x}\bar{x}})_{j-1/2}^{n-1/2} - (1-\tilde{\nu}^2)_j^n\left[\delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2} - \delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2}\right]\right\}
\tag{7.97}
$$

and

$$
\begin{aligned}
\delta(\nu_{\bar{x}\bar{x}\bar{x}})_j^n \doteq {}& \left[\frac{1}{2(1+\tau_1)}\right]_j^n \left[\delta(\nu_{\bar{x}\bar{x}})_{j+1/2}^{n-1/2} - \delta(\nu_{\bar{x}\bar{x}})_{j-1/2}^{n-1/2}\right] \\
& - \left[\frac{1-\tau_1+2\tilde{\nu}}{2(1+\tau_1)}\right]_j^n \delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2} - \left[\frac{1-\tau_1-2\tilde{\nu}}{2(1+\tau_1)}\right]_j^n \delta(\nu_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2}
\end{aligned}
\tag{7.98}
$$

In deriving Eqs. (7.96) and (7.98) from Eqs. (7.62) and (7.64), the reader should bear in mind the simplifying convention established earlier for the parameters $\tau_1$ and $\tau_2$. As an example, we have

$$
\left[\frac{\nu}{2(1+\tau_2)}\right]_{j\pm1/2}^{n-1/2} \stackrel{\text{def}}{=} \frac{\nu_{j\pm1/2}^{n-1/2}}{2\left[1+(\tau_2)_j^n\right]}
\tag{7.99}
$$

Thus, by using Eqs. (7.75) and (7.78), Eq. (7.99) implies that

$$
\delta\left[\frac{\nu}{2(1+\tau_2)}\right]_{j\pm1/2}^{n-1/2} \doteq \left[\frac{1}{2(1+\tau_2)}\right]_j^n \delta\nu_{j\pm1/2}^{n-1/2} - \left[\frac{\tilde{\nu}}{2(1+\tau_2)^2}\right]_j^n \delta(\tau_2)_j^n
\tag{7.100}
$$

In turn, Eq. (7.100) implies that

$$
\delta\left[\frac{\nu}{2(1+\tau_2)}\right]_{j+1/2}^{n-1/2} - \delta\left[\frac{\nu}{2(1+\tau_2)}\right]_{j-1/2}^{n-1/2} \doteq \left[\frac{1}{2(1+\tau_2)}\right]_j^n \left(\delta\nu_{j+1/2}^{n-1/2} - \delta\nu_{j-1/2}^{n-1/2}\right)
\tag{7.101}
$$

The expression on the right side of Eq. (7.101) is the first term on the right side of Eq. (7.96).

Eqs. (7.91) and (7.96)–(7.98) can be cast into the matrix form

$$
\delta\vec{q}(j,n) \doteq \left[\widehat{Q}_-(\tau_1,\tau_2,\tilde{\nu})\right]_j^n \delta\vec{q}(j+1/2, n-1/2) + \left[\widehat{Q}_+(\tau_1,\tau_2,\tilde{\nu})\right]_j^n \delta\vec{q}(j-1/2, n-1/2) \quad (j,n)\in\Omega
\tag{7.102}
$$

46

where

$$\delta \vec{q}(j,n) \stackrel{\text{def}}{=} \begin{pmatrix} \delta \nu_j^n \\ \delta(\nu_{\bar{x}})_j^n \\ \delta(\nu_{\bar{x}\bar{x}})_j^n \\ \delta(\nu_{\bar{x}\bar{x}\bar{x}})_j^n \end{pmatrix} \qquad (j,n) \in \Omega \qquad (7.103)$$

and

$$\left[\widehat{Q}_\pm(\tau_1, \tau_2, \tilde{\nu})\right]_j^n \stackrel{\text{def}}{=} \begin{pmatrix} \hat{q}_{11}^\pm & \hat{q}_{12}^\pm & \hat{q}_{13}^\pm & \hat{q}_{14}^\pm \\ \hat{q}_{21}^\pm & \hat{q}_{22}^\pm & \hat{q}_{23}^\pm & \hat{q}_{24}^\pm \\ 0 & 0 & \hat{q}_{33}^\pm & \hat{q}_{34}^\pm \\ 0 & 0 & \hat{q}_{43}^\pm & \hat{q}_{44}^\pm \end{pmatrix}_j^n \qquad (j,n) \in \Omega \qquad (7.104)$$

with

$$\hat{q}_{11}^\pm \stackrel{\text{def}}{=} \frac{1 \pm \tilde{\nu}}{2}, \quad \hat{q}_{12}^\pm \stackrel{\text{def}}{=} \pm\frac{(1 - \tilde{\nu}^2)}{2}, \quad \hat{q}_{13}^\pm \stackrel{\text{def}}{=} \mp\frac{\tilde{\nu}(1 - \tilde{\nu}^2)}{3}, \quad \hat{q}_{14}^\pm \stackrel{\text{def}}{=} \mp\frac{(1 - \tilde{\nu}^2)^2}{6} \qquad (7.105)$$

$$\hat{q}_{21}^\pm \stackrel{\text{def}}{=} \frac{\mp 1}{2(1 + \tau_2)}, \quad \hat{q}_{22}^\pm \stackrel{\text{def}}{=} -\frac{(1 - \tau_2 \mp 2\tilde{\nu})}{2(1 + \tau_2)}, \quad \hat{q}_{23}^\pm \stackrel{\text{def}}{=} \pm\frac{(1 + \tau_2)^2}{12(1 + \tau_1)} \mp \frac{(1 - \tau_2 \mp 2\tilde{\nu})^2}{4(1 + \tau_2)}$$

$$\hat{q}_{24}^\pm \stackrel{\text{def}}{=} \frac{(1 + \tau_2)^2(1 - \tau_1 \mp 2\tilde{\nu})}{12(1 + \tau_1)} - \frac{(1 - \tau_2 \mp 2\tilde{\nu})^3}{12(1 + \tau_2)} \qquad (7.106)$$

and

$$\hat{q}_{33}^\pm \stackrel{\text{def}}{=} \frac{1 \pm \tilde{\nu}}{2}, \quad \hat{q}_{34}^\pm \stackrel{\text{def}}{=} \pm\frac{(1 - \tilde{\nu}^2)}{2}, \quad \hat{q}_{43}^\pm \stackrel{\text{def}}{=} \frac{\mp 1}{2(1 + \tau_1)}, \quad \hat{q}_{44}^\pm \stackrel{\text{def}}{=} -\frac{(1 - \tau_1 \mp 2\tilde{\nu})}{2(1 + \tau_1)} \qquad (7.107)$$

A comparison of Eqs. (7.102)–(7.107) with Eqs. (6.7), (5.54), and (6.8)–(6.11) reveals a striking similarity in the structures of these two sets of equations. On the other hand, it was explained in Sec. 5.4 that: (i) computational instability would occur if round-off errors are amplified without bound as they propagate down the subsequent time levels; and (ii) for linear schemes like the $a$-$\epsilon$-4 and $c$-$\tau$-4 schemes, the time evolution of the mesh variables and that of the associated round-off errors are governed by the same system of equations. As a result of the facts stated above, one now arrives at the conclusion that **the stability conditions of the nonlinear inviscid Burgers $c$-$\tau$-4 scheme can be obtained from those of the linear $c$-$\tau$ scheme by replacing the constant coefficients $\nu$, $\tau_1$, and $\tau_2$ with the mesh-point dependent coefficients $\tilde{\nu}_j^n$, $(\tau_1)_j^n$, and $(\tau_2)_j^n$, respectively.**

### 7.3. The inviscid Burgers $c$-$\tau^*$-4 schemes

Let $h_1(s)$ and $h_2(s)$ be any two functions that meet the same requirements imposed on the function $h(s)$ introduced in Sec. 3.2. Then, because of the close relations among the $c$-$\tau$, $c$-$\tau$-4 and current schemes established above, here an "ideal" inviscid Burgers $c$-$\tau^*$-4 scheme is defined by an extension of Eq. (6.21), i.e.,

$$(\tau_1)_j^n = h_1\big((\tilde{\nu}_j^n)^2\big) \quad \text{and} \quad (\tau_2)_j^n = h_2\big((\tilde{\nu}_j^n)^2\big) \qquad (|\tilde{\nu}_j^n| < 1) \qquad (7.108)$$

According to Eq. (3.22), one can assume $h_1(s) = h_2(s) = \sqrt{s}$. As such a special inviscid Burgers $c$-$\tau^*$-4 scheme can be defined by the relation

$$(\tau_1)_j^n = (\tau_2)_j^n = |\tilde{\nu}_j^n| \qquad (|\tilde{\nu}_j^n| < 1) \qquad (7.109)$$

47

For robust real-world applications, we may also introduce a special inviscid Burgers $c$-$\tau$-4 scheme which is defined by an extension of Eq. (6.23), i.e.,

$$(\tau_1)_j^n = \beta_1 \, |\tilde{\nu}_j^n| \quad \text{and} \quad (\tau_2)_j^n = \beta_2 \, |\tilde{\nu}_j^n| \qquad\qquad (\beta_1 \geq 1; \ \beta_2 \geq 1; \ |\tilde{\nu}_j^n| < 1) \qquad (7.110)$$

where $\beta_1 \geq 1$ and $\beta_2 \geq 1$ are two adjustable parameters. In the rest of this section, we consider only the special schemes which satisfy at least one of Eqs. (7.59), (7.60), and (7.108)–(7.110). As a result of Eq. (7.43), for these schemes,

$$(\tau_1)_j^n \to 0 \quad \text{and} \quad (\tau_2)_j^n \to 0 \quad \text{as} \quad \Delta t \to 0 \qquad (7.111)$$

## 7.4. Shock capturing schemes

As explained in Sec. 2.3, the current Burgers solver needed to be modified so that it is capable of resolving shocks. Specifically, we need to turn the expression on the right side of Eq. (7.52) and/or that of Eq. (7.58) into proper weighted averages discussed in Sec. 2.3 and Sec. 4.

First we consider Eq. (7.52). Let

$$u_{\bar{x}\bar{x}}(P^+) \stackrel{\text{def}}{=} \left(\frac{\Delta x}{4}\right)^2 u_{xx}(P^+) \quad \text{and} \quad u_{\bar{x}\bar{x}}(P^-) \stackrel{\text{def}}{=} \left(\frac{\Delta x}{4}\right)^2 u_{xx}(P^-) \qquad (7.112)$$

where $u_{xx}(P^+)$ and $u_{xx}(P^-)$ are defined by Eqs. (7.47), (7.48), (7.50), and (7.51). Then it can be shown that

$$u_{\bar{x}\bar{x}}(P^+) = \left[(1 - 6\nu_{\bar{x}})u_{\bar{x}\bar{x}} - (1 - \tau_1 + 2\nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j+1/2}^{n-1/2} \qquad (7.113)$$

and

$$u_{\bar{x}\bar{x}}(P^-) = \left[(1 - 6\nu_{\bar{x}})u_{\bar{x}\bar{x}} + (1 - \tau_1 - 2\nu)u_{\bar{x}\bar{x}\bar{x}}\right]_{j-1/2}^{n-1/2} \qquad (7.114)$$

Thus Eq. (7.52) implies that

$$(u_{\bar{x}\bar{x}\bar{x}})_j^n = \frac{u_{\bar{x}\bar{x}}(P^+) - u_{\bar{x}\bar{x}}(P^-)}{2\left[1 + (\tau_1)_j^n\right]} = \frac{1}{2}\left(u_{\bar{x}\bar{x}\bar{x}-} + u_{\bar{x}\bar{x}\bar{x}+}\right)_j^n \qquad (7.115)$$

where

$$(u_{\bar{x}\bar{x}\bar{x}-})_j^n \stackrel{\text{def}}{=} \frac{(u_{\bar{x}\bar{x}})_j^n - u_{\bar{x}\bar{x}}(P^-)}{1 + (\tau_1)_j^n} \quad \text{and} \quad (u_{\bar{x}\bar{x}\bar{x}+})_j^n \stackrel{\text{def}}{=} \frac{u_{\bar{x}\bar{x}}(P^+) - (u_{\bar{x}\bar{x}})_j^n}{1 + (\tau_1)_j^n} \qquad (7.116)$$

with the understanding that $(u_{\bar{x}\bar{x}})_j^n$ is that defined in Eq. (7.44). Eqs. (7.115) and (7.116) can be considered as the substitution forms of Eqs. (3.10) and (4.1)–(4.3).

According to Eq. (7.115), $(u_{\bar{x}\bar{x}\bar{x}})_j^n$ evaluated using Eq. (7.52) is the simple average of $(u_{\bar{x}\bar{x}\bar{x}-})_j^n$ and $(u_{\bar{x}\bar{x}\bar{x}+})_j^n$. Moreover, by using Eqs. (7.111), (7.113)–(7.116) and

$$\nu_j^n \to 0, \quad (\nu_{\bar{x}})_j^n \to 0, \quad ,(\nu_{\bar{x}\bar{x}})_j^n \to 0, \quad \text{and} \quad (\nu_{\bar{x}\bar{x}\bar{x}})_j^n \to 0 \qquad \text{as} \quad \Delta t \to 0 \qquad (7.117)$$

(which follows from Eq. (7.43)), one can show that

$$\lim_{\Delta t \to 0} (u_{\bar{x}\bar{x}\bar{x}-})_j^n = \lim_{\Delta t \to 0} (u_{\bar{x}\bar{x}\bar{x}+})_j^n = \lim_{\Delta t \to 0} (u_{\bar{x}\bar{x}\bar{x}})_j^n = \frac{1}{2}\left[(u_{\bar{x}\bar{x}} - u_{\bar{x}\bar{x}\bar{x}})_{j+1/2}^{n-1/2} - (u_{\bar{x}\bar{x}} + u_{\bar{x}\bar{x}\bar{x}})_{j-1/2}^{n-1/2}\right] \qquad (7.118)$$

Because $(u_{\bar{x}\bar{x}\bar{x}-})_j^n$ and $(u_{\bar{x}\bar{x}\bar{x}+})_j^n$ satisfy Eq. (7.118), the simple average in Eq. (7.115) can be turned into an weighted average using any of the techniques described in Sec. 4. As an example, consider the technique presented in Sec. 4.4. The current counterparts of Eqs. (4.22)–(4.24) are

$$\left[(\eta_1)_\pm\right]_j^n \stackrel{\text{def}}{=} \frac{|(u_{\bar{x}\bar{x}\bar{x}\pm})_j^n|^{\alpha_1}}{\min\{|(u_{\bar{x}\bar{x}\bar{x}-})_j^n|^{\alpha_1}, |(u_{\bar{x}\bar{x}\bar{x}+})_j^n|^{\alpha_1}\}} - 1 \qquad (\alpha_1 > 0 \text{ and } |(u_{\bar{x}\bar{x}\bar{x}-})_j^n| + |(u_{\bar{x}\bar{x}\bar{x}+})_j^n| > 0) \qquad (7.119)$$

$$\left[(\tilde{w}_1)_+\right]_j^n \stackrel{\text{def}}{=} \left\{ \frac{1 + \sigma_1 (\eta_1)_-}{2 + \sigma_1[(\eta_1)_+ + (\eta_1)_-]} \right\}_j^n \qquad (\tilde{\nu}_j^n \neq 0) \tag{7.120}$$

and

$$\left[(\tilde{w}_1)_-\right]_j^n \stackrel{\text{def}}{=} \left\{ \frac{1 + \sigma_1 (\eta_1)_+}{2 + \sigma_1[(\eta_1)_+ + (\eta_1)_-]} \right\}_j^n \qquad (\tilde{\nu}_j^n \neq 0) \tag{7.121}$$

respectively. Here (i) $\alpha_1 > 0$ is a preset parameter in the order of 1, and (ii)

$$(\sigma_1)_j^n \stackrel{\text{def}}{=} \frac{(\sigma_1)_o}{|\tilde{\nu}_j^n|} \qquad ((\sigma_1)_o > 0 \text{ and } \tilde{\nu}_j^n \neq 0) \tag{7.122}$$

with $(\sigma_1)_o$ being a preset parameter in the order of 1. With the above definitions, the current counterpart of Eq. (4.30) can be expressed as

$$(u_{\bar{x}\bar{x}\bar{x}})_j^n = \left[(\tilde{w}_1)_-\right]_j^n (u_{\bar{x}\bar{x}\bar{x}-})_j^n + \left[(\tilde{w}_1)_+\right]_j^n (u_{\bar{x}\bar{x}\bar{x}+})_j^n \tag{7.123}$$

Next consider Eq. (7.58). Note that, with the aid of Eq. (7.52) and (7.57), it can be shown that Eq. (7.58) $\Leftrightarrow$

$$(u_{\bar{x}})_j^n = \frac{u^{4*}(P^+) - u^{4*}(P^-)}{2\left[1 + (\tau_2)_j^n\right]} - \frac{\left[1 + (\tau_2)_j^n\right]^2}{6}(u_{\bar{x}\bar{x}\bar{x}})_j^n \tag{7.124}$$

In the following, with the aid of Eq. (7.124), $(u_{\bar{x}\pm})_j^n$ will be defined so that

$$(u_{\bar{x}})_j^n = \frac{1}{2}(u_{\bar{x}-} + u_{\bar{x}+})_j^n \tag{7.125}$$

and

$$\lim_{\Delta t \to 0} (u_{\bar{x}-})_j^n = \lim_{\Delta t \to 0} (u_{\bar{x}+})_j^n = \lim_{\Delta t \to 0} (u_{\bar{x}})_j^n = \frac{1}{2}\left[\left(u - u_{\bar{x}} + \frac{1}{3}u_{\bar{x}\bar{x}}\right)_{j+1/2}^{n-1/2} - \left(u + u_{\bar{x}} + \frac{1}{3}u_{\bar{x}\bar{x}}\right)_{j-1/2}^{n-1/2}\right] \tag{7.126}$$

Note that the validity of the last equality sign follows directly from Eqs. (7.58), (7.111), and (7.117).

To facilitate the development, note that here (i) $P^+$ and $P^-$ are the points depicted in Fig. 3, *except that the parameter $\tau$ shown in the figure be replaced by the mesh-point dependent parameter* $(\tau_2)_j^n$; and (ii) $u^{4*}(P^+)$ and $u^{4*}(P^-)$ defined in Eq. (7.56) are 4th-order Taylor's approximations of the values of $u$ at points $P^+$ and $P^-$, respectively. Alternatively, these two values can also be approximated by

$$\tilde{u}^{4*}(P^\pm) \stackrel{\text{def}}{=} u^*\left(x(P^\pm), t^n; j, n\right) \tag{7.127}$$

respectively. With the aid of Eqs. (7.35), (7.42), (7.53), and (7.54), Eq. (7.127) implies that

$$\tilde{u}^{4*}(P^\pm) = u_j^n \pm \left[1 + (\tau_2)_j^n\right](u_{\bar{x}})_j^n + \frac{\left[1 + (\tau_2)_j^n\right]^2}{2}(u_{\bar{x}\bar{x}})_j^n \pm \frac{\left[1 + (\tau_2)_j^n\right]^3}{6}(u_{\bar{x}\bar{x}\bar{x}})_j^n \tag{7.128}$$

In turn, Eq. (7.128) implies that

$$\frac{\tilde{u}^{4*}(P^+) - u_j^n}{1 + (\tau_2)_j^n} - \frac{1 + (\tau_2)_j^n}{2}(u_{\bar{x}\bar{x}})_j^n = \frac{u_j^n - \tilde{u}^{4*}(P^-)}{1 + (\tau_2)_j^n} + \frac{1 + (\tau_2)_j^n}{2}(u_{\bar{x}\bar{x}})_j^n \tag{7.129}$$

Moreover, by using Eqs. (7.44), (7.46), (7.52), (7.57), (7.58), (7.111), (7.117) and (7.128), one concludes that

$$\lim_{\Delta t \to 0} \tilde{u}^{4*}(P^\pm) = \lim_{\Delta t \to 0} u^{4*}(P^\pm) = \left(u \mp u_{\bar{x}} + \frac{1}{2}u_{\bar{x}\bar{x}} \mp \frac{1}{6}u_{\bar{x}\bar{x}\bar{x}}\right)_{j\pm1/2}^{n-1/2} \tag{7.130}$$

49

An immediate result of Eqs. (7.129) and (7.130) is

$$\lim_{\Delta t \to 0} (u'_{\bar{x}+})^n_j = \lim_{\Delta t \to 0} (u'_{\bar{x}-})^n_j \tag{7.131}$$

where

$$(u'_{\bar{x}+})^n_j \overset{\text{def}}{=} \frac{u^{4*}(P^+) - u^n_j}{1 + (\tau_2)^n_j} - \frac{1 + (\tau_2)^n_j}{2} (u_{\bar{x}\bar{x}})^n_j \tag{7.132}$$

and

$$(u'_{\bar{x}-})^n_j \overset{\text{def}}{=} \frac{u^n_j - u^{4*}(P^-)}{1 + (\tau_2)^n_j} + \frac{1 + (\tau_2)^n_j}{2} (u_{\bar{x}\bar{x}})^n_j \tag{7.133}$$

As a result of Eqs. (7.132) and (7.133), one has

$$(u'_{\bar{x}+})^n_j + (u'_{\bar{x}-})^n_j = \frac{u^{4*}(P^+) - u^{4*}(P^-)}{1 + (\tau_2)^n_j} \tag{7.134}$$

Combining Eqs. (7.124), (7.131) and (7.134), one concludes that the requirements Eqs. (7.125) and (7.126) are met by the definition

$$(u_{\bar{x}\pm})^n_j \overset{\text{def}}{=} (u'_{\bar{x}\pm})^n_j - \frac{\left[1 + (\tau_2)^n_j\right]^2}{6} (u_{\bar{x}\bar{x}\bar{x}})^n_j \tag{7.135}$$

Because $(u_{\bar{x}+})^n_j$ and $(u_{\bar{x}-})^n_j$ meet the condition that they have the same value in the limit of $\Delta t \to 0$, the simple average on the right side of Eq. (7.125) can also be converted into an weighted average using any of the techniques described in Sec. 4. As an example, consider the technique presented in Sec. 4.4. The current counterparts of Eqs. (7.119)–(7.121) are

$$\left[(\eta_2)_\pm\right]^n_j \overset{\text{def}}{=} \frac{|(u_{\bar{x}\pm})^n_j|^{\alpha_2}}{\min\{|(u_{\bar{x}-})^n_j|^{\alpha_2}, |(u_{\bar{x}+})^n_j|^{\alpha_2}\}} - 1 \qquad (\alpha_2 > 0 \text{ and } |(u_{\bar{x}-})^n_j| + |(u_{\bar{x}+})^n_j| > 0) \tag{7.136}$$

$$\left[(\tilde{w}_2)_+\right]^n_j \overset{\text{def}}{=} \left\{ \frac{1 + \sigma_2 (\eta_2)_-}{2 + \sigma_2[(\eta_2)_+ + (\eta_2)_-]} \right\}^n_j \qquad (\tilde{\nu}^n_j \neq 0) \tag{7.137}$$

and

$$\left[(\tilde{w}_2)_-\right]^n_j \overset{\text{def}}{=} \left\{ \frac{1 + \sigma_2 (\eta_2)_+}{2 + \sigma_2[(\eta_2)_+ + (\eta_2)_-]} \right\}^n_j \qquad (\tilde{\nu}^n_j \neq 0) \tag{7.138}$$

respectively. Here (i) $\alpha_2 > 0$ is a preset parameter in the order of 1, and (ii)

$$(\sigma_2)^n_j \overset{\text{def}}{=} \frac{(\sigma_2)_o}{|\tilde{\nu}^n_j|} \qquad ((\sigma_2)_o > 0 \text{ and } \tilde{\nu}^n_j \neq 0) \tag{7.139}$$

with $(\sigma_2)_o$ being a preset parameter in the order of 1. With the above definitions, the current counterpart of Eq. (7.123) can be expressed as

$$(u_{\bar{x}})^n_j = \left[(\tilde{w}_2)_-\right]^n_j (u_{\bar{x}-})^n_j + \left[(\tilde{w}_2)_+\right]^n_j (u_{\bar{x}+})^n_j \tag{7.140}$$

Note that a result of Eqs. (7.125) and (7.135) is

$$(u_{\bar{x}})^n_j = \frac{1}{2} \left(u'_{\bar{x}+} + u'_{\bar{x}-}\right)^n_j - \frac{\left[1 + (\tau_2)^n_j\right]^2}{6} (u_{\bar{x}\bar{x}\bar{x}})^n_j \tag{7.141}$$

Thus an alternative shock-capturing form of Eq. (7.125) can be obtained by replacing the simple-average term on the right side of Eq. (7.141) with an weighted average of $(u'_{\bar{x}+})^n_j$ and $(u'_{\bar{x}-})^n_j$.

50

## 8. A sketch of multidimensional extensions

Consider the 2D advection equation

$$\frac{\partial u}{\partial t} + a_x \frac{\partial u}{\partial x} + a_y \frac{\partial u}{\partial y} = 0 \tag{8.1}$$

where $a_x$, and $a_y$ are constants. Let $x_1 = x$, $x_2 = y$, and $x_3 = t$ be the coordinates of a three-dimensional Euclidean space $E_3$. By using Gauss' divergence theorem in the space-time $E_3$, it can be shown that Eq. (8.1) is the differential form of the integral conservation law

$$\oint_{S(V)} \vec{h} \cdot d\vec{s} = 0 \tag{8.2}$$

Here (i) $S(V)$ is the boundary of an arbitrary space-time region $V$ in $E_3$, (ii)

$$\vec{h} \overset{\text{def}}{=} (a_x u, \, a_y u, \, u) \tag{8.3}$$

is a current density vector in $E_3$, 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 $S(V)$.

In addition, by using an argument similar to that was used to obtain the conservation condition Eq. (5.9), it can be shown that, for a smooth solution $u(x, y, t)$ of Eq. (8.1), we have

$$\oint_{S(V)} \vec{h}_{xx} \cdot d\vec{s} = 0, \quad \oint_{S(V)} \vec{h}_{xy} \cdot d\vec{s} = 0, \quad \text{and} \quad \oint_{S(V)} \vec{h}_{yy} \cdot d\vec{s} = 0 \tag{8.4}$$

where

$$\vec{h}_{xx} \overset{\text{def}}{=} \frac{\partial^2 \vec{h}}{\partial x^2} \equiv \left( a_x \frac{\partial^2 u}{\partial x^2}, a_y \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial x^2} \right), \quad \vec{h}_{xy} \overset{\text{def}}{=} \frac{\partial^2 \vec{h}}{\partial x \partial y} \equiv \left( a_x \frac{\partial^2 u}{\partial x \partial y}, a_y \frac{\partial^2 u}{\partial x \partial y}, \frac{\partial^2 u}{\partial x \partial y} \right),$$

$$\text{and} \quad \vec{h}_{yy} \overset{\text{def}}{=} \frac{\partial^2 \vec{h}}{\partial y^2} \equiv \left( a_x \frac{\partial^2 u}{\partial y^2}, a_y \frac{\partial^2 u}{\partial y^2}, \frac{\partial^2 u}{\partial y^2} \right) \tag{8.5}$$

In a 2D CESE scheme [13], a spatial domain can be divided into triangles of arbitrary shape with the understanding that any two neighboring triangles share a common side. Each triangle and the associated spatial solution point is assigned with an integer cell number $j$. The $x$- and $y$- coordinates of a spatial solution point $j$ are denoted by $x_j$ and $y_j$, respectively. In addition, a space-time solution point $(j, n)$ is defined to be the point with $x = x_j$, $y = y_j$, and $t = t^n \overset{\text{def}}{=} n\Delta t$. As in the 1D case, each $(j, n)$ is associated with a solution element $\text{SE}(j, n)$. However, corresponding to the three sides of a triangle, each $(j, n)$ is associated with three basic space-time conservation elements $\text{CE}_r(j, n)$, $r = 1, 2, 3$. Again these BCEs are constructed so that the boundary of each BCE can be divided into component parts with each of them belonging to a unique SE. The details are described in [13].

To construct a 4th-order scheme, for any $(x, y, t) \in \text{SE}(j, n)$, the dependent variable $u$ is approximated by a numerical analogue of the third-order Taylor's expansion, i.e.,

$$
\begin{aligned}
u^*(x, y, t; j, n) &\overset{\text{def}}{=} u_j^n + (u_x)_j^n(x - x_j) + (u_y)_j^n(y - y_j) + (u_t)_j^n(t - t^n) + \frac{1}{2}(u_{xx})_j^n(x - x_j)^2 \\
&+ \frac{1}{2}(u_{yy})_j^n(y - y_j)^2 + \frac{1}{2}(u_{tt})_j^n(t - t^n)^2 + (u_{xy})_j^n(x - x_j)(y - y_j) + (u_{xt})_j^n(x - x_j)(t - t^n) \\
&+ (u_{yt})_j^n(y - y_j)(t - t^n) + \frac{1}{6}(u_{xxx})_j^n(x - x_j)^3 + \frac{1}{6}(u_{yyy})_j^n(y - y_j)^3 + \frac{1}{6}(u_{ttt})_j^n(t - t^n)^3 \\
&+ \frac{1}{2}(u_{xxy})_j^n(x - x_j)^2(y - y_j) + \frac{1}{2}(u_{xyy})_j^n(x - x_j)(y - y_j)^2 + \frac{1}{2}(u_{xxt})_j^n(x - x_j)^2(t - t^n) \\
&+ \frac{1}{2}(u_{yyt})_j^n(y - y_j)^2(t - t^n) + \frac{1}{2}(u_{xtt})_j^n(x - x_j)(t - t^n)^2 + \frac{1}{2}(u_{ytt})_j^n(y - y_j)(t - t^n)^2 \\
&+ (u_{xyt})_j^n(x - x_j)(y - y_j)(t - t^n)
\end{aligned} \tag{8.6}
$$

Let $u = u^*(x, y, t\,; j, n)$ satisfy Eq. (8.1) uniformly in SE$(j, n)$, i.e.,

$$\frac{\partial u^*(x, y, t\,; j, n)}{\partial t} + a_x \frac{\partial u^*(x, y, t\,; j, n)}{\partial x} + a_y \frac{\partial u^*(x, y, t\,; j, n)}{\partial y} \equiv 0 \qquad (x, y, t) \in \text{ SE}(j, n) \qquad (8.7)$$

Then it can be shown that

$$(u_t)_j^n = -(a_x u_x + a_y u_y)_j^n \tag{8.8}$$

$$(u_{xt})_j^n = -(a_x u_{xx} + a_y u_{xy})_j^n \tag{8.9}$$

$$(u_{yt})_j^n = -(a_x u_{xy} + a_y u_{yy})_j^n \tag{8.10}$$

$$(u_{tt})_j^n = -(a_x u_{xt} + a_y u_{yt}) = \left[ (a_x)^2 u_{xx} + 2 a_x a_y u_{xy} + (a_y)^2 u_{yy} \right]_j^n \tag{8.11}$$

$$(u_{xxt})_j^n = -(a_x u_{xxx} + a_y u_{xxy})_j^n \tag{8.12}$$

$$(u_{xyt})_j^n = -(a_x u_{xxy} + a_y u_{xyy})_j^n \tag{8.13}$$

$$(u_{yyt})_j^n = -(a_x u_{xyy} + a_y u_{yyy})_j^n \tag{8.14}$$

$$(u_{xtt})_j^n = -(a_x u_{xxt} + a_y u_{xyt})_j^n = \left[ (a_x)^2 u_{xxx} + 2 a_x a_y u_{xxy} + (a_y)^2 u_{xyy} \right]_j^n \tag{8.15}$$

$$(u_{ytt})_j^n = -(a_x u_{xyt} + a_y u_{yyt})_j^n = \left[ (a_x)^2 u_{xxy} + 2 a_x a_y u_{xyy} + (a_y)^2 u_{yyy} \right]_j^n \tag{8.16}$$

and

$$(u_{ttt})_j^n = -(a_x u_{xtt} + a_y u_{ytt})_j^n = -\left[ (a_x)^3 u_{xxx} + 3(a_x)^2 a_y u_{xxy} + 3 a_x (a_y)^2 u_{xyy} + (a_y)^3 u_{yyy} \right]_j^n \tag{8.17}$$

Substituting Eqs. (8.8)–(8.17) into Eq. (8.6), one has

$$\begin{aligned}
u^*(x, y, t\,; j, n) &= u_j^n + (u_x)_j^n \left[ (x - x_j) - a_x(t - t^n) \right] + (u_y)_j^n \left[ (y - y_j) - a_y(t - t^n) \right] \\
&+ \frac{1}{2}(u_{xx})_j^n \left[ (x - x_j) - a_x(t - t^n) \right]^2 + \frac{1}{2}(u_{yy})_j^n \left[ (y - y_j) - a_y(t - t^n) \right]^2 \\
&+ (u_{xy})_j^n \left[ (x - x_j) - a_x(t - t^n) \right] \left[ (y - y_j) - a_y(t - t^n) \right] + \frac{1}{6}(u_{xxx})_j^n \left[ (x - x_j) - a_x(t - t^n) \right]^3 \\
&\frac{1}{6}(u_{yyy})_j^n \left[ (y - y_j) - a_y(t - t^n) \right]^3 + \frac{1}{2}(u_{xxy})_j^n \left[ (x - x_j) - a_x(t - t^n) \right]^2 \left[ (y - y_j) - a_y(t - t^n) \right] \\
&+ \frac{1}{2}(u_{xyy})_j^n \left[ (x - x_j) - a_x(t - t^n) \right] \left[ (y - y_j) - a_y(t - t^n) \right]^2
\end{aligned} \tag{8.18}$$

Thus $u^*(x, y, t\,; j, n)$ is a 3rd-order polynomial function of $\left[ (x - x_j) - a_x(t - t^n) \right]$ and $\left[ (y - y_j) - a_y(t - t^n) \right]$ specified by the coefficients $u_j^n$, $(u_x)_j^n$, $(u_y)_j^n$, $(u_{xx})_j^n$, $(u_{yy})_j^n$, $(u_{xy})_j^n$, $(u_{xxx})_j^n$, $(u_{yyy})_j^n$, $(u_{xxy})_j^n$, and $(u_{xyy})_j^n$. For a reason to be given later, in the current extension, only eight of these ten coefficients are considered as the independent mesh variables at $(j, n)$ while two of them (i.e., $(u_{xxy})_j^n$ and $(u_{xyy})_j^n$) are considered as functions of another four independent mesh variables at $(j, n)$. In other words, there are twelve independent mesh variables at each solution point $(j, n)$.

To construct the $a$-4 scheme for Eq. (8.1), we will impose the following three conservation conditions:

$$\oint_{S(CE_r(j,n))} \vec{h}^* \cdot d\vec{s} = 0, \qquad r = 1, 2, 3 \tag{8.19}$$

where

$$\vec{h}^*(x, y, t\,; j, n) \stackrel{\text{def}}{=} \left[ a_x u^*(x, y, t\,; j, n),\ a_y u^*(x, y, t\,; j, n),\ u^*(x, y, t\,; j, n) \right] \tag{8.20}$$

Eqs. (8.19) and (8.20) are the numerical analogues of Eqs. (8.2) and (8.3), respectively.

Note that Eq. (8.18) implies that

$$\frac{\partial^2 u^*(x,y,t\,;j,n)}{\partial x^2} = (u_{xx})_j^n + (u_{xxx})_j^n \left[(x-x_j) - a_x(t-t^n)\right] + (u_{xxy})_j^n \left[(y-y_j) - a_y(t-t^n)\right] \qquad (8.21)$$

$$\frac{\partial^2 u^*(x,y,t\,;j,n)}{\partial x \partial y} = (u_{xy})_j^n + (u_{xxy})_j^n \left[(x-x_j) - a_x(t-t^n)\right] + (u_{xyy})_j^n \left[(y-y_j) - a_y(t-t^n)\right] \qquad (8.22)$$

and

$$\frac{\partial^2 u^*(x,y,t\,;j,n)}{\partial y^2} = (u_{yy})_j^n + (u_{xyy})_j^n \left[(x-x_j) - a_x(t-t^n)\right] + (u_{yyy})_j^n \left[(y-y_j) - a_y(t-t^n)\right] \qquad (8.23)$$

Thus, at each $(j,n)$, imposition of conservation conditions Eq. (8.19) and the numerical analogue of Eq. (8.4), i.e.,

$$\oint_{S(CE_r(j,n))} \vec{h}_{xx}^* \cdot d\vec{s} = 0, \qquad r = 1,2,3 \qquad (8.24)$$

$$\oint_{S(CE_r(j,n))} \vec{h}_{xy}^* \cdot d\vec{s} = 0, \qquad r = 1,2,3 \qquad (8.25)$$

and

$$\oint_{S(CE_r(j,n))} \vec{h}_{yy}^* \cdot d\vec{s} = 0, \qquad r = 1,2,3 \qquad (8.26)$$

would result in twelve conditions for the ten coefficients that appear on the right side of Eq. (8.18) if one assumes that (i)

$$\vec{h}_{xx}^*(x,y,t;j,n) \overset{\text{def}}{=} \left[a_x u_{xx}^*(x,y,t;j,n),\, a_y u_{xx}^*(x,y,t;j,n),\, u_{xx}^*(x,y,t;j,n)\right] \qquad (8.27)$$

$$\vec{h}_{xy}^*(x,y,t;j,n) \overset{\text{def}}{=} \left[a_x u_{xy}^*(x,y,t;j,n),\, a_y u_{xy}^*(x,y,t;j,n),\, u_{xy}^*(x,y,t;j,n)\right] \qquad (8.28)$$

and

$$\vec{h}_{yy}^*(x,y,t;j,n) \overset{\text{def}}{=} \left[a_x u_{yy}^*(x,y,t;j,n),\, a_y u_{yy}^*(x,y,t;j,n),\, u_{yy}^*(x,y,t;j,n)\right] \qquad (8.29)$$

and (ii)

$$u_{xx}^*(x,y,t;j,n) = \frac{\partial^2 u^*(x,y,t\,;j,n)}{\partial x^2} \qquad (8.30)$$

$$u_{xy}^*(x,y,t;j,n) = \frac{\partial^2 u^*(x,y,t\,;j,n)}{\partial x \partial y} \qquad (8.31)$$

and

$$u_{yy}^*(x,y,t;j,n) = \frac{\partial^2 u^*(x,y,t\,;j,n)}{\partial y^2} \qquad (8.32)$$

To overcome the above problem of over determinacy, we replace Eqs. (8.30)–(8.32) with

$$u_{xx}^*(x,y,t;j,n) = (u_{xx})_j^n + (u_{xxx})_j^n \left[(x-x_j) - a_x(t-t^n)\right] + (u_{\underline{xxy}})_j^n \left[(y-y_j) - a_y(t-t^n)\right] \qquad (8.33)$$

$$u_{xy}^*(x,y,t;j,n) = (u_{xy})_j^n + (u_{\underline{xyx}})_j^n \left[(x-x_j) - a_x(t-t^n)\right] + (u_{\underline{xyy}})_j^n \left[(y-y_j) - a_y(t-t^n)\right] \qquad (8.34)$$

and

$$u_{yy}^*(x,y,t;j,n) = (u_{yy})_j^n + (u_{\underline{yyx}})_j^n \left[(x-x_j) - a_x(t-t^n)\right] + (u_{yyy})_j^n \left[(y-y_j) - a_y(t-t^n)\right] \qquad (8.35)$$

respectively. Note that: (i) the expression on the right sides of Eq. (8.33) is identical to that of Eq. (8.21) except that the symbol $(u_{xxy})_j^n$ in Eq. (8.21) is replaced by $(u_{\underline{xxy}})_j^n$ in Eq. (8.33); (ii) the expression on the right side of Eq. (8.34) is identical to that of Eq. (8.22) except that the symbols $(u_{xxy})_j^n$ and $(u_{xyy})_j^n$ in Eq. (8.22) are replaced by $(u_{\underline{xyx}})_j^n$ and $(u_{\underline{xyy}})_j^n$, respectively; and (iii) the expression on the right side of Eq. (8.35) is identical to that of Eq. (8.23) except that the symbol $(u_{xyy})_j^n$ in Eq. (8.23) is replaced by $(u_{\underline{yyx}})_j^n$. In spite of the facts that (i) $(u_{xxy})_j^n$, $(u_{\underline{xxy}})_j^n$, and $(u_{\underline{xyx}})_j^n$ are the "numerical analogues" of the value of the same mixed spatial derivative $\partial^3 u/\partial x^2 \partial y$ at $(j, n)$; and (ii) $(u_{xyy})_j^n$, $(u_{\underline{xyy}})_j^n$, and $(u_{\underline{yyx}})_j^n$ are the "numerical analogues" of the value of the same mixed spatial derivative $\partial^3 u/\partial x \partial y^2$ at $(j, n)$, here **(i)** $(u_{xxy})_j^n$, $(u_{\underline{xxy}})_j^n$, **and** $(u_{\underline{xyx}})_j^n$ **are treated as different numerical parameters and thus can assume different values; and (ii)** $(u_{xyy})_j^n$, $(u_{\underline{xyy}})_j^n$, **and** $(u_{\underline{yyx}})_j^n$ **are also treated as different numerical parameters and thus can assume different values**. In fact, $(u_{\underline{xxy}})_j^n$, $(u_{\underline{xyx}})_j^n$, $(u_{\underline{xyy}})_j^n$, **and** $(u_{\underline{yyx}})_j^n$ **are considered as independent mesh variables at** $(j, n)$ **while (i)** $(u_{xxy})_j^n$ **is considered as a function of** $(u_{\underline{xxy}})_j^n$ **and** $(u_{\underline{xyx}})_j^n$; **and (ii)** $(u_{xyy})_j^n$ **is considered as a function of** $(u_{\underline{xyy}})_j^n$ **and** $(u_{\underline{yyx}})_j^n$. In fact, one may assume that

$$(u_{xxy})_j^n = \frac{1}{2}(u_{\underline{xxy}} + u_{\underline{xyx}})_j^n \quad \text{and} \quad (u_{xyy})_j^n = \frac{1}{2}(u_{\underline{xyy}} + u_{\underline{yyx}})_j^n \tag{8.36}$$

As such, at each $(j, n)$, there are twelve independent mesh variables, i.e., $u_j^n$, $(u_x)_j^n$, $(u_y)_j^n$, $(u_{xx})_j^n$, $(u_{xxx})_j^n$, $(u_{\underline{xxy}})_j^n$, $(u_{xy})_j^n$, $(u_{\underline{xyx}})_j^n$, $(u_{\underline{xyy}})_j^n$, $(u_{yy})_j^n$, $(u_{\underline{yyx}})_j^n$, and $(u_{yyy})_j^n$. They can be determined in terms of the known mesh variables by imposing the twelve conservations conditions given in Eqs. (8.19) and (8.24)–(8.26). In fact, with the aid of Eqs. (8.27) and (8.33), the three mesh variables $(u_{xx})_j^n$, $(u_{xxx})_j^n$, and $(u_{\underline{xxy}})_j^n$ can be determined in terms of the known mesh variables by imposing the three conditions given in Eq. (8.24). Similarly, $(u_{xy})_j^n$, $(u_{\underline{xyx}})_j^n$, $(u_{\underline{xyy}})_j^n$ can be solved by using Eq. (8.25) while $(u_{yy})_j^n$, $(u_{\underline{yyx}})_j^n$, $(u_{yyy})_j^n$ can be solved by using Eq. (8.26). After the nine 2nd-order and 3rd-order independent mesh variables at $(j, n)$ are determined, the last three independent mesh variables, i.e., $u_j^n$, $(u_x)_j^n$, and $(u_y)_j^n$ can be solved using Eq. (8.19). Thus the $a$-4 scheme for Eq. (8.1) can be constructed easily. Moreover, with the aid of Eqs. (8.33)–(8.36), the dissipative extensions of the 2D $a$-4 scheme can also be constructed easily using well-established CESE techniques.

We conclude this section with the following remarks:

(a) The identities

$$\frac{\partial^2}{\partial x^2}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2}{\partial x \partial y}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial^2}{\partial y \partial x}\left(\frac{\partial u}{\partial x}\right) \tag{8.37}$$

and

$$\frac{\partial^2}{\partial y^2}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial^2}{\partial x \partial y}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2}{\partial y \partial x}\left(\frac{\partial u}{\partial y}\right) \tag{8.38}$$

are derived from a limiting process in which $\Delta x, \Delta y \to 0$. Because $\Delta x$ and $\Delta y$ are finite in a numerical simulation, logically the numerical analogues of the third-order mixed spatial derivatives are not required to satisfy the numerical analogues of Eqs. (8.37) and (8.38).

(b) Because $(u_{xxy})_j^n$ and $(u_{xyy})_j^n$ appear as 3rd-order terms on the right side of Eq. (8.18), the 4th-order accuracy requirement of $u^*(x, t; j, n)$ demands only that $(u_{xxy})_j^n$ and $(u_{xyy})_j^n$ be 1st-order in accuracy, i.e., the same orders of accuracy achieved by $(u_x)_j^n$ and $(u_y)_j^n$ in a 2nd-order 2D CESE scheme. This rather low requirement for accuracy further justifies the above introduction of $(u_{\underline{xxy}})_j^n$, $(u_{\underline{xyx}})_j^n$, $(u_{\underline{xyy}})_j^n$, and $(u_{\underline{yyx}})_j^n$.

(c) For the 3D CESE method, tetrahedrons are the most natural building blocks for structured or unstructured spatial meshes. As it turns out, the high order CESE framework described here can be naturally extended in a 3D space using spatial meshes built from tetrahedrons.

## 9. Numerical results

To assess the stability and accuracy of the $a$-$\epsilon$-4 and $c$-$\tau$-4 schemes, consider the model problem defined by the PDE Eq. (2.1) and the exact solution

$$u = u_e(x, t) \overset{\text{def}}{=} \sin\left[2\pi(x - t)\right] \tag{9.1}$$

We have

$$a = \lambda = T = 1 \tag{9.2}$$

where $\lambda$ = wavelength and $T$ = period. Let (i)

$$u_{xe}(x, t) \overset{\text{def}}{=} \frac{\partial u_e(x, t)}{\partial x}, \quad u_{xxe}(x, t) \overset{\text{def}}{=} \frac{\partial^2 u_e(x, t)}{\partial x^2}, \quad \text{and} \quad u_{xxxe}(x, t) \overset{\text{def}}{=} \frac{\partial^3 u_e(x, t)}{\partial x^3} \tag{9.3}$$

and (ii) the spatial domain of unit length be divided into $K$ uniform intervals. Thus

$$\Delta x = 1/K, \quad \Delta t = \nu \Delta x \quad \text{and} \quad t = n\Delta t \tag{9.4}$$

where $n$ = number of marching cycles (each cycle is formed by two consecutive marching steps), and $t$ = total marching time.

In Tables 1-14, the numerical errors of several computations using the $a$-$\epsilon$-4 and $c$-$\tau$-4 schemes are presented in terms of the $L_2$ error norms

$$E(K, n, \nu) \overset{\text{def}}{=} \sqrt{\frac{1}{K} \sum_{j=1}^{K} [u_j^n - u_e(x_j, t^n)]^2} \tag{9.5}$$

$$E_x(K, n, \nu) \overset{\text{def}}{=} \sqrt{\frac{1}{K} \sum_{j=1}^{K} [(u_x)_j^n - u_{xe}(x_j, t^n)]^2} \tag{9.6}$$

$$E_{xx}(K, n, \nu) \overset{\text{def}}{=} \sqrt{\frac{1}{K} \sum_{j=1}^{K} [(u_{xx})_j^n - u_{xxe}(x_j, t^n)]^2} \tag{9.7}$$

and

$$E_{xxx}(K, n, \nu) \overset{\text{def}}{=} \sqrt{\frac{1}{K} \sum_{j=1}^{K} [(u_{xxx})_j^n - u_{xxxe}(x_j, t^n)]^2} \tag{9.8}$$

Note that $(u_x)_j^n$, $(u_{xx})_j^n$, and $(u_{xxx})_j^n$ which appear in Eqs. (9.5)–(9.8) are *not* the normalized mesh variables defined in Eq. (7.42).

The numerical errors of the $a$-$\epsilon$-4 and $c$-$\tau$-4 schemes with periodic boundary conditions are shown in Tables 1–7. In all simulations, the values of $n$ are adjusted so that $t = 5.4$ for all the results presented.

The results with $\nu = 1$ generated using the $a$-$\epsilon$-4 scheme are given in Table 1 ($\epsilon_1 = \epsilon_2 = 0$) and Table 2 ($\epsilon_1 = \epsilon_2 = 0.5$). For these cases where the value of $\nu$ is right at the stability boundary, the errors for $u_j^n$ and $(u_{xx})_j^n$ are machine zero. On the other hand, the values of $E_x$ and $E_{xxx}$ are reduced by factors of about 16 and 4, respectively, as both $K$ and $n$ double their values. Thus, for both the cases (a) $\nu = 1$ and $\epsilon_1 = \epsilon_2 = 0$; and (b) $\nu = 1$ and $\epsilon_1 = \epsilon_2 = 0.5$, the $a$-$\epsilon$-4 scheme is 4th-order and 2nd-order in accuracy for $(u_x)_j^n$ and $(u_{xxx})_j^n$, respectively.

The results with $\nu = \epsilon_1 = \epsilon_2 = 0.5$ generated using the $a$-$\epsilon$-4 scheme are shown in Table 3. It is seen that, for this case, the $a$-$\epsilon$-4 scheme is 4th-order in accuracy for both $u_j^n$ and $(u_x)_j^n$ while it is 2nd-order in accuracy for both $(u_{xx})_j^n$ and $(u_{xxx})_j^n$.

The results with $\nu = 0.05$ and $\epsilon_1 = \epsilon_2 = 0.5$ generated using the $a$-$\epsilon$-4 scheme are shown in Table 4. It is seen that, for this case with a very small value of $\nu$, the $a$-$\epsilon$-4 scheme has an accuracy higher than

4th-order for both $u_j^n$ and $(u_x)_j^n$ when the meshes used are relatively coarse. For the coarse mesh cases, it also has an accuracy higher than 2nd-order for both $(u_{xx})_j^n$ and $(u_{xxx})_j^n$. According to the discussions given in Sec. 2 and 5, the special scheme with $\epsilon_1 = \epsilon_2 = 0.5$ (referred to as the $c$ scheme) has a huge advantage in computational efficiency over other $a$-$\epsilon$-4 schemes. As such the $c$ scheme is emphasized in our numerical study of the $a$-$\epsilon$-4 scheme.

The numerical results of the three cases with (a) $\nu = \tau_1 = \tau_2 = 1$, (b) $\nu = \tau_1 = \tau_2 = 0.5$, and (c) $\nu = \tau_1 = \tau_2 = 0.05$, are presented in Tables 5–7, respectively. According to a discussion given in Sec. 6, all these cases are the so-called $c$-$\tau^*$-4 schemes which form a special class of the $c$-$\tau$-4 schemes and are computational efficient and also Courant number insensitive. For case (a), the errors for $u_j^n$ and $(u_{xx})_j^n$ are again machine zero while it is 4th-order and 2nd-order in accuracy for $(u_x)_j^n$ and $(u_{xxx})_j^n$, respectively. For case (b), again the scheme is 4th-order in accuracy for $u_j^n$. Also it is 3rd-order, 2nd-order and 1st-order in accuracy for $(u_x)_j^n$, $(u_{xx})_j^n$, and $(u_{xxx})_j^n$ when the meshes used become finer and finer. In fact, for this case we have (i) $E = 2.45 \times 10^{-8}$, $E_x = 2.14 \times 10^{-6}$, $E_{xx} = 7.31 \times 10^{-3}$, and $E_{xxx} = 0.299$ when $K = 200$ and $n = 2160$; and (ii) $E = 1.53 \times 10^{-9}$, $E_x = 2.76 \times 10^{-7}$, $E_{xx} = 1.83 \times 10^{-3}$, and $E_{xxx} = 0.161$ when $K = 400$ and $n = 4320$. For case (c), as the meshes used become finer and finer, the scheme converges to 4th-order, 3rd-order, 2nd-order, and 1st-order in accuracy for $u_j^n$, $(u_x)_j^n$, $(u_{xx})_j^n$, and $(u_{xxx})_j^n$, respectively.

According to discussions presented in Sec. 2, 3, 5, and 6, in the case where $|\nu| \ll 1$, the $c$-$\tau^*$-4 schemes should be much more accurate than the 4th-order $c$ scheme. This conclusion is indeed consistent with the numerical results presented in Tables 4 and 7.

The results presented in Tables 8–14 are for the cases which are identical to those associated with the results presented in Tables 1–7 except that the periodic boundary conditions used in the latter cases are replaced in the former cases by the boundary conditions in which (i) the exact solution Eq. (9.1) is specified at the left spatial boundary $x = 0$, and (ii) the 4th-order non-reflecting boundary conditions Eqs. (5.83), (5.86), (5.88), and (5.89) are imposed on the right spatial boundary $x = 1$. As shown by the results presented in Tables 8–14, this changes of boundary conditions has no impact on the orders of accuracy for the $a$-$\epsilon$-4 and $c$-$\tau^*$-4 schemes. In fact, this is true even for the numerical results associated with the right spatial boundary $x = 1$.

The last numerical results to be presented is a shock solution for the problem defined by Eq. (7.1) and the initial condition

$$u = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x > 0 \end{cases} \qquad (t = 0) \qquad (9.9)$$

The exact solution of this problem is

$$u_e(x, t) = \begin{cases} 1 & \text{if } x < t/2 \\ 0 & \text{if } x > t/2 \end{cases} \qquad (9.10)$$

The numerical solution is generated using a scheme described in Sec. 7. Specifically, we assume Eqs. (7.44), (7.46), (7.110), (7.119)–(7.123) and the alternative shock-capturing form of Eq. (7.125) described following Eqs. (7.141) (the same procedure Eqs. (7.136)–(7.140) is used to obtain the weighted average of $(u'_{\bar{x}+})_j^n$ and $(u'_{\bar{x}-})_j^n$). We also assume that

$$\beta_1 = \beta_2 = \alpha_1 = \alpha_2 = 2 \quad \text{and} \quad (\sigma_1)_o = (\sigma_2)_o = 1 \qquad (9.11)$$

The spatial computational domain is defined by $-0.505 \leq x \leq 0.505$ and divided into 101 uniform interval, i.e $\Delta x = 0.01$. We also assume that $\Delta t = 0.009$ and $n = 50$, i.e., $t = n\Delta t = 0.45$. According to Eq. (9.10),

$$u_e(x, 0.45) = \begin{cases} 1 & \text{if } x < 0.225 \\ 0 & \text{if } x > 0.225 \end{cases} \qquad (9.12)$$

56

The numerical solution is shown against the exact solution in Fig. 4. Note that even though the maximal $CFL$ number $= 0.9$ for the exact solution, numerically, the maximal $CFL$ number reached during the simulation is 0.967.

## 10. Conclusions and discussions

A thorough and rigorous discussion of a new approach for constructing highly stable high-order CESE schemes has been presented. Each of the linear and nonlinear 4th-order schemes described in Sec. 5–7 has the same stencil and same stability conditions of its 2nd-order version and still retains all other advantages of the latter scheme. In particular, the $c$-$\tau$-4 scheme and its extensions for solving the inviscid Burgers equation Eq. (7.1) are completely explicit, i.e., implementation of these schemes does not require inverting systems of linear/nonlinear algebraic equations locally or globally. As such the new approach can overcome several common shortcomings of traditional high order schemes that were listed in the abstract.

Because the new 4th-order schemes are built from their 2nd-order versions, in addition to providing a review of the CESE core ideas and advantages in Sec. 1, a rather extensive review of several key 2nd-order CESE schemes were provided in Sec. 2–4. The special CESE techniques (e.g., the advanced weighted-average techniques described in Sec. 4) and the stability conditions presented in this review form a foundation for the 4th-order schemes discussed in Sec. 5–7.

To explain clearly how a 4th-order scheme can have the same stencil and same stability conditions of the 2nd-order scheme from which it is built, the conceptual leap that underlines the current development is clearly explained early in Sec. 5. This key idea is then thoroughly validated in the rest of the paper. The validation includes rigorous discussions of the stability conditions of the new 4th-order schemes. In particular, in Sec. 7.2, we present a stability study of a nonlinear scheme based on a rather exhaustive analysis of the propagation of round-off errors. As a result of this study, the reader can understand why the underlying idea is applicable even in a nonlinear case.

To complement the new 4th-order schemes, a set of simple 4th-order non-reflecting boundary conditions for linear schemes is presented in Sec. 5.5. Its potency is validated in Sec. 9. The extensions of these boundary conditions for nonlinear schemes are under development.

By introducing new independent mesh variables (e.g., those appear in Eqs. (8.33)–(8.35)) and the concept of dependent mesh variables (e.g., those defined in Eq. (8.36)), it is shown in Sec. 8 that the framework described in Sec. 5–7 can be extended to construct multidimensional high order schemes.

Accuracy and stability of the new 4th-order CESE schemes are rigorously assessed using the numerical results presented in Sec. 9. It is shown that the theoretical predictions made in Sec. 5–7 are consistent with the numerical results.

Even though they are not presented in the paper, from the details provided here, in a straightforward manner, the reader can build from a given 2nd-order CESE scheme its 6th-, 8th-,... order versions which have the same stencil and same stability conditions of the 2nd-order scheme.

Finally note that the current approach can also be used to construct high-order schemes from classical 2nd-order schemes. In fact, by using a "regular" space-time mesh in which (i) the set of mesh points used include all $(j, n)$ with $j, n = 0, \pm 1, \pm 2, \ldots$, and (ii) $x_j = j\Delta x$ and $t^n = n\Delta t$, a 3rd-order hybrid scheme has been built from the Lax-Wendroff and the second-order $a$ scheme. To construct this scheme, Eq. (5.4) is replaced by

$$u^*(x, t; j, n) = u_j^n + (u_x)_j^n \left[(x - x_j) - a\,(t - t^n)\right] + \frac{1}{2}(u_{xx})_j^n \left[(x - x_j) - a\,(t - t^n)\right]^2 \qquad (10.1)$$

Also, to account for the new mesh structure, we have

$$\nu \stackrel{\text{def}}{=} a\frac{\Delta t}{\Delta x}, \quad (u_{\bar{x}})_j^n \stackrel{\text{def}}{=} \frac{\Delta x}{2}(u_x)_j^n, \quad \text{and} \quad (u_{\bar{x}\bar{x}})_j^n \stackrel{\text{def}}{=} \left(\frac{\Delta x}{2}\right)^2 (u_{xx})_j^n, \qquad j, n = 0 \pm 1, \pm 2, \ldots \qquad (10.2)$$

Then the new hybrid scheme is formed by

$$(u_{\bar{x}\bar{x}})_j^n = \frac{\nu(\nu + 1)}{2}(u_{\bar{x}\bar{x}})_{j-1}^{n-1} + (1 - \nu^2)(u_{\bar{x}\bar{x}})_j^{n-1} + \frac{\nu(\nu - 1)}{2}(u_{\bar{x}\bar{x}})_{j+1}^{n-1} \qquad (10.3)$$

$$(u_{\bar{x}})_j^n = \frac{1}{2} \left[ (s_+)_{j+1}^{n-1} - (s_-)_{j-1}^{n-1} \right] - \frac{2\nu}{3} (u_{\bar{x}\bar{x}})_j^n \tag{10.4}$$

and

$$u_j^n = \frac{1}{2} \left[ (1 - \nu)(s_+)_{j+1}^{n-1} + (1 + \nu)(s_-)_{j-1}^{n-1} \right] - \frac{2}{3} (u_{\bar{x}\bar{x}})_j^n \tag{10.5}$$

where

$$(s_+)_j^n \overset{\text{def}}{=} \left[ u - (1 + \nu)u_{\bar{x}} + \frac{2(1 + \nu + \nu^2)}{3} u_{\bar{x}\bar{x}} \right]_j^n \qquad j, n = 0, \pm 1, \pm 2, \ldots \tag{10.6}$$

and

$$(s_-)_j^n \overset{\text{def}}{=} \left[ u + (1 - \nu)u_{\bar{x}} + \frac{2(1 - \nu + \nu^2)}{3} u_{\bar{x}\bar{x}} \right]_j^n \qquad j, n = 0, \pm 1, \pm 2, \ldots \tag{10.7}$$

As expected, the hybrid scheme is third order in accuracy and is stable if $\nu^2 < 1$.

## References

1. S.C. Chang and W.M. To, *A New Numerical Framework for Solving Conservation Laws–The Method of Space-Time Conservation Element and Solution Element*, NASA TM 104495, August 1991.
2. S.C. Chang, On An Origin of Numerical Diffusion: Violation of Invariance under Space-Time Inversion, in *Proceedings, 23rd Conference on Modeling and simulation, April 30-May 1, 1992, Pittsburgh, PA, USA*, edited by W.G. Volt and M.H. Mickle, Part 5, p. 2727. Also published as NASA TM 105776.
3. S.C. Chang and W.M. To, A Brief Description of a New Numerical Framework for Solving Conservation Laws—The Method of Space-Time Conservation Element and Solution Element, in *Proceedings of the Thirteenth International Conference on Numerical Methods in Fluid Dynamics, Rome, Italy, 1992*, edited by M. Napolitano and F. Sabetta, Lecture Notes in Physics 414, (Springer-Verlag, New York/Berlin, 1992), p. 396.
4. S.C. Chang, *New Developments in the Method of Space-Time Conservation Element and Solution Element—Applications to the Euler and Navier-Stokes Equations*, NASA TM 106226.
5. S.C. Chang, The method of space-time conservation element and solution Element—A new approach for solving the Navier-Stokes and Euler equations, *J. Comput. Phys.*, **119**, 295 (1995).
6. S.T. Yu and S.C. Chang, *Treatments of Stiff Source Terms in Conservation Laws by the Method of Space-Time Conservation Element and Solution Element*, AIAA Paper 97-0435 (1997).
7. S.C. Chang, S.T. Yu, A. Himansu, X.Y. Wang, C.Y. Chow, and C.Y. Loh, The method of space-time conservation element and solution element—A new paradigm for numerical solution of conservation laws, in *Computational Fluid Dynamics Review 1998* edited by M.M. Hafez and K. Oshima (World Scientific, Singapore), Vol. 1, p. 206.
8. T. Molls and F. Molls, Space-Time Conservation Method Applied to Saint Venant Equations, *J. of Hydraulic Engr.*, **124(5)**, 501 (1998).
9. S.T. Yu, S.C. Chang, P.C.E. Jorgenson, S.J. Park and M.C. Lai, "Treating Stiff Source Terms in Conservation Laws by the Space-Time Conservation Element and Solution Element Method," in *Proceedings of the 16th International Conference on Numerical Method in Fluid Dynamics, Arcachon, France, 6-10 July, 1998*, edited by C.H. Bruneau, (Springer-Verlag Berlin Heidelberg 1998), p. 433.
10. S.C. Chang, X.Y. Wang, and C.Y. Chow, *The Space-Time Conservation Element and Solution Element Method—A New High-Resolution and Genuinely Multidimensional Paradigm for Solving Conservation Laws. I. The Two-Dimensional Time Marching Schemes*, NASA TM—1998-208843, December, 1998.
11. S.C. Chang, X.Y. Wang, and C.Y. Chow, *The Space-Time Conservation Element and Solution Element Method—A New High-Resolution and Genuinely Multidimensional Paradigm for Solving Conservation Laws. II. Numerical Simulation of Shock Waves and Contact Discontinuities*, NASA TM—1998-208844, December, 1998.
12. S.C. Chang, X.Y. Wang, and C.Y. Chow, The space-time conservation element and solution element method: A new high-resolution and genuinely multidimensional paradigm for solving conservation laws, *J. Comput. Phys.*, **156**, 89 (1999).

13. X.Y. Wang, and S.C. Chang, A 2D non-splitting unstructured triangular mesh Euler solver based on the space-time conservation element and solution element method, *Computational Fluid Dynamics Journal*, **8(2)**, 309 (1999).

14. C. Zoppou and S. Roberts, Space-Time Conservation Method Applied to Saint Venant Equations: A Discussion, *J. of Hydraulic Engr.*, **125(8)**, 891 (1999).

15. N.S. Liu and K.H. Chen, *Flux: An Alternative Flow Solver for the National Combustion Code*, AIAA Paper 99-1079.

16. X.Y. Wang and S.C. Chang, A 3D structured/unstructured Euler solver based on the space-time conservation element and solution element method, in *A Collection of Technical Papers, 14th AIAA CFD Conference, June 28–July 1, 1999, Norfolk, Virginia*, AIAA Paper 99-3278.

17. S.C. Chang, X.Y. Wang and W.M. To, Application of the space-time conservation element and solution element method to one-dimensional convection-diffusion problems, *J. Comput. Phys.*, **165**, 189 (2000).

18. J. Qin, S.T. Yu, Z.C. Zhang, and M.C. Lai, Direct Calculations of Cavitating Flows by the Space-Time CESE Method, *J. Fuels & Lubricants, SAE Transc.*, **108(4)**, 1720 (2000).

19. C.Y. Loh, L.S. Hultgren and S.C. Chang, Wave computation in compressible flow using the space-time conservation element and solution element method, *AIAA J.*, **39(5)**, 794 (2001).

20. Z.C. Zhang, S.T. Yu, and S.C. Chang, A Space-Time Conservation Element and Solution Element Method for Solving the Two- and Three-Dimensional Unsteady Euler Equations Using Quadrilateral and Hexahedral Meshes, *J. Comput. Phys.*, **175**, 168 (2002).

21. K.B.M.Q. Zaman, M.D. Dahl, T.J. Bencic, and C.Y. Loh, Investigation of A 'Transonic Resonance' with Convergent-Divergent Nozzles, *J. Fluid Mech.*, **463**, 313 (2002).

22. C.Y. Loh and K.B.M.Q. Zaman, Numerical Investigation of 'Transonic Resonance' with A Convergent-Divergent Nozzle, *AIAA J.*, **40(12)**, 2393 (2002).

23. S. Motz, A. Mitrovic, and E.-D. Gilles, Comparison of Numerical Methods for the Simulation of Dispersed Phase Systems, *Chemical Engineering Science*, **57**, 4329 (2002).

24. S. Cioc and T.G. Keith, Application of the CESE Method to One-Dimensional Flow in Fluid Film Bearings, *STLE Tribology Transactions*, **45**, 167 (2002).

25. S.C. Chang, *Courant Number Insensitive CESE Schemes*, AIAA Paper 2002-3890 (2002).

26. S. Cioc and T.G. Keith, Application of the CESE Method to Two-Dimensional Flow in Fluid Film Bearings, *International J. of Numer. Methods for Heat & Fluid Flow*, **13(2)**, 216 (2003).

27. S. Cioc, F. Dimofte, T.G. Keith, and D.P. Fleming, Computation of Pressurized Gas Bearings Using the CESE Method, *STLE Tribology Transactions*, **46(1)**, 128 (2003).

28. S. Cioc, F. Dimofte, and T.G. Keith, Application of the CESE Method to Wave Journal Bearings, *STLE Tribology Transactions*, **46(2)**, 179 (2003).

29. A. Ayasoufi and T.G. Keith, Application of the Conservation Element and Solution Element Method in Numerical Modeling of Heat Conduction with Melting and/or Freezing, *International J. of Numer. Methods for Heat & Fluid Flow*, **13(4)**, 448 (2003).

30. I.S. Chang, *Unsteady Underexpanded Jet Flows*, AIAA Paper 2003-3885.

31. S.C. Chang and X.Y. Wang, *Multidimensional Courant Number Insensitive CESE Euler Solvers for Applications Involving Highly Nonuniform Meshes*, AIAA Paper 2003-5280.

32. A. Ayasoufi and T.G. Keith, Application of the Conservation Element and Solution Element Method in Numerical Modeling of Axisymmetric Heat Conduction with Melting and/or Freezing, *JSME International J. Series B*, **47(1)**, 115 (2004).

33. A. Ayasoufi and T.G. Keith, Application of the Conservation Element and Solution Element Method in Numerical Modeling of Three-dimensional Heat Conduction with Melting and/or Freezing, *Transactions of the ASME, J. of Heat Transfer*, **126(6)**, 937 (2004).

34. Y.I. Lim, S.C. Chang, and S.B. Jorgensen, A Novel Partial Differential Algebraic Equation (PDAE) Solver: Iterative Space-Time Conservation Element/Solution Element (CESE) Method, *Computers and Chemical Engineering*, **28**, 1309 (2004)

35. Y.I. Lim and S.B. Jorgensen, A Fast and Accurate Numerical Method for Solving Simulated Moving Bed (SMB) Chromatographic Separation Problems, *Chemical Engineering Science*, **59**, 1931 (2004).

36. Y.I. Lim, An Optimization Strategy for Nonlinear Simulated Moving Bed Chromatography: Multi-level Optimization Procedure (MLOP), *Korea J. Chem. Eng.*, 21(4), 836 (2004).

37. C.K. Kim, S.T. John Yu, and Z.C. Zhang, Cavity Flow in Scramjet Engine by the Space-Time Conservation Element and Solution Element Method, *AIAA J.*, **42(5)**, 912 (2004).

38. M. Zhang, S.T. John Yu, S.C. Lin, S.C. Chang, and I. Blankson, Solving Magnetohydrodynamic Equations Without Special Treatment for Divergence-Free Magnetic Field, *AIAA J.*, **42(12)**, 2605 (2004).

39. K.S. Im, M.C. Lai, S.T. John Yu, and Robert R. Matheson, Jr., Simulation of Spray Transfer Process in Electrostatic Rotary Bell Sprayer, *ASME J. of Fluid Engineering*, **126(3)**, 449 (2004).

40. S. Jerez, J.V. Romero, and M.D. Rosello, A Semi-Implicit Space-Time CESE Method to Improve Mass Conservation through Tapered Ducts in Internal Combustion Engines, *Math. and Computer Modeling*, **40**, 941 (2004).

41. B.S. Venkatachari, G.C. Cheng, and S.C. Chang, *Development of A Transient Viscous Flow Solver Based on Conservation Element-Solution Element Framework*, AIAA Paper 2004-3413.

42. S.C. Chang, Y. Wu, V. Yang, and X.Y. Wang, Local Time Stepping Procedures for the Space-Time Conservation Element and Solution Element Method, *International J. Comput. Fluid Dynamics*, **19(5)**, 359 (2005).

43. Y.I. Lim, S.B. Jorgensen, and I.H. Kim, Computer-Aided Model Analysis for Ionic Strength-Dependent Effective Charge of Protein in Ion-Exchange Chromatography, Bio-Chem. Eng. J., **25(2)**, 125 (2005).

44. T.I. Tseng and R.J. Yang, Simulation of the Mach Reflection in Supersonic Flows by the CESE Method, *Shock Waves*, **14(4)**, 307 (2005).

45. B. Wang, H. He, and S.-T.J. Yu, Direct Calculation of Wave Implosion for Detonation Initiation, *AIAA J.*, **43(10)**, 2157 (2005).

46. S.C. Chang, *Explicit von Neumann Stability Conditions for the c-τ Scheme—A Basic Scheme in the Development of the CESE Courant Number Insensitive Schemes*, NASA TM 2005-213627, April 2005.

47. J.C. Yen and D.A. Wagner, *Computational Aeroacoustics Using a Simplified Courant Number Insensitive CESE Method*, AIAA Paper 2005-2820.

48. I.S. Chang, C.L. Chang, and S.C. Chang, *Unsteady Navier-Stokes Rocket Nozzle Flows*, AIAA Paper 2005-4353.

49. S.C. Chang, *Courant Number and Mach Number Insensitive CESE Euler Solvers*, AIAA Paper 2005-4355

50. S.C. Chang, A. Himansu, C.Y. Loh, X.Y. Wang, and S.T. Yu, Robust and Simple Non-Reflecting Boundary Conditions for the Euler Equations–A New Approach Based on the Space-Time CESE Method, in *Proceedings, NSF-CBMS Regional Research Conference on Mathematical Methods in Nonlinear Wave Propagation, North Carolina A&T State University, Greensboro, North Carolina, May 15-19, 2002*, edited by D.P. Clemence and G. Tang, p. 155-190, Vol. 379 in *Contemporary Mathematics*, American Mathematical Society (2005). Also published as NASA/TM-2003-212495/Rev1.

51. B.S. Venkatachari, G.C. Cheng, and S.C. Chang, *Courant Number Insensitive Transient Viscous Flow Solver Based on CESE Framework*, AIAA Paper 2005-00931.

52. T.I. Tseng and R.J. Yang, Numerical Simulation of Vorticity Production in Shock Diffraction, *AIAA J.*, **44(5)**, 1040 (2006).

53. M. Zhang, S.-T. Yu, S.C.H. Lin, S.C. Chang, and I. Blankson, Solving the MHD Equations By the Space-Time Conservation Element and Solution Element Method, *J. Comput. Phys.*, **214**, 599 (2006).

54. I.S. Chang, C.L. Chang, and S.C. Chang, *3D Unsteady Navier-Stokes Rocket Nozzle Flows*, AIAA Paper 2006-4775.

55. C.L. Chang, *Time-accurate, Unstructured-Mesh Navier-Stokes Computations with the Space-time CESE Method*, AIAA Paper 2006-4780.

56. S.C. Chang, *On Space-Time Inversion Invariance and Its Relation to Non-Dissipativeness of a CESE Core Scheme*, AIAA Paper 2006-4779. Also published as NASA TM 2006-214385.

57. X. Feng, Y. Zhou, and S.T. Wu, A Novel Numerical Implementation for Solar Wind Modeling by the Modified Conservation Element/Solution Element Method, *Astrophys. J.*, **655(1)**, 1110 (2007).

58. C.L. Chang, *Three-Dimensional Navier-Stokes Calculations Using the Modified Space-Time CESE Method*, AIAA Paper 2007-5818.

59. S.C. Chang, *The $a(3)$ Scheme—A Fourth-Order Neutrally Stable CESE Solver*, AIAA Paper 2007-4321.

60. S.C. Chang, *The $a(4)$ Scheme—A High-Order Neutrally Stable CESE Solver*, AIAA Paper 2007-5820.

61. S.C. Chang, *The $a(3)$ Scheme—A Fourth-Order Space-Time Flux-Conserving and Neutrally Atable CESE Solver* (A revised and expanded version of AIAA Paper 2007-4321), NASA TM 2008-215138, April 2008.

62. S.C. Chang, *Courant Number Insensitive Schemes and New Classes of Wiggle-Suppressing Weighted-Averaging Techniques*, in preparation.

63. Other CESE references are posted on: http://www.grc.nasa.gov/www/microbus.

64. B. van Leer, Toward the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection, *J. Comput. Phys.*, **23**, 276 (1977).

65. G.D. van Albada, B. van Leer, and W.W. Robert, A Comparative Study of Computational Methods in Cosmic Gas Dynamics, *Astronom. and Astrophys.*, **108**, 76 (1982).

66. R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

Table 1. Numerical results of the $a$-$\varepsilon$-4 scheme with periodic boundary conditions

| $v = 1.0, \varepsilon_1 = \varepsilon_2 = 0, t = 5.4$ | | | |
|---|---|---|---|
| | K=25, n=135 | K=50, n=270 | K=100, n=540 |
| $E$ | 1.07E-15 | 1.38E-15 | 1.43E-15 |
| $E_X$ | 2.91E-05 | 1.81E-06 | 1.13E-07 |
| $E_{xx}$ | 4.21E-14 | 5.49E-14 | 5.66E-14 |
| $E_{xxx}$ | 1.09 | 0.272 | 6.79E-02 |

Table 2. Numerical results of the $a$-$\varepsilon$-4 scheme with periodic boundary conditions

| $v = 1.0, \varepsilon_1 = \varepsilon_2 = 0.5, t = 5.4$ | | | |
|---|---|---|---|
| | K=25, n=135 | K=50, n=270 | K=100, n=540 |
| $E$ | 1.07E-15 | 1.38E-15 | 1.43E-15 |
| $E_X$ | 2.48E-05 | 1.54E-06 | 9.62E-08 |
| $E_{xx}$ | 4.21E-14 | 5.49E-14 | 5.66E-14 |
| $E_{xxx}$ | 0.929 | 0.231 | 5.77E-02 |

Table 3. Numerical results of the $a$-$\varepsilon$-4 scheme with periodic boundary conditions

| $v = 0.5, \varepsilon_1 = \varepsilon_2 = 0.5, t = 5.4$ | | | |
|---|---|---|---|
| | K=25, n=270 | K=50, n=540 | K=100, n=1080 |
| $E$ | 1.59E-03 | 9.95E-05 | 6.21E-06 |
| $E_X$ | 1.00E-02 | 6.27E-04 | 3.91E-05 |
| $E_{xx}$ | 1.86 | 0.467 | 0.117 |
| $E_{xxx}$ | 11.7 | 2.94 | 0.735 |

Table 4. Numerical results of the $a$-$\varepsilon$-4 scheme with periodic boundary conditions

| $v = 0.05, \varepsilon_1 = \varepsilon_2 = 0.5, t = 5.4$ | | | |
|---|---|---|---|
| | K=25, n=2700 | K=50, n=5400 | K=100, n=10800 |
| $E$ | 1.14E-02 | 3.30E-04 | 1.35E-05 |
| $E_X$ | 7.30E-02 | 2.14E-03 | 8.69E-05 |
| $E_{xx}$ | 4.83 | 0.842 | 0.171 |
| $E_{xxx}$ | 30.6 | 5.37 | 1.09 |

Table 5. Numerical results of the $c$-$\tau$-4 scheme with periodic boundary conditions

| $v = \tau_1 = \tau_2 = 1.0, t = 5.4$ | | | |
|---|---|---|---|
| | K=25, n=135 | K=50, n=270 | K=100, n=540 |
| $E$ | 1.07E-15 | 1.38E-15 | 1.43E-15 |
| $E_X$ | 2.48E-05 | 1.54E-06 | 9.62E-08 |
| $E_{xx}$ | 4.21E-14 | 5.49E-14 | 5.66E-14 |
| $E_{xxx}$ | 0.929 | 0.231 | 5.77E-02 |

Table 6.  Numerical results of the $c$-$\tau$-4 scheme with periodic boundary conditions

| | $v = \tau_1 = \tau_2 = 0.5, t = 5.4$ | | |
| --- | --- | --- | --- |
| | K=25, n=270 | K=50, n=540 | K=100, n=1080 |
| $E$ | 1.05E-04 | 6.37E-06 | 3.94E-07 |
| $E_X$ | 5.58E-04 | 1.08E-04 | 1.59E-05 |
| $E_{xx}$ | 4.76E-01 | 1.18E-01 | 2.93E-02 |
| $E_{xxx}$ | 0.366 | 0.64 | 0.505 |

Table 7.  Numerical results of the $c$-$\tau$-4 scheme with periodic boundary conditions

| | $v = \tau_1 = \tau_2 = 0.05, t = 5.4$ | | |
| --- | --- | --- | --- |
| | K=25, n=2700 | K=50, n=5400 | K=100, n=10800 |
| $E$ | 5.10E-04 | 3.17E-05 | 1.98E-06 |
| $E_X$ | 3.67E-03 | 2.61E-04 | 2.02E-05 |
| $E_{xx}$ | 1.06 | 0.265 | 6.61E-02 |
| $E_{xxx}$ | 7.16 | 1.92 | 0.546 |

Table 8.  Numerical results of the $a$-$\varepsilon$-4 scheme with non-reflecting boundary conditions

| | $v = 1.0, \varepsilon_1 = \varepsilon_2 = 0, t = 5.4$ | | |
| --- | --- | --- | --- |
| | K=25, n=135 | K=50, n=270 | K=100, n=540 |
| $E$ | 2.18E-15 | 2.34E-15 | 2.48E-15 |
| $E_X$ | 4.21E-05 | 2.67E-06 | 1.67E-07 |
| $E_{xx}$ | 8.57E-14 | 9.21E-14 | 9.79E-14 |
| $E_{xxx}$ | 1.58 | 0.4 | 0.1 |

Table 9.  Numerical results of the $a$-$\varepsilon$-4 scheme with non-reflecting boundary conditions

| | $v = 1.0, \varepsilon_1 = \varepsilon_2 = 0.5, t = 5.4$ | | |
| --- | --- | --- | --- |
| | K=25, n=135 | K=50, n=270 | K=100, n=540 |
| $E$ | 2.18E-15 | 2.34E-15 | 2.48E-15 |
| $E_X$ | 2.46E-05 | 1.54E-06 | 9.60E-08 |
| $E_{xx}$ | 8.57E-14 | 9.21E-14 | 9.79E-14 |
| $E_{xxx}$ | 0.923 | 0.23 | 5.76E-02 |

Table 10.  Numerical results of the $a$-$\varepsilon$-4 scheme with non-reflecting boundary conditions

| | $v = \varepsilon_1 = \varepsilon_2 = 0.5, t = 5.4$ | | |
| --- | --- | --- | --- |
| | K=25, n=270 | K=50, n=540 | K=100, n=1080 |
| $E$ | 2.34E-05 | 1.30E-06 | 7.76E-08 |
| $E_X$ | 5.30E-04 | 4.98E-05 | 4.58E-06 |
| $E_{xx}$ | 0.227 | 5.63E-02 | 1.40E-02 |
| $E_{xxx}$ | 1.44 | 0.507 | 0.18 |

Table 11.  Numerical results of the $a$-$\varepsilon$-4 scheme with non-reflecting boundary conditions

| | $v = 0.05,\ \varepsilon_1 = \varepsilon_2 = 0.5,\ t = 5.4$ | | |
|---|---|---|---|
| | K=25, n=2700 | K=50, n=5400 | K=100, n=10800 |
| $E$ | 2.99E-04 | 8.25E-06 | 2.76E-07 |
| $E_X$ | 3.21E-03 | 1.06E-04 | 7.88E-06 |
| $E_{xx}$ | 0.522 | 9.20E-02 | 1.98E-02 |
| $E_{xxx}$ | 5.28 | 1.59 | 0.586 |

Table 12.  Numerical results of the $c$-$\tau$-4 scheme with non-reflecting boundary conditions

| | $v = \tau_1 = \tau_2 = 1.0,\ t = 5.4$ | | |
|---|---|---|---|
| | K=25, n=135 | K=50, n=270 | K=100, n=540 |
| $E$ | 2.18E-15 | 2.34E-15 | 2.48E-15 |
| $E_X$ | 2.46E-05 | 1.56E-06 | 9.60E-08 |
| $E_{xx}$ | 8.57E-14 | 9.21E-14 | 9.80E-14 |
| $E_{xxx}$ | 0.923 | 0.23 | 5.76E-02 |

Table 13.  Numerical results of the $c$-$\tau$-4 scheme with non-reflecting boundary conditions

| | $v = \tau_1 = \tau_2 = 0.5,\ t = 5.4$ | | |
|---|---|---|---|
| | K=25, n=270 | K=50, n=540 | K=100, n=1080 |
| $E$ | 2.64E-06 | 1.49E-07 | 8.84E-09 |
| $E_X$ | 2.35E-04 | 2.66E-05 | 3.09E-06 |
| $E_{xx}$ | 6.53E-02 | 1.57E-02 | 3.84E-03 |
| $E_{xxx}$ | 2.94 | 1.45 | 0.709 |

Table 14.  Numerical results of the $c$-$\tau$-4 scheme with non-reflecting boundary conditions

| | $v = \tau_1 = \tau_2 = 0.05,\ t = 5.4$ | | |
|---|---|---|---|
| | K=25, n=2700 | K=50, n=5400 | K=100, n=10800 |
| $E$ | 9.11E-06 | 5.46E-07 | 3.27E-08 |
| $E_X$ | 8.25E-04 | 6.75E-05 | 5.69E-06 |
| $E_{xx}$ | 0.125 | 3.12E-02 | 7.83E-03 |
| $E_{xxx}$ | 2.86 | 1.07 | 0.393 |
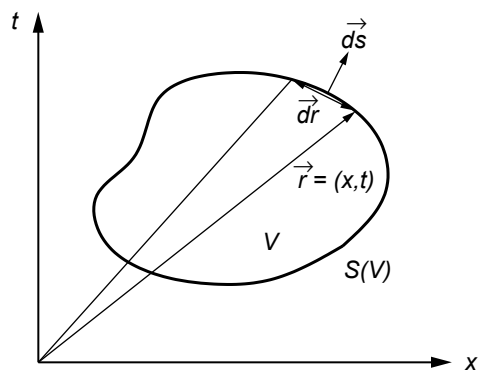
Figure 1.—A surface element on the boundary
$S(V)$ of an arbitrary space-time volume $V$.
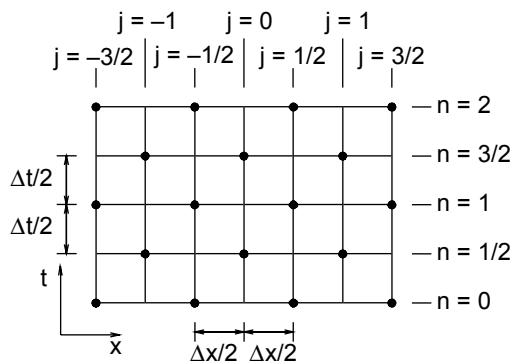


2(a)—The space-time mesh.
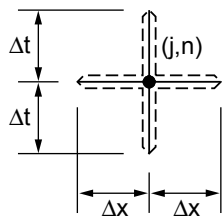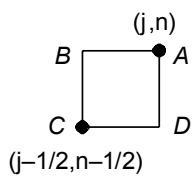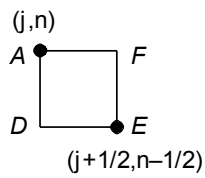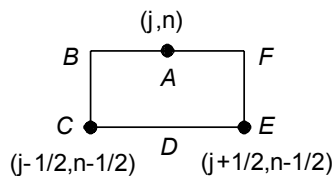


2(b) SE(j,n).          2(c) CE_(j,n).          2(d) CE$_+$(j,n).



2(e) CE(j,n)

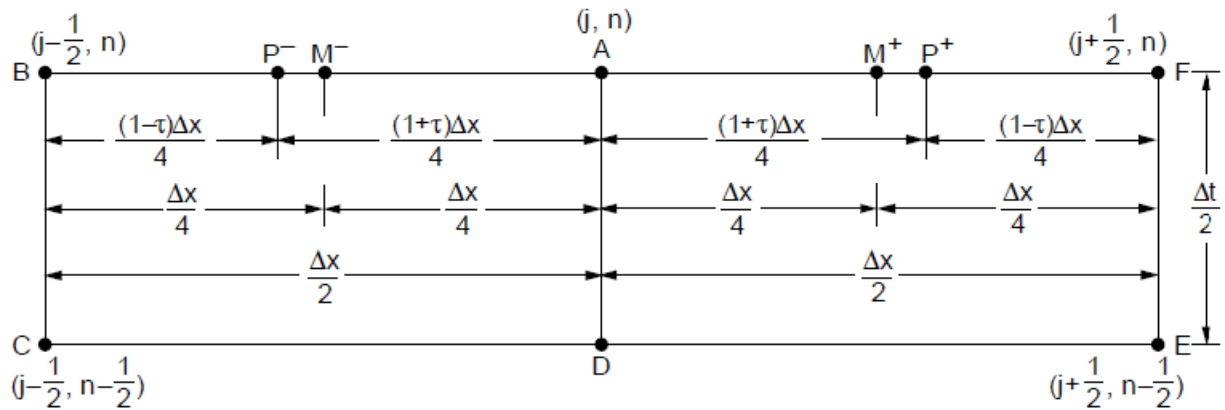Figure 2.—The SEs and CEs.

Figure 3.—Definition of $P^-$, $M^-$, $M^+$, and $P^+$.
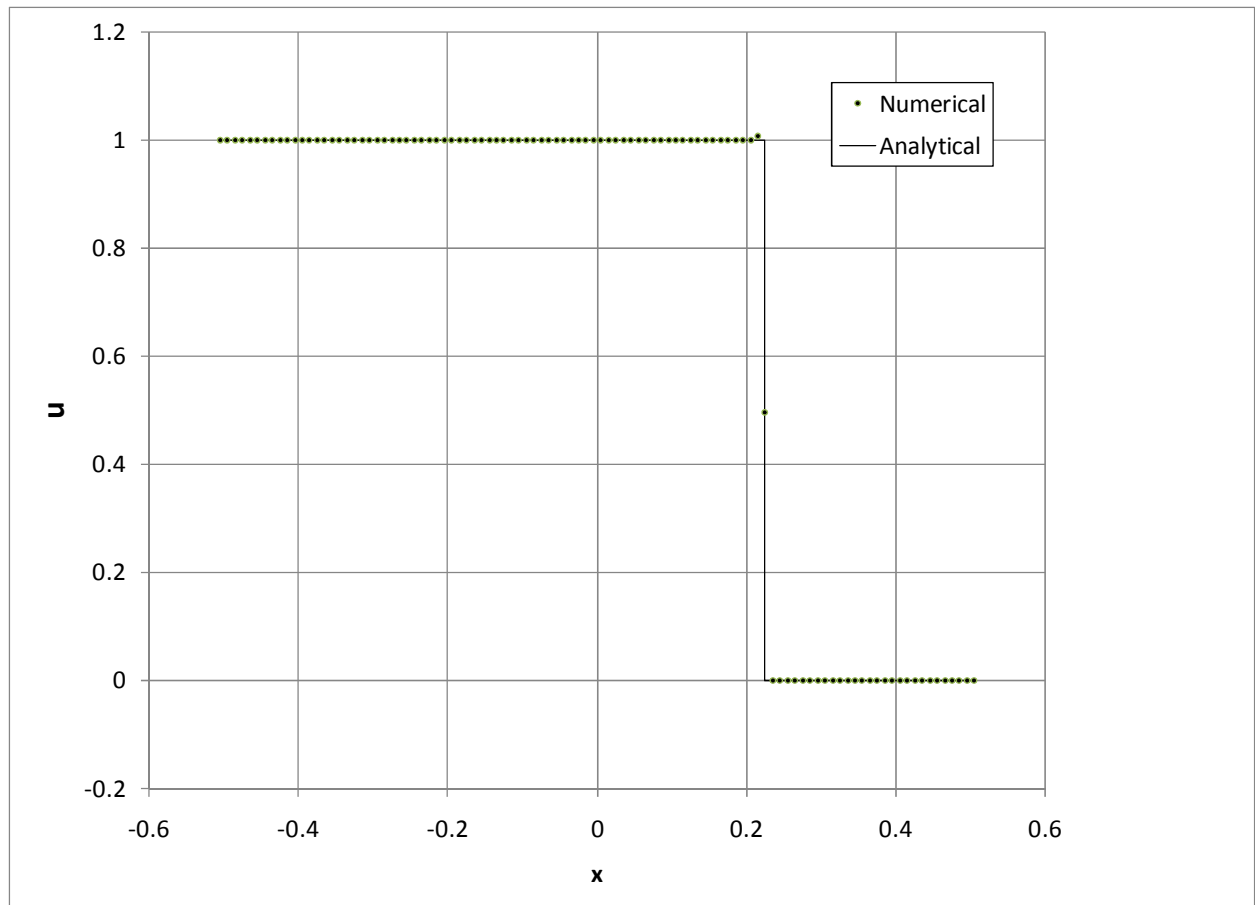


Figure 4.—A Burgers solution.

| 1. REPORT DATE (DD-MM-YYYY) 01-08-2010 | 2. REPORT TYPE Technical Memorandum | 3. DATES COVERED (From - To) |
|---|---|---|

**4. TITLE AND SUBTITLE**
A New Approach for Constructing Highly Stable High Order CESE Schemes

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Chang, Sin-Chung

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**
WBS 599489.02.07.03.04.03.01

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
National Aeronautics and Space Administration
John H. Glenn Research Center at Lewis Field
Cleveland, Ohio 44135-3191

**8. PERFORMING ORGANIZATION REPORT NUMBER**
E-17391

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
National Aeronautics and Space Administration
Washington, DC 20546-0001

**10. SPONSORING/MONITOR'S ACRONYM(S)**
NASA

**11. SPONSORING/MONITORING REPORT NUMBER**
NASA/TM-2010-216766

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
Unclassified-Unlimited
Subject Category: 64
Available electronically at http://gltrs.grc.nasa.gov
This publication is available from the NASA Center for AeroSpace Information, 443-757-5802

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
A new approach is devised to construct high order CESE schemes which would avoid the common shortcomings of traditional high order schemes including: (a) susceptibility to computational instabilities; (b) computational inefficiency due to their local implicit nature (i.e., at each mesh points, need to solve a system of linear/nonlinear equations involving all the mesh variables associated with this mesh point); (c) use of large and elaborate stencils which complicates boundary treatments and also makes efficient parallel computing much harder; (d) difficulties in applications involving complex geometries; and (e) use of problem-specific techniques which are needed to overcome stability problems but often cause undesirable side effects. In fact it will be shown that, with the aid of a conceptual leap, one can build from a given 2nd-order CESE scheme its 4th-, 6th-, 8th-,... order versions which have the same stencil and same stability conditions of the 2nd-order scheme, and also retain all other advantages of the latter scheme. A sketch of multidimensional extensions will also be provided.

**15. SUBJECT TERMS**
CESE method; High-order schemes

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON STI Help Desk (email:help@sti.nasa.gov) |
|---|---|---|---|---|---|
| a. REPORT U | b. ABSTRACT U | c. THIS PAGE U | UU | 72 | 19b. TELEPHONE NUMBER (include area code) 443-757-5802 |