NASA/TM—2008-215138

The *a*(3) Scheme—A Fourth-Order Space-Time Flux-Conserving and Neutrally Stable CESE Solver

Sin-Chung Chang
Glenn Research Center, Cleveland, Ohio

April 2008

NASA/TM—2008-215138

The *a*(3) Scheme—A Fourth-Order Space-Time Flux-Conserving and Neutrally Stable CESE Solver

Sin-Chung Chang
Glenn Research Center, Cleveland, Ohio

A revised and expanded version of AIAA–2007–4321, presented at the
18th Computational Fluid Dynamics Conference
sponsored by the American Institute of Aeronautics and Astronautics
Miami, Florida, June 25–28, 2007

National Aeronautics and
Space Administration

Glenn Research Center
Cleveland, Ohio 44135

April 2008

This work was sponsored by the Fundamental Aeronautics Program
at the NASA Glenn Research Center.

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

Available from

NASA Center for Aerospace Information
7115 Standard Drive
Hanover, MD 21076–1320

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161

Available electronically at http://gltrs.grc.nasa.gov

THE $a(3)$ SCHEME—A FOURTH-ORDER SPACE-TIME FLUX-CONSERVING AND NEUTRALLY STABLE CESE SOLVER

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

Abstract

The CESE development is driven by a belief that a solver should (i) enforce conservation laws in both space and time, and (ii) be built from a non-dissipative (i.e., neutrally stable) core scheme so that the numerical dissipation can be controlled effectively. To provide a solid foundation for a systematic CESE development of high order schemes, in this paper we describe the $a(3)$ scheme—a new 4th-order space-time flux-conserving and neutrally stable CESE solver of the advection equation $\partial u/\partial t + a\partial u/\partial x = 0$. The space-time stencil of this two-level explicit scheme is formed by one point at the upper time level and three points at the lower time level. Because it is associated with three independent mesh variables u_j^n, $(u_x)_j^n$, and $(u_{xx})_j^n$ (the numerical analogues of u, $\partial u/\partial x$, and $\partial^2 u/\partial x^2$, respectively) and three equations per mesh point, the new scheme is referred to as the $a(3)$ scheme. As in the case of other similar CESE neutrally stable solvers, the $a(3)$ scheme enforces conservation laws in space-time locally and globally, and it has the basic, forward marching, and backward marching forms. These forms are equivalent and satisfy a space-time inversion invariant property which is shared by the advection equation. (In physics, space-time inversion invariance is referred to as PT invariance where P denotes a parity, i.e., mirror-image or spatial-reflection, operation while T denote a time-reversal operation.) Based on the concept of PT invariance, a set of algebraic relations involving the coefficient matrices of the $a(3)$ scheme is developed. As it turns out, in the von Neumann analysis, these relations lead to the conclusion that the $a(3)$ scheme must be neutrally stable when it is stable. Also, in the same analysis, it is proved rigorously that: (i) all three amplification factors (i.e., the eigenvalues of the 3×3 amplification matrix) of the $a(3)$ scheme are of unit magnitude for all phase angles θ of the Fourier modes considered in the von Neumann analysis if and only if $|\nu| \leq 1/2$ ($\nu = a\Delta t/\Delta x$); (ii) the $a(3)$ scheme is stable if and only if $|\nu| < 1/2$; and (iii) the $a(3)$ scheme is linearly unstable (in a sense to be defined) if $|\nu| = 1/2$. These theoretical results have been confirmed numerically. Moreover, through numerical experiments, it is established that the $a(3)$ scheme generally is (i) 4th-order accurate for the mesh variables u_j^n and $(u_x)_j^n$; and (ii) 2nd-order accurate for $(u_{xx})_j^n$. However, in some exceptional cases, the scheme can achieve perfect accuracy aside from round-off errors. Finally the phase errors of the principal amplification factor of the $a(3)$ scheme are evaluated numerically and shown to be $O(\theta^4)$, a sharp reduction from those of the a scheme (the original neutrally stable CESE solver) which are $O(\theta^2)$.

1. Introduction

The space-time conservation element and solution element (CESE) method is a high-resolution and genuinely multidimensional method for solving conservation laws [1–73]. 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 novel 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 non-dissipative 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 marching 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 several 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 or other special techniques, since its inception in 1991 [1], the unstructured-mesh compatible 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 [51]. 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, electrodynamics, MHD vortex, hydraulic jump, crystal growth, and chromatographic problems [3–73]. In particular, its unexpected simple non-reflecting boundary conditions [9,68] and rather unique capability to resolve both strong shocks and small disturbances (e.g., acoustic waves) simultaneously [13,15,16] makes the CESE method an effective tool for attacking computational aeroacoustics (CAA) problems. Note that the fact that the second-order CESE schemes can solve CAA problems accurately is an exception to the commonly-held belief that a second-order scheme is not adequate for solving 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 potent capabilities, the core ideas of the CESE method are simple. In fact, all of its key features are the inescapable results of an honest 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 CESE requirement. Obviously, this requirement is also very helpful in 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, schemes with high-order accuracy generally have high accuracy and low numerical dissipation. However, they are susceptible to instability. In fact, in dealing with complicated real-world problems, stability of these schemes often is difficult to maintain without resorting to ad hoc treatments. To confront this issue head-on, in CESE development, generally it is required that a solver be built from a non-dissipative (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 non-dissipative. As such numerical dissipation can be controlled effectively if the deviation of a solver from its non-dissipative core scheme can be adjusted using some built-in parameters. Note that the above idea also plays an essential role in the recent successful development of a family of Courant number insensitive schemes [59,61,64,65,67].

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; (iii) triangular and tetrahedral meshes be used in 2D and 3D cases, respectively, so that the method is compatible to the simplest unstructured meshes and thus can be 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, as it turns out, the realization of the above lesser ideas (i)–(iv) follows effortlessly from that of the first two core ideas.

The first model equation considered in the CESE development is the simple advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{1.1}$$

where the advection speed $a \neq 0$ is a constant. Let $x_1 = x$, and $x_2 = t$ be considered as the coordinates of a two-dimensional Euclidean space E_2. Then, because Eq. (1.1) can be expressed as $\nabla \cdot \vec{h} = 0$ with $\vec{h} \overset{\text{def}}{=} (au, u)$, Gauss' divergence theorem in the space-time E_2 implies that Eq. (1.1) is the differential form of the integral conservation law

$$\oint_{S(V)} \vec{h} \cdot d\vec{s} = 0 \tag{1.2}$$

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. (1.2) 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.

It is well known that a solution to Eq. (1.1) represents *non-dissipative* data propagation along its characteristic lines defined by $dx/dt = a$. Moreover, Eq. (1.1) is invariant under space-time inversion, i.e., it transforms back to itself if x and t are replaced by $-x$ and $-t$, respectively. (In physics, space-time inversion invariance generally is referred to as PT invariance where P denotes a parity, i.e., mirror-image or spatial-reflection, operation while T denotes a time-reversal operation.) Thus a solution to Eq. (1.1) possesses the following properties: (i) it is completely determined by the data specified at an initial time level; (ii) its value at a space-time point has a finite domain of dependence (a point) at the initial time level; and (iii) the space-time inversion image of a solution to Eq. (1.1) is also a solution and vice versa. As such, in the initial CESE development, the focus is on the construction of an ideal core solver of Eq. (1.1) that enforces the conservation law Eq. (1.2) and also possesses properties similar to those of Eq. (1.1), i.e., it is a two-level, explicit, non-dissipative, and PT invariant solver. An in-depth account of this development and the resulting "a" scheme is given in [71]. As it turns out, the 2nd-order accurate a scheme (i) has a space-time stencil formed by one mesh point at the upper time level and two mesh points at the lower time level; and (ii) it is neutrally stable if $\nu^2 < 1$ where $\nu = a\Delta t/\Delta x$. Also, at each space-time mesh point (j, n), the a scheme is associated with two independent mesh variables u_j^n and $(u_x)_j^n$ (the numerical analogues of u and $\partial u/\partial x$, respectively) and two equations.

Until recently, with one exception (a three-level and 3rd-order accurate scheme reported on p. 80 of [1]), all CESE solvers of Eq. (1.1) are two-level and 2nd-order accurate extensions of the a scheme. To initiate a systematic CESE development of high-order schemes, in this paper we describe a new 4th-order accurate, space-time flux conserving, and neutrally stable CESE solver of Eq. (1.1). As will be shown, the space-time stencil of this two-level explicit scheme is formed by one point at the upper time level and three points at the lower time level. Because it is associated with three independent mesh variables u_j^n, $(u_x)_j^n$ and $(u_{xx})_j^n$ (the numerical analogues of u, $\partial u/\partial x$, and $\partial^2 u/\partial x^2$, respectively) and three equations at each mesh point, hereafter the new scheme is referred to as the $a(3)$ scheme.

The rest of the paper is organized as follows: In sec. 2, it is explained how the concepts of flux conservation and PT invariance along with a requirement to minimize the truncation error of its principal component equation uniquely define the $a(3)$ scheme. Also, for the $a(3)$ scheme, we present (i) several of its equivalent forms; (ii) the truncation errors of its three component equations; and (iii) a set of PT-invariance induced algebraic relations involving the coefficient matrices of its component equations.

A von Neumann analysis for the $a(3)$ scheme is presented in Sec. 3. Specifically, we provide rather rigorous and thorough discussions on the properties of the 3×3 amplification matrix and its eigenvalues (i.e., the amplification factors). In particular, it is proved that: (i) the $a(3)$ scheme must be neutrally stable if it is stable; (ii) all three amplification factors are of unit magnitude for all phase angles θ of the Fourier modes considered in the von Neumann analysis if and only if $|\nu| \leq 1/2$ ($\nu = a\Delta t/\Delta x$); (iii) the $a(3)$ scheme is stable if and only if $|\nu| < 1/2$; and (iv) the $a(3)$ scheme is linearly unstable (in a sense to be defined) if $|\nu| = 1/2$.

In addition to numerically verifying the theoretical predictions made in Sec. 3, in Sec. 4 it is shown that the $a(3)$ scheme generally is (i) 4th-order accurate for the mesh variables u_j^n and $(u_x)_j^n$; and (ii) 2nd-order accurate for $(u_{xx})_j^n$. However, as predicted from theoretical considerations, in some exceptional cases the scheme can achieve perfect accuracy aside from round-off errors. Moreover, it is shown that the phase errors of the principal amplification factor of the $a(3)$ scheme are $O(\theta^4)$ if $|\nu| < 1/2$, a sharp reduction from those of the dual a scheme [71] which are $O(\theta^2)$ if $|\nu| < 1$.

Conclusions and discussions are given in Sec. 5. Finally, several theorems and trigonometric identities used in Secs. 2 and 3 are proved in Appendices A and B while the three Fortran codes from which the numerical results presented in Sec. 4 are generated are listed in Appendices C, D, and E.

2. The $a(3)$ scheme

To proceed, consider the set Ω of space-time mesh points (j,n) (marked by dots and crosses in Fig. 2(a)) where

$$\Omega \overset{\text{def}}{=} \{(j,n)|j,n = 0, \pm 1, \pm 2, \pm 3, \ldots\} \tag{2.1}$$

We have

$$\Omega = \Omega_1 \cup \Omega_2 \tag{2.2}$$

where Ω_1 and Ω_2 are two disjoint sets defined by

$$\Omega_1 \overset{\text{def}}{=} \{(j,n)|j,n = 0, \pm 1, \pm 2, \pm 3, \ldots, \text{ and } (j+n) \text{ is an odd integer}\} \tag{2.3}$$

$$\Omega_2 \overset{\text{def}}{=} \{(j,n)|j,n = 0, \pm 1, \pm 2, \pm 3, \ldots, \text{ and } (j+n) \text{ is an even integer}\} \tag{2.4}$$

In Fig. 2(a), the mesh points $\in \Omega_1$ are marked by dots while those $\in \Omega_2$ are marked by crosses. Hereafter Ω_2 is referred to as the complement set of Ω_1 and vice versa. Obviously each of Ω_1 and Ω_2 represents a set of space-time staggered mesh points.

Each $(j,n) \in \Omega$ is associated with (i) a solution element (SE), denoted by SE(j,n) (see Fig. 2(b) where $(j,n) \in \Omega_1$ is assumed), and (ii) two conservation elements (CEs), denoted by CE$_-(j,n)$ and CE$_+(j,n)$ (see Figs. 2(c) and 2(d) where $(j,n) \in \Omega_1$ is assumed), respectively. Each SE is the *interior* of a *space-time* region that includes a horizontal line segment, a vertical line segment, and their immediate neighborhood. On the other hand, each CE is a rectangular space-time region. Hereafter, (i) SEs or CEs associated with mesh points $\in \Omega_1$ ($\in \Omega_2$) may be referred to simply as SEs or CEs associated with Ω_1 (Ω_2).

As a preliminary for the following development, note that (see Figs. 2(a)–(d)):

(a) Two CEs which are associated with two mesh points, one $\in \Omega_1$ while another $\in \Omega_2$ may occupy the same space-time region. As an example, (i) CE$_-(j,n)$ and CE$_+(j-1,n)$ occupy the same space-time region; and (ii) $(j,n) \in \Omega_1 \Leftrightarrow (j-1,n) \in \Omega_2$. Hereafter the symbol"$\Leftrightarrow$" is used as a shorthand for the statement "if and only if".

(b) A pair of diagonally opposite vertices of a CE both belong to the same set Ω_1 or Ω_2 while another pair both belong to the complement set. As an example, points A and C belong to Ω_1 while points B and D belong to Ω_2.

(c) The CEs associated with each of Ω_1 and Ω_2 by themselves are nonoverlapping and can fill the space-time E_2.

(d) Among the line segments forming the boundary of the same space-time region occupied by both CE$_-(j,n)$ and CE$_+(j-1,n)$, (i) $\overline{AB}$ and $\overline{AD} \subset$ SE(j,n); (ii) $\overline{CB}$ and $\overline{CD} \subset$ SE$(j-1,n-1)$; (iii) $\overline{BA}$ and $\overline{BC} \subset$ SE$(j-1,n)$; and (iv) $\overline{DA}$ and $\overline{DC} \subset$ SE$(j,n-1)$. Because $\overline{AB}$ and $\overline{BA}$ represent the same line segment, one can see that any line segment on this boundary is a subset of two SEs with one of them being associated with Ω_1 and another associated with Ω_2. *Hereafter, this ambiguity is removed by the following SE designation rule: any line segment designated as a boundary of a CE associated with Ω_1 (Ω_2) is designated as a subset of a SE associated with Ω_1 (Ω_2).* As an example, if $\overline{AB}$, $\overline{AD}$, $\overline{CB}$, and $\overline{CD}$ are designated as boundaries of CE$_-(j,n)$, then because points A and C belong to Ω_1, the above rule implies that: (i) both $\overline{AB}$ and $\overline{AD}$ are designated as subsets of SE(j,n); and (ii) both $\overline{CB}$ and $\overline{CD}$ are designated as subsets of SE$(j-1,n-1)$. On the other hand, if $\overline{BA}$, $\overline{BC}$, $\overline{DA}$, and $\overline{DC}$ are designated as boundaries of CE$_+(j-1,n)$, then: (i) both $\overline{BA}$ and $\overline{BC}$ are designated as subsets of SE$(j-1,n)$; and (ii) both $\overline{DA}$ and $\overline{DC}$ are designated as subsets of SE$(j,n-1)$.

Let $(x,t) \in$ SE(j,n). Then Eqs. (1.1) and (1.2) will be simulated numerically assuming that $u(x,t)$ and $\vec{h}(x,t)$, respectively, are approximated by

$$u^*(x,t\,;j,n) \overset{\text{def}}{=} u_j^n + (u_x)_j^n(x-x_j) + (u_t)_j^n(t-t^n) + \frac{1}{2}(u_{xx})_j^n(x-x_j)^2 + (u_{xt})_j^n(x-x_j)(t-t^n) + \frac{1}{2}(u_{tt})_j^n(t-t^n)^2 \tag{2.5}$$

and

$$\vec{h}^*(x,t\,;j,n) \overset{\text{def}}{=} \big(au^*(x,t\,;j,n),\, u^*(x,t\,;j,n)\big) \tag{2.6}$$

Note that: (i) u_j^n, $(u_x)_j^n$, $(u_t)_j^n$, $(u_{xx})_j^n$, $(u_{xt})_j^n$, and $(u_{tt})_j^n$ are constants in SE(j,n), and the numerical analogues of the values of u, $\partial u/\partial x$, $\partial u/\partial t$, $\partial^2 u/\partial x^2$, $\partial^2 u/\partial x \partial t$, and $\partial^2 u/\partial t^2$ at the mesh point (j,n), respectively; (ii) (x_j, t^n) are the coordinates of the mesh point (j,n) where $x_j = j\Delta x$ and $t^n = n\Delta t$; (iii) $u^*(x,t\,;j,n)$ represents a 2nd-order Taylor's approximation of u; and (iv) Eq. (2.6) is the numerical analogy of the definition $\vec{h} = (au, u)$.

For any $(j,n) \in \Omega$, let $u = u^*(x,t\,;j,n)$ satisfy Eq. (1.1) for all $(x,t) \in$ SE(j,n). Then one has

$$(u_t)_j^n = -a(u_x)_j^n, \quad (u_{xt})_j^n = -a(u_{xx})_j^n, \quad \text{and} \quad (u_{tt})_j^n = a^2(u_{xx})_j^n, \qquad (j,n) \in \Omega \tag{2.7}$$

Substituting Eq. (2.7) into Eq. (2.5), one has

$$u^*(x,t\,;j,n) = u_j^n + (u_x)_j^n \left[(x - x_j) - a\,(t - t^n)\right] + \frac{1}{2}(u_{xx})_j^n \left[(x - x_j) - a\,(t - t^n)\right]^2, \quad (j,n) \in \Omega \tag{2.8}$$

i.e., u_j^n, $(u_x)_j^n$, and $(u_{xx})_j^n$ are the only independent mesh variables associated with (j,n).

With the above preliminaries, next we derive the flux conservation relations that underline the $a(3)$ scheme.

2.1. Flux conservation relations

Let the flux of $\vec{h}^*$ conserve over all CEs, i.e.,

$$\oint_{S(CE_-(j,n))} \vec{h}^* \cdot d\vec{s} = 0, \qquad (j,n) \in \Omega \tag{2.9}$$

and

$$\oint_{S(CE_+(j,n))} \vec{h}^* \cdot d\vec{s} = 0, \qquad (j,n) \in \Omega \tag{2.10}$$

Because (i) with respect to CE$_-(j,n)$, the outward unit normal vectors $\vec{n}$ at $\overline{AB}$, $\overline{AD}$, $\overline{CD}$, and $\overline{CB}$ are $(0,1)$, $(1,0)$, $(0,-1)$, and $(-1,0)$, respectively; and (ii) with respect to CE$_+(j,n)$, the vectors $\vec{n}$ at $\overline{AF}$, $\overline{AD}$, $\overline{ED}$, and $\overline{EF}$ are $(0,1)$, $(-1,0)$, $(0,-1)$, and $(1,0)$, respectively, by using (i) the definitions given following Eq. (1.2), (ii) the above SE designation rule, and (iii) Eqs. (2.6) and (2.8), it can be shown that Eqs. (2.9) and (2.10) are equivalent to

$$(1+\nu)\left[u - (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_j^n = (1+\nu)\left[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j-1}^{n-1}, \quad (j,n) \in \Omega \tag{2.11}$$

and

$$(1-\nu)\left[u + (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_j^n = (1-\nu)\left[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j+1}^{n-1}, \quad (j,n) \in \Omega \tag{2.12}$$

respectively. Here: (i) $\nu \overset{\text{def}}{=} a\Delta t/\Delta x$ is the Courant number; (ii)

$$(u_{\bar{x}})_j^n \overset{\text{def}}{=} \frac{\Delta x}{2}(u_x)_j^n \quad \text{and} \quad (u_{\bar{x}\bar{x}})_j^n \overset{\text{def}}{=} \frac{(\Delta x)^2}{4}(u_{xx})_j^n \tag{2.13}$$

and (iii) 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. (2.12) as an example, i.e.,

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

At this juncture, note that:

(a) Because

$$\frac{\partial u}{\partial \bar{x}} = \frac{\Delta x}{2}\frac{\partial u}{\partial x} \quad \text{and} \quad \frac{\partial^2 u}{\partial \bar{x}^2} = \frac{(\Delta x)^2}{4}\frac{\partial^2 u}{\partial x^2} \qquad \text{if} \qquad \bar{x} \overset{\text{def}}{=} \frac{x}{\Delta x/2}$$

the normalized parameters $(u_{\bar{x}})_j^n$ and $(u_{\bar{x}\bar{x}})_j^n$, respectively, can be interpreted as the numerical analogues of the values at (j,n) of the first and second derivatives of u with respect to the normalized coordinate $\bar{x}$.

(b) By definition, points B and D depicted in Fig. 2(c) do not belong to either SE(j,n) or SE$(j-1,n-1)$. This fact, however, does not pose a problem for flux evaluation over $S(CE_-(j,n))$ because the values of $\vec{h}^*$ at isolated points do not contribute to the flux of $\vec{h}^*$ over a finite line segment. Similarly, the fact that points D and F depicted in Fig. 2(d) do not belong to SE(j,n) and SE$(j+1,n-1)$ does not pose a problem for flux evaluation over $S(CE_+(j,n))$.

(c) According to the SE designation rule, each line segment such as $\overline{AB}$ depicted in Fig. 2(c) can be assigned with two different fluxes of $\vec{h}^*$, one is associated with Ω_1 (hereafter referred to as the Ω_1-flux) and another associated with Ω_2 (hereafter referred to as the Ω_2-flux). As such, among those local conservation relations Eqs. (2.9) and (2.10), those associated with $(j,n) \in \Omega_1$ are completely decoupled from those associated with $(j,n) \in \Omega_2$. Because Eqs. (2.9) and (2.10) are equivalent to Eqs. (2.11) and (2.12), respectively, it follows that each of the two systems of equations defined by Eqs. (2.11) and (2.12) is formed by two decoupled subsystems, one is associated with Ω_1 while another associated with Ω_2.

(d) Moreover, because (i) the vector $\vec{h}^*$ at any interface separating two neighboring CEs associated with the same set Ω_1 (Ω_2) is evaluated using the information from the same SE, and (ii) the unit outward normal vector on the surface element pointing outward from one of these two neighboring CEs is exactly the negative of that pointing outward from another CE, one concludes that the flux leaving one of these CEs through the interface is the negative of that leaving another CE through the same interface. Due to this interface flux cancelation and the fact that the CEs associated with each of Ω_1 and Ω_2 by themselves are nonoverlapping and can fill the space-time E_2, the local conservation relations Eqs. (2.9) and (2.10) associated with $(j,n) \in \Omega_1$ ($(j,n) \in \Omega_2$) lead to a global conservation relation, i.e., *the total Ω_1- (Ω_2-) flux of $\vec{h}^*$ leaving the boundary of any space-time region that is the union of any combination of CEs associated with the same set Ω_1 (Ω_2) vanishes.*

Let $1-\nu^2 \neq 0$, i.e. $1+\nu \neq 0$ and $1-\nu \neq 0$. Then Eqs. (2.11) and (2.12) reduce to

$$\left[u - (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_j^n = \left[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j-1}^{n-1}, \quad (j,n) \in \Omega \tag{2.14}$$

and

$$\left[u + (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_j^n = \left[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j+1}^{n-1}, \quad (j,n) \in \Omega \tag{2.15}$$

respectively. Obviously, each of the two systems of equations defined by Eqs. (2.14) and (2.15) is also formed by two decoupled subsystems. Moreover, each component equation in Eq. (2.14) represents a stronger condition than the corresponding equation in Eq. (2.11) in the sense that the former implies the latter for any given ν while the latter implies the former only if an extra condition (i.e., $\nu \neq -1$ for this case) is imposed. Similarly, each component equation in Eq. (2.15) represents a stronger condition than the corresponding equation in Eq. (2.12). These stronger conditions will be used in the construction of the $a(3)$ scheme.

As a preliminary to a later development, next we will take a side tour and introduce the concept of invariance under space-time inversion.

2.2. Invariance under space-time inversion

Let $u = u(x,t)$ be a solution to Eq. (1.1) in the domain $-\infty < x,t < +\infty$, i.e.,

$$\frac{\partial u(x,t)}{\partial t} + a\frac{\partial u(x,t)}{\partial x} \equiv 0, \qquad -\infty < x,t < +\infty \tag{2.16}$$

Let

$$x' \overset{\text{def}}{=} -x \quad \text{and} \quad t' \overset{\text{def}}{=} -t \tag{2.17}$$

and

$$\hat{u}(x,t) \overset{\text{def}}{=} u(-x,-t) \tag{2.18}$$

Then (i) Eq. (2.16) $\Leftrightarrow$

$$\frac{\partial u(x',t')}{\partial t'} + a\frac{\partial u(x',t')}{\partial x'} \equiv 0, \qquad -\infty < x',t' < +\infty \tag{2.19}$$

and (ii)

$$\frac{\partial}{\partial t'} = -\frac{\partial}{\partial t} \quad \text{and} \quad \frac{\partial}{\partial x'} = -\frac{\partial}{\partial x} \tag{2.20}$$

Thus Eq. (2.16) $\Leftrightarrow$

$$\frac{\partial \hat{u}(x,t)}{\partial t} + a\frac{\partial \hat{u}(x,t)}{\partial x} \equiv 0, \qquad -\infty < x,t < +\infty \tag{2.21}$$

In other words, if $u = u(x,t)$ is a solution to Eq. (1.1), so must be $u = \hat{u}(x,t)$ and vice versa. Because the one-to-one mapping

$$(x,t) \leftrightarrow (-x,-t), \qquad -\infty < x,t < +\infty \tag{2.22}$$

represents a space-time inversion (PT) operation, hereafter (i) a pair of functions such as u and $\hat{u}$ will be referred to as the PT images of each other; and (ii) a partial differential equation (PDE) such as Eq. (1.1) is said to be PT invariant if the PT image of a solution is also a solution and vice versa.

Next let

$$u^{(k,\ell)}(x,t) \overset{\text{def}}{=} \frac{\partial^{k+\ell} u(x,t)}{\partial x^k \partial t^\ell} \quad \text{and} \quad \hat{u}^{(k,\ell)}(x,t) \overset{\text{def}}{=} \frac{\partial^{k+\ell} \hat{u}(x,t)}{\partial x^k \partial t^\ell}, \qquad -\infty < x,t < +\infty;\ k,\ell = 0,1,2,\ldots \tag{2.23}$$

Then, with the aid of the chain rule, Eqs. (2.17), (2.18), and (2.23) imply that

$$\begin{aligned}\hat{u}^{(k,\ell)}(x,t) &= \frac{\partial^{k+\ell} u(-x,-t)}{\partial x^k \partial t^\ell} = (-1)^{k+\ell}\frac{\partial^{k+\ell} u(x',t')}{\partial x'^k \partial t'^\ell} \qquad -\infty < x,t < +\infty;\ k,\ell = 0,1,2,\ldots \\ &= (-1)^{k+\ell} u^{(k,\ell)}(x',t') = (-1)^{k+\ell} u^{(k,\ell)}(-x,-t)\end{aligned} \tag{2.24}$$

i.e.,

$$\hat{u}^{(k,\ell)}(x,t) = \begin{cases} u^{(k,\ell)}(-x,-t) & \text{if } (k+\ell) \text{ is even} \\ -u^{(k,\ell)}(-x,-t) & \text{if } (k+\ell) \text{ is odd} \end{cases} \tag{2.25}$$

According to Eq. (2.23), $u^{(0,0)} = u$ and $\hat{u}^{(0,0)} = \hat{u}$. Thus Eq. (2.18) is a special case of Eq. (2.24) with $k = \ell = 0$.

In the following, the concept of PT invariance will be introduced for the $a(3)$ scheme. As a preliminary, note that: (i)

$$(j,n) \leftrightarrow (-j,-n) \tag{2.26}$$

is the numerical analogue of the PT mapping Eq. (2.22); and (ii) u_j^n, $(u_x)_j^n$, $(u_t)_j^n$, $(u_{xx})_j^n$, $(u_{xt})_j^n$, and $(u_{tt})_j^n$ are the numerical analogues of the values of u, $\partial u/\partial x$, $\partial u/\partial t$, $\partial^2 u/\partial x^2$, $\partial^2 u/\partial x \partial t$, and $\partial^2 u/\partial t^2$, at the mesh point (j,n), respectively. Thus, motivated by Eq. (2.25), the one-to-one mapping

$$\begin{matrix} u_j^n \leftrightarrow u_{-j}^{-n}; & (u_x)_j^n \leftrightarrow -(u_x)_{-j}^{-n}; & (u_t)_j^n \leftrightarrow -(u_t)_{-j}^{-n} \\ (u_{xx})_j^n \leftrightarrow (u_{xx})_{-j}^{-n}; & (u_{xt})_j^n \leftrightarrow (u_{xt})_{-j}^{-n}; & (u_{tt})_j^n \leftrightarrow (u_{tt})_{-j}^{-n} \end{matrix} \qquad (j,n) \in \Omega \tag{2.27}$$

is taken as the numerical analogue of the one-to-one mapping

$$u^{(k,\ell)}(x,t) \leftrightarrow \hat{u}^{(k,\ell)}(x,t), \qquad -\infty < x,t < +\infty;\ k,\ell = 0,1,2,3 \tag{2.28}$$

For the independent mesh variables, by using Eq. (2.13), Eq. (2.27) reduces to

$$\begin{pmatrix} u_j^n \\ (u_{\bar{x}})_j^n \\ (u_{\bar{x}\bar{x}})_j^n \end{pmatrix} \leftrightarrow \begin{pmatrix} u_{-j}^{-n} \\ -(u_{\bar{x}})_{-j}^{-n} \\ (u_{\bar{x}\bar{x}})_{-j}^{-n} \end{pmatrix}, \qquad (j,n) \in \Omega \tag{2.29}$$

Eq. (2.29) can be expressed as

$$\vec{q}(j,n) \leftrightarrow U\vec{q}(-j,-n), \qquad (j,n) \in \Omega \tag{2.30}$$

where

$$\vec{q}(j,n) \overset{\text{def}}{=} \begin{pmatrix} u_j^n \\ (u_{\bar{x}})_j^n \\ (u_{\bar{x}\bar{x}})_j^n \end{pmatrix}, \qquad (j,n) \in \Omega \tag{2.31}$$

and

$$U \overset{\text{def}}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.32}$$

The matrix U is unitary. In fact it is a real matrix with

$$U = U^{-1} \tag{2.33}$$

Hereafter (i) M^{-1} denotes the inverse of any nonsingular square matrix M; (ii) for each (j,n), $U\vec{q}(-j,-n)$ is referred to as the PT image of $\vec{q}(j,n)$; and (iii) the set formed by $U\vec{q}(-j,-n)$, $(j,n) \in \Omega$ is also referred to as the image of the set formed by $\vec{q}(j,n)$, $(j,n) \in \Omega$. According to Eq. (2.33), $\vec{q}(j,n) = UU\vec{q}(-(-j),-(-n))$. Thus $\vec{q}(j,n)$ is the PT image of $U\vec{q}(-j,-n)$ as an individual (j,n) or as the set defined over Ω. In the following, we will show that by itself each of the four subsystems of equations associated with Eqs. (2.14) and (2.15) is PT invariant, i.e., *the subsystem maps onto an equivalent subsystem under the mapping Eq. (2.29)*.

As an example, consider the subsystem of equations formed by the component equations associated with Ω_1 in Eq. (2.14). Let it be denoted as Eq. (2.14a). Under the mapping Eq. (2.29), Eq. (2.14a) maps onto

$$\left[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{-j}^{-n} = \left[u - (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{-(j-1)}^{-(n-1)}, \quad (j,n) \in \Omega_1 \tag{2.34}$$

At this juncture, note that, in addition to changing the sign of each $u_{\bar{x}}$, mapping Eq. (2.29) requires that the upper and lower indices j, n, $j-1$, and $n-1$ in Eq. (2.14a) be replaced by their negatives, respectively. *This is different from simply replacing the symbols j and n everywhere with $-j$ and $-n$, respectively*. Moreover, to simplify argument, hereafter system B is referred to as the PT image of system A if A maps onto B under the mapping Eq. (2.29), e.g., the subsystem Eq. (2.34) is the PT image of Eq. (2.14a). Let

$$j^* \overset{\text{def}}{=} 1 - j \quad \text{and} \quad n^* \overset{\text{def}}{=} 1 - n, \qquad (j,n) \in \Omega_1 \tag{2.35}$$

Then, by using the fact that $(j^* + n^*) + (j + n) = 2$ and therefore $(j^*, n^*) \in \Omega_1 \Leftrightarrow (j,n) \in \Omega_1$, Eq. (2.34) can be cast into the form

$$\left[u - (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j^*}^{n^*} = \left[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j^*-1}^{n^*-1}, \quad (j^*, n^*) \in \Omega_1 \tag{2.36}$$

By comparing Eqs. (2.14a) and (2.36), one can see that the subsystem Eq. (2.14a) is identical to its PT image Eq. (2.34) (which is identical to Eq. (2.36)). Thus, under the mapping Eq. (2.29), Eq. (2.14a) maps onto itself, i.e., the subsystem Eq. (2.14a) is PT invariant. QED.

The PT invariance of another three subsystems associated with Eqs. (2.14) and (2.15) can be established in a similar manner. As such the system formed by all component equations in each of Eqs. (2.14) and (2.15) is PT invariant.

The three mesh variables at any $(j,n) \in \Omega$ are linked to those at $(j-1,n-1)$ and $(j+1,n-1)$ by two component equations in Eqs. (2.14) and (2.15), respectively. In order that the three mesh variables at (j,n) can be determined in terms of those mesh variables at the $(n-1)$th time level, in the next subsection we introduce an extra PT invariant condition that links the mesh variables at (j,n) with those at the mesh point $(j,n-1)$.

2.3. A family of PT invariant solvers

Consider the following system of equations:

$$[u+\alpha u_{\bar{x}}+\beta u_{\bar{x}\bar{x}}]_j^n = [u-\alpha u_{\bar{x}}+\beta u_{\bar{x}\bar{x}}]_j^{n-1}, \qquad (j,n)\in\Omega \tag{2.37}$$

where α and β are parameters independent of (j,n). By definition, $(j,n)\in\Omega_1$ $(\Omega_2) \Leftrightarrow (j,n-1)\in\Omega_2$ (Ω_1). Thus, unlike Eqs. (2.14) and (2.15), *the mesh variables associated with Ω_1 are linked to those associated with Ω_2 through Eq. (2.37)*. However, as will be shown, like a subsystem associated with Eq. (2.14) or Eq. (2.15), the system of equations Eq. (2.37) is PT invariant for any pair of α and β.

The PT image of the system Eq. (2.37) is

$$[u-\alpha u_{\bar{x}}+\beta u_{\bar{x}\bar{x}}]_{-j}^{-n} = [u+\alpha u_{\bar{x}}+\beta u_{\bar{x}\bar{x}}]_{-j}^{-(n-1)}, \qquad (j,n)\in\Omega \tag{2.38}$$

Let

$$j' \overset{\text{def}}{=} -j \quad \text{and} \quad n' \overset{\text{def}}{=} 1-n, \qquad (j,n)\in\Omega \tag{2.39}$$

Then because $(j',n')\in\Omega \Leftrightarrow (j,n)\in\Omega$, Eq. (2.38) can be cast into the form

$$[u+\alpha u_{\bar{x}}+\beta u_{\bar{x}\bar{x}}]_{j'}^{n'} = [u-\alpha u_{\bar{x}}+\beta u_{\bar{x}\bar{x}}]_{j'}^{n'-1}, \qquad (j',n')\in\Omega \tag{2.40}$$

By comparing Eqs. (2.37) and (2.40), one can see that the system Eq. (2.37) is identical to its PT image Eq. (2.38) (which is identical to Eq. (2.40)). Thus, under the mapping Eq. (2.29), Eq. (2.37) maps onto itself, i.e., the system Eq. (2.37) is PT invariant. QED.

Because each of Eqs. (2.14) and (2.15) is PT invariant. one can see that, for any pair of α and β, the system formed by Eqs. (2.14), (2.15), and (2.37) is PT invariant.

Next, the three mesh variables at any $(j,n)\in\Omega$ will be solved in terms of those at $(j-1,n-1)$, $(j,n-1)$ and $(j+1,n-1)$ using Eqs. (2.14), (2.15), and (2.37). Let

$$\triangle \overset{\text{def}}{=} \frac{4}{3}(1+\alpha\nu)-2\beta \tag{2.41}$$

and assume $\triangle \neq 0$. Then it can be shown that Eqs. (2.14), (2.15), and (2.37) $\Leftrightarrow$

$$\begin{aligned} u_j^n = {} & \frac{4}{3\triangle}\left[u-\alpha u_{\bar{x}}+\beta u_{\bar{x}\bar{x}}\right]_j^{n-1} \\ & + \frac{1}{\triangle}\left[\left(\frac{2\alpha\nu}{3}-\beta\right)(1-\nu)-\frac{2\alpha}{3}\right]\left[u-(1+\nu)u_{\bar{x}}+\frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j+1}^{n-1} \\ & + \frac{1}{\triangle}\left[\left(\frac{2\alpha\nu}{3}-\beta\right)(1+\nu)+\frac{2\alpha}{3}\right]\left[u+(1-\nu)u_{\bar{x}}+\frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j-1}^{n-1} \end{aligned} \qquad (j,n)\in\Omega \tag{2.42}$$

$$
\begin{aligned}
(u_{\bar{x}})_j^n = {} & \frac{4\nu}{3\Delta}\Big[u - \alpha u_{\bar{x}} + \beta u_{\bar{x}\bar{x}}\Big]_j^{n-1} \\
& + \frac{1}{\Delta}\Big[\frac{2(1-\nu+\nu^2)}{3} - \beta\Big]\Big[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\Big]_{j+1}^{n-1} \qquad (j,n) \in \Omega \\
& - \frac{1}{\Delta}\Big[\frac{2(1+\nu+\nu^2)}{3} - \beta\Big]\Big[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\Big]_{j-1}^{n-1}
\end{aligned}
\tag{2.43}
$$

and

$$
\begin{aligned}
(u_{\bar{x}\bar{x}})_j^n = {} & -\frac{2}{\Delta}\Big[u - \alpha u_{\bar{x}} + \beta u_{\bar{x}\bar{x}}\Big]_j^{n-1} \\
& + \frac{1-\nu+\alpha}{\Delta}\Big[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\Big]_{j+1}^{n-1} \qquad (j,n) \in \Omega \\
& + \frac{1+\nu-\alpha}{\Delta}\Big[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\Big]_{j-1}^{n-1}
\end{aligned}
\tag{2.44}
$$

For any pair of α and β with $\Delta \neq 0$, Eqs. (2.42)–(2.44) represent a solver for Eq. (1.1). In the next subsection, we pick out the pair of α and β with which the solver will have the smallest truncation error (i.e., the highest order of truncation error) for Eq. (2.42).

2.4. A study of truncation error

Because, at each (j,n), Eqs. (2.42)–(2.44) represent a system of three equations for three *independent* mesh variables, Eqs. (2.42)–(2.44) represent a numerical analogue of a system of three coupled partial differential equations (PDEs) with three dependent variables. (Eq. (1.1) is one of these PDEs). As such, in the following study, three different symbols $\tilde{u}$, $\tilde{v}$, and $\tilde{w}$ will be used to denote the analytical versions of u_j^n, and the *non-normalized* variables $(u_x)_j^n$ and $(u_{xx})_j^n$, respectively. Specifically, let $\tilde{u}(x,t)$, $\tilde{v}(x,t)$, and $\tilde{w}(x,t)$ be functions having all the derivatives needed. Thus one can define

$$
\hat{v}(x,t) \overset{\text{def}}{=} \tilde{v}(x,t) - \frac{\partial \tilde{u}(x,t)}{\partial x} \quad \text{and} \quad \hat{w}(x,t) \overset{\text{def}}{=} \tilde{w}(x,t) - \frac{\partial^2 \tilde{u}(x,t)}{\partial x^2}
\tag{2.45}
$$

Also, as an example, one can define

$$
\left(\frac{\partial^{\ell+m}\tilde{u}}{\partial x^\ell \partial t^m}\right)_j^n \overset{\text{def}}{=} \frac{\partial^{\ell+m}\tilde{u}}{\partial x^\ell \partial t^m}(j\Delta x, n\Delta t) \qquad \ell, m = 0, 1, 2, \ldots
\tag{2.46}
$$

Next we will consider the "analytical" version of Eq. (2.42) which results from replacing (i) u_j^n, $(u_x)_j^n$, and $(u_{xx})_j^n$, respectively, with $\tilde{u}_j^n$, $\tilde{v}_j^n$, and $\tilde{w}_j^n$, for each (j,n); and (ii) the index n with $n+1$ everywhere. By using Eq. (2.13) and the fact that $(j, n+1) \in \Omega \Leftrightarrow (j,n) \in \Omega$, the analytical form can be expressed as

$$
\begin{aligned}
(e_1)_j^n \overset{\text{def}}{=} {} & \frac{1}{\Delta t}\Bigg\{\tilde{u}_j^{n+1} - \frac{4}{3\Delta}\Big[\tilde{u} - \frac{\alpha\Delta x}{2}\tilde{v} + \frac{\beta(\Delta x)^2}{4}\tilde{w}\Big]_j^n \\
& - \frac{1}{\Delta}\Big[\Big(\frac{2\alpha\nu}{3} - \beta\Big)(1-\nu) - \frac{2\alpha}{3}\Big]\Big[\tilde{u} - \frac{(1+\nu)\Delta x}{2}\tilde{v} + \frac{(1+\nu+\nu^2)\Delta x^2}{6}\tilde{w}\Big]_{j+1}^n \\
& - \frac{1}{\Delta}\Big[\Big(\frac{2\alpha\nu}{3} - \beta\Big)(1+\nu) + \frac{2\alpha}{3}\Big]\Big[\tilde{u} + \frac{(1-\nu)\Delta x}{2}\tilde{v} + \frac{(1-\nu+\nu^2)\Delta x^2}{6}\tilde{w}\Big]_{j-1}^n\Bigg\} = 0 \\
& (j,n) \in \Omega;\ \Delta \neq 0
\end{aligned}
\tag{2.47}
$$

By applying Taylor's formula, it can be shown that

$$\begin{aligned}
(e_1)_j^n \overset{\text{def}}{=} \Bigg\{ & \Big(\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\Big) + \frac{4(\alpha-\nu)}{3\Delta}\frac{\partial \tilde{u}}{\partial x}\frac{\Delta x}{\Delta t} + \Big(\frac{\partial^2 \tilde{u}}{\partial t^2} - a^2\frac{\partial^2 \tilde{u}}{\partial x^2}\Big)\frac{\Delta t}{2} - \frac{1}{\Delta}\Big[\frac{2\alpha\nu^3}{3} + \beta(1-\nu^2)\Big]\frac{\partial \hat{v}}{\partial x}\frac{(\Delta x)^2}{\Delta t} \\
& + \frac{2\nu(\nu-\alpha)}{3\Delta}\frac{\partial^2 \tilde{u}}{\partial x^2}\frac{(\Delta x)^2}{\Delta t} + \Big(\frac{\partial^3 \tilde{u}}{\partial t^3} + a^3\frac{\partial^3 \tilde{u}}{\partial x^3}\Big)\frac{(\Delta t)^2}{6} - \frac{\alpha}{3\Delta}\frac{\partial^2 \hat{v}}{\partial x^2}\frac{(\Delta x)^3}{\Delta t} \\
& + \frac{1}{3\Delta}\Big[\frac{2\alpha}{3}(1+\nu^2+\nu^4) - \beta\nu^3\Big]\frac{\partial \hat{w}}{\partial x}\frac{(\Delta x)^3}{\Delta t} + \frac{\alpha(1+4\nu^2) - 3\beta\nu - 2\nu^3}{9\Delta}\frac{\partial^3 \tilde{u}}{\partial x^3}\frac{(\Delta x)^3}{\Delta t} \\
& + \Big(\frac{\partial^4 \tilde{u}}{\partial t^4} - a^4\frac{\partial^4 \tilde{u}}{\partial x^4}\Big)\frac{(\Delta t)^3}{24} - \frac{1}{6\Delta}\Big[\frac{2\alpha\nu^3}{3} + \beta(1-\nu^2)\Big]\frac{\partial^3 \hat{v}}{\partial x^3}\frac{(\Delta x)^4}{\Delta t} + \frac{\beta}{6\Delta}\frac{\partial^2 \hat{w}}{\partial x^2}\frac{(\Delta x)^4}{\Delta t} \\
& + \frac{1}{12\Delta}\Big[\frac{2\nu^4}{3} + (1+2\nu^2-\nu^4)\Big(\beta - \frac{2\alpha\nu}{3}\Big)\Big]\frac{\partial^4 \tilde{u}}{\partial x^4}\frac{(\Delta x)^4}{\Delta t}\Bigg\}_j^n + O\big[(\Delta t)^4\big] \\
& + \frac{1}{\Delta}\Big[\frac{2\alpha}{3}(1-\nu+\nu^2) + \beta(1-\nu)\Big]\Big[O\big[(\Delta x)^5\big]/\Delta t + O\big[(\Delta x)^4\big] + O\big[\Delta t(\Delta x)^3\big]\Big] \\
& + \frac{1}{\Delta}\Big[\frac{2\alpha}{3}(1+\nu+\nu^2) - \beta(1+\nu)\Big]\Big[O\big[(\Delta x)^5\big]/\Delta t + O\big[(\Delta x)^4\big] + O\big[\Delta t(\Delta x)^3\big]\Big] \\
& (j,n) \in \Omega\ ;\ \Delta \neq 0
\end{aligned} \tag{2.48}$$

Note that $(e_1)_j^n$ defined in Eq. (2.47) is normalized by the factor $(1/\Delta t)$ so that the lowest-order terms in the above Taylor's expansion contain the leading term $(\partial\tilde{u}/\partial t + a\partial\tilde{u}/\partial x)$ which is independent of Δt and Δx. Also, in Eq. (2.48) a term is denoted by $O[(\Delta t)^{\ell_1}(\Delta x)^{\ell_2}]$ if there exists a constant $C > 0$ and two fixed integers $\ell_1 \geq 0$ and $\ell_2 \geq 0$ such that the absolute value of this term $< C(\Delta t)^{\ell_1}(\Delta x)^{\ell_2}$ for all sufficiently small Δt and Δx. Note that, *in determining the order of magnitude of a term such as $O\left[(\Delta x)^5\right]$ in Eq. (2.48), the parameters α and β are not assumed to be constants independent of Δt and Δx.* In fact, to reduce the truncation error of the $a(3)$ scheme, they will be chosen to be functions of ν (see Eqs. (2.58)) and thus vary with the ratio $\Delta t/\Delta x$.

In the following, let $u = \tilde{u}(x,t)$, $v = \tilde{v}(x,t)$, and $w = \tilde{w}(x,t)$ be a solution to the system of PDEs formed by Eq. (1.1) and

$$v - \frac{\partial u}{\partial x} = 0 \quad \text{and} \quad w - \frac{\partial^2 u}{\partial x^2} = 0 \tag{2.49}$$

i.e.,

$$\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x} \equiv 0, \quad \tilde{v} - \frac{\partial \tilde{u}}{\partial x} \equiv 0, \quad \text{and} \quad \tilde{w} - \frac{\partial^2 \tilde{u}}{\partial x^2} \equiv 0 \tag{2.50}$$

In other words, here the scheme Eqs. (2.42)–(2.44) is considered as a solver of the system of PDEs Eqs. (1.1) and (2.49). Eqs. (2.45) and (2.50) imply that

$$\frac{\partial^{\ell+m}\hat{v}}{\partial x^\ell \partial t^m} \equiv 0 \quad \text{and} \quad \frac{\partial^{\ell+m}\hat{w}}{\partial x^\ell \partial t^m} \equiv 0 \qquad \ell, m = 0, 1, 2, \ldots \tag{2.51}$$

$$\frac{\partial^2 \tilde{u}}{\partial t^2} - a^2\frac{\partial^2 \tilde{u}}{\partial x^2} \equiv \left(\frac{\partial}{\partial t} - a\frac{\partial}{\partial x}\right)\left(\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\right) \equiv 0 \tag{2.52}$$

$$\frac{\partial^3 \tilde{u}}{\partial t^3} + a^3\frac{\partial^3 \tilde{u}}{\partial x^3} \equiv \left(\frac{\partial^2}{\partial t^2} - a\frac{\partial^2}{\partial t \partial x} + a^2\frac{\partial^2}{\partial x^2}\right)\left(\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\right) \equiv 0 \tag{2.53}$$

$$\frac{\partial^4 \tilde{u}}{\partial t^4} - a^4\frac{\partial^4 \tilde{u}}{\partial x^4} \equiv \left(\frac{\partial^2}{\partial t^2} + a^2\frac{\partial^2}{\partial x^2}\right)\left(\frac{\partial}{\partial t} - a\frac{\partial}{\partial x}\right)\left(\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\right) \equiv 0 \tag{2.54}$$

Note that the first equation in Eq. (2.50), and Eqs. (2.52)–(2.54) are all special cases of

$$\frac{\partial^{\ell+m}}{\partial x^\ell \partial t^m}\left[\frac{\partial^k \tilde{u}}{\partial t^k} + (-1)^{k-1} a^k \frac{\partial^k \tilde{u}}{\partial x^k}\right] \equiv 0, \qquad \ell, m = 0, 1, 2, \ldots;\ k = 1, 2, 3 \ldots \tag{2.55}$$

With the hint provided by Eqs. (2.52)–(2.54), Eqs. (2.55) can be proved using the first equation in Eq. (2.50) and elementary algebra.

By using Eqs. (2.46) and (2.51)–(2.54), one can see that $(e_1)_j^n$ reduces to

$$\begin{aligned}(e_1)_j^n = \Bigg\{ & \frac{4(\alpha-\nu)}{3\Delta}\frac{\partial \tilde{u}}{\partial x}\frac{\Delta x}{\Delta t} + \frac{2\nu(\nu-\alpha)}{3\Delta}\frac{\partial^2 \tilde{u}}{\partial x^2}\frac{(\Delta x)^2}{\Delta t} + \frac{\alpha(1+4\nu^2) - 3\beta\nu - 2\nu^3}{9\Delta}\frac{\partial^3 \tilde{u}}{\partial x^3}\frac{(\Delta x)^3}{\Delta t} \\ & + \frac{1}{12\Delta}\left[\frac{2\nu^4}{3} + (1+2\nu^2-\nu^4)\left(\beta - \frac{2\alpha\nu}{3}\right)\right]\frac{\partial^4 \tilde{u}}{\partial x^4}\frac{(\Delta x)^4}{\Delta t}\Bigg\}_j^n + O\big[(\Delta t)^4\big] \\ & + \frac{1}{\Delta}\left[\frac{2\alpha}{3}(1-\nu+\nu^2) + \beta(1-\nu)\right]\Big[O\big[(\Delta x)^5\big]/\Delta t + O\big[(\Delta x)^4\big] + O\big[\Delta t(\Delta x)^3\big]\Big] \\ & + \frac{1}{\Delta}\left[\frac{2\alpha}{3}(1+\nu+\nu^2) - \beta(1+\nu)\right]\Big[O\big[(\Delta x)^5\big]/\Delta t + O\big[(\Delta x)^4\big] + O\big[\Delta t(\Delta x)^3\big]\Big] \\ & (j,n) \in \Omega\ ; \Delta \neq 0 \end{aligned} \tag{2.56}$$

By definition, the expression on the right side of Eq. (2.56) represents the truncation error of Eqs. (2.42) if the scheme Eqs. (2.42)–(2.44) are considered as a solver of the system of PDEs Eqs. (1.1) and (2.49). Here the values of α and β will be chosen so that the truncation error will reach the highest order. From Eq. (2.56), one can see that the coefficients of the three lowest-order terms in the truncation error vanish if

$$\alpha - \nu = 0 \quad \text{and} \quad \alpha(1+4\nu^2) - 3\beta\nu - 2\nu^3 = 0 \tag{2.57}$$

For the case $\nu \neq 0$, Eq. (2.57) $\Leftrightarrow$

$$\alpha = \nu \quad \text{and} \quad \beta = \frac{1+2\nu^2}{3} \tag{2.58}$$

Next the $a(3)$ scheme will be defined as the special solver with α and β being chosen according to Eq. (2.58).

2.5. The basic and forward marching forms of the $a(3)$ scheme

Assuming Eq. (2.58), Eqs. (2.37), (2.41)–(2.44) and (2.56) reduce to

$$\left[u + \nu u_{\bar{x}} + \frac{1+2\nu^2}{3} u_{\bar{x}\bar{x}}\right]_j^n = \left[u - \nu u_{\bar{x}} + \frac{1+2\nu^2}{3} u_{\bar{x}\bar{x}}\right]_j^{n-1} \tag{2.59}$$

$$\Delta = 2/3 \tag{2.60}$$

$$\begin{aligned} u_j^n = {} & 2\Big[u - \nu u_{\bar{x}} + \frac{1+2\nu^2}{3} u_{\bar{x}\bar{x}}\Big]_j^{n-1} - \frac{1+\nu}{2}\left[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3} u_{\bar{x}\bar{x}}\right]_{j+1}^{n-1} \\ & - \frac{1-\nu}{2}\left[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3} u_{\bar{x}\bar{x}}\right]_{j-1}^{n-1} \qquad (j,n) \in \Omega \end{aligned} \tag{2.61}$$

$$\begin{aligned} (u_{\bar{x}})_j^n = {} & 2\nu\Big[u - \nu u_{\bar{x}} + \frac{1+2\nu^2}{3} u_{\bar{x}\bar{x}}\Big]_j^{n-1} + \frac{1-2\nu}{2}\left[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3} u_{\bar{x}\bar{x}}\right]_{j+1}^{n-1} \\ & - \frac{1+2\nu}{2}\left[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3} u_{\bar{x}\bar{x}}\right]_{j-1}^{n-1} \qquad (j,n) \in \Omega \end{aligned} \tag{2.62}$$

$$
\begin{aligned}
(u_{\bar{x}\bar{x}})^n_j = {}& -3\big[u - \nu u_{\bar{x}} + \frac{1+2\nu^2}{3} u_{\bar{x}\bar{x}}\big]^{n-1}_j + \frac{3}{2}\Big[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3} u_{\bar{x}\bar{x}}\Big]^{n-1}_{j+1} \\
& + \frac{3}{2}\Big[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3} u_{\bar{x}\bar{x}}\Big]^{n-1}_{j-1} \qquad (j,n) \in \Omega
\end{aligned}
\tag{2.63}
$$

and

$$
\begin{aligned}
(e_1)^n_j = {}& \frac{1}{24}\left(\frac{\partial^4 \tilde{u}}{\partial x^4}\right)^n_j \Big[\frac{(\Delta x)^4}{\Delta t} + 2a^2 \Delta t (\Delta x)^2 + a^4 (\Delta t)^3\Big] + O\big[(\Delta t)^4\big] + O\big[(\Delta x)^4\big] \\
& + O\big[\Delta t (\Delta x)^3\big] + O\big[(\Delta t)^2 (\Delta x)^2\big] + \frac{O\big[(\Delta x)^5\big]}{\Delta t} \qquad (j,n) \in \Omega
\end{aligned}
\tag{2.64}
$$

Note that: (i) the forms of the last four terms in Eq. (2.64) have been simplified using the definition $\nu = a\Delta t/\Delta x$; and (ii) the expression on the right side of Eq. (2.64) represents the truncation error of Eq. (2.61) if the scheme formed by Eqs. (2.61)–(2.63) is considered as a solver of the system of PDEs Eqs. (1.1) and (2.49).

Next we convert Eq. (2.62) into its analytical form by replacing (i) u^n_j, $(u_x)^n_j$, and $(u_{xx})^n_j$, respectively, with $\tilde{u}^n_j$, $\tilde{v}^n_j$, and $\tilde{w}^n_j$, for each (j,n); and (ii) the index n with $n+1$ everywhere. By using (i) Eq. (2.13), (ii) $\nu = a\Delta t/\Delta x$, and (iii) the fact that $(j, n+1) \in \Omega \Leftrightarrow (j,n) \in \Omega$, then after a normalization by the factor 1/2, the analytical form can be expressed as

$$
\begin{aligned}
(e_2)^n_j \overset{\text{def}}{=} {}& \frac{1}{2}\tilde{v}^{n+1}_j - \frac{2a\Delta t}{(\Delta x)^2}\Big[\tilde{u} - \frac{a\Delta t}{2}\tilde{v} + \frac{(\Delta x)^2 + 2a^2(\Delta t)^2}{12}\tilde{w}\Big]^n_j \\
& - \frac{1}{2\Delta x}\Big(1 - \frac{2a\Delta t}{\Delta x}\Big)\Big[\tilde{u} - \frac{\Delta x + a\Delta t}{2}\tilde{v} + \frac{(\Delta x)^2 + a\Delta t\Delta x + a^2(\Delta t)^2}{6}\tilde{w}\Big]^n_{j+1} \\
& + \frac{1}{2\Delta x}\Big(1 + \frac{2a\Delta t}{\Delta x}\Big)\Big[\tilde{u} + \frac{\Delta x - a\Delta t}{2}\tilde{v} + \frac{(\Delta x)^2 - a\Delta t\Delta x + a^2(\Delta t)^2}{6}\tilde{w}\Big]^n_{j-1} = 0 \qquad (j,n) \in \Omega
\end{aligned}
\tag{2.65}
$$

Similarly, the analytical form of Eq. (2.63) can be expressed as

$$
\begin{aligned}
(e_3)^n_j \overset{\text{def}}{=} {}& \frac{1}{6}\tilde{w}^{n+1}_j + \frac{2}{(\Delta x)^2}\Big[\tilde{u} - \frac{a\Delta t}{2}\tilde{v} + \frac{(\Delta x)^2 + 2a^2(\Delta t)^2}{12}\tilde{w}\Big]^n_j \\
& - \frac{1}{(\Delta x)^2}\Big[\tilde{u} - \frac{\Delta x + a\Delta t}{2}\tilde{v} + \frac{(\Delta x)^2 + a\Delta t\Delta x + a^2(\Delta t)^2}{6}\tilde{w}\Big]^n_{j+1} \\
& - \frac{1}{(\Delta x)^2}\Big[\tilde{u} + \frac{\Delta x - a\Delta t}{2}\tilde{v} + \frac{(\Delta x)^2 - a\Delta t\Delta x + a^2(\Delta t)^2}{6}\tilde{w}\Big]^n_{j-1} = 0 \qquad (j,n) \in \Omega
\end{aligned}
\tag{2.66}
$$

By using Taylor's formula and Eq. (2.45), Eqs. (2.65) and (2.66) imply that

$$
\begin{aligned}
(e_2)^n_j = {}& \Big\{\hat{v} + \Big[\frac{\partial \hat{v}}{\partial t} - a\frac{\partial \hat{v}}{\partial x} + \frac{\partial}{\partial x}\Big(\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\Big)\Big]\frac{\Delta t}{2} + \frac{\partial^2 \hat{v}}{\partial t^2}\frac{(\Delta t)^2}{4} + \frac{\partial^2 \hat{v}}{\partial x^2}\frac{\big[(\Delta x)^2 - 2a^2(\Delta t)^2\big]}{4} \\
& + \frac{\partial \hat{w}}{\partial x}\frac{\big[a^2(\Delta t)^2 - (\Delta x)^2\big]}{6} + \frac{\partial}{\partial x}\Big(\frac{\partial^2 \tilde{u}}{\partial t^2} - a^2\frac{\partial^2 \tilde{u}}{\partial x^2}\Big)\frac{(\Delta t)^2}{4} - \frac{(1+\nu^2)}{12}\frac{\partial^3 \tilde{u}}{\partial x^3}(\Delta x)^2\Big\}^n_j \\
& + O\big[(\Delta t)^3\big] + O\big[(\Delta t)^2 \Delta x\big] + O\big[\Delta t(\Delta x)^2\big] + O\big[(\Delta x)^3\big] \qquad (j,n) \in \Omega
\end{aligned}
\tag{2.67}
$$

and

$$
\begin{aligned}
(e_3)_j^n = \Big\{ & \frac{\partial \hat{v}}{\partial x} + \Big[\frac{\partial^2}{\partial x^2}\Big(\frac{\partial \tilde{u}}{\partial t} + a\frac{\partial \tilde{u}}{\partial x}\Big) + \frac{\partial \hat{w}}{\partial t} - 2a\frac{\partial \hat{w}}{\partial x} + 3a\frac{\partial^2 \hat{v}}{\partial x^2}\Big]\frac{\Delta t}{6} \\
& + \Big[\frac{\partial^2}{\partial x^2}\Big(\frac{\partial^2 \tilde{u}}{\partial t^2} - a^2\frac{\partial^2 \tilde{u}}{\partial x^2}\Big) + \frac{\partial^2 \hat{w}}{\partial t^2} - 2a^2\frac{\partial^2 \hat{w}}{\partial x^2}\Big]\frac{(\Delta t)^2}{12} + \Big(\frac{\partial^3 \hat{v}}{\partial x^3} - \frac{\partial^2 \hat{w}}{\partial x^2}\Big)\frac{(\Delta x)^2}{6} \qquad (j,n) \in \Omega \qquad (2.68) \\
& - \frac{(1+\nu^2)}{12}\frac{\partial^4 \tilde{u}}{\partial x^4}(\Delta x)^2 \Big\}_j^n + O\big[(\Delta t)^3\big] + O\big[(\Delta t)^2 \Delta x\big] + O\big[\Delta t (\Delta x)^2\big] + O\big[(\Delta x)^3\big]
\end{aligned}
$$

Assuming Eqs. (2.51) and (2.55), $(e_2)_j^n$ and $(e_3)_j^n$ reduce to

$$
(e_2)_j^n = -\left(\frac{\partial^3 \tilde{u}}{\partial x^3}\right)_j^n \frac{\big[(\Delta x)^2 + a^2(\Delta t)^2\big]}{12} + O\big[(\Delta t)^3\big] + O\big[(\Delta t)^2 \Delta x\big] + O\big[\Delta t (\Delta x)^2\big] + O\big[(\Delta x)^3\big] \quad (j,n) \in \Omega \quad (2.69)
$$

and

$$
(e_3)_j^n = -\left(\frac{\partial^4 \tilde{u}}{\partial x^4}\right)_j^n \frac{\big[(\Delta x)^2 + a^2(\Delta t)^2\big]}{12} + O\big[(\Delta t)^3\big] + O\big[(\Delta t)^2 \Delta x\big] + O\big[\Delta t (\Delta x)^2\big] + O\big[(\Delta x)^3\big] \quad (j,n) \in \Omega \quad (2.70)
$$

respectively.

Hereafter, for any ν, let *the system of equations* defined by Eqs. (2.14), (2.15), and (2.59) be referred to as the basic form of the $a(3)$ scheme while that defined by Eqs. (2.61)–(2.63) be referred to as the forward marching form of the $a(3)$ scheme. Because (i) Eqs. (2.14), (2.15), and (2.37) $\Leftrightarrow$ Eqs. (2.42)–(2.44) if $\Delta \neq 0$, and (ii) Eqs. (2.59)–(2.63) are special cases of Eqs. (2.37), and (2.41)–(2.44), respectively, the basic form of the $a(3)$ scheme $\Leftrightarrow$ its forward marching form. Thus the essential conditions represented by these or other equivalent forms may be referred to simply as the $a(3)$ scheme.

With the above definitions, the expressions on the right sides of Eqs. (2.64), (2.69) and (2.70) represent the truncation errors of Eqs. (2.61)–(2.63), respectively, if the forward marching form of the $a(3)$ scheme is considered as a solver of the system of PDEs Eqs. (1.1) and (2.49). According to Eqs. (2.69) and (2.70), $(e_2)_j^n \to 0$ and $(e_3)_j^n \to 0$ as $\Delta t, \Delta x \to 0$, regardless how Δt and Δx are related when $\Delta t, \Delta x \to 0$. On the other hand, Eq. (2.64) implies that $(e_1)_j^n \to 0$ as $\Delta t, \Delta x \to 0$ only if the mesh refinement procedure is subjected to the condition

$$
\frac{(\Delta x)^4}{\Delta t} \to 0 \quad \text{as} \quad \Delta t, \Delta x \to 0 \qquad (2.71)
$$

Thus the $a(3)$ scheme is consistent with the system of PDEs Eqs. (1.1) and (2.49) if and only if Eq. (2.71) is satisfied.

At this juncture, we offer the following remarks:

(a) Let $\Delta t/\Delta x$ be held as constant as $\Delta t, \Delta x \to 0$. Then for this mesh refinement procedure, Eqs. (2.64), (2.69), and (2.70) imply that the truncation errors for Eqs. (2.61)–(2.63), respectively, are third order, second order, and second order in Δt and Δx. On the other hand, according to the numerical results presented in Sec. 4, the $a(3)$ scheme generally is 4th order in accuracy for both u_j^n and $(u_x)_j^n$ while only 2nd order in accuracy for $(u_{xx})_j^n$. *Note that order of truncation error and order of accuracy represent total different concepts (see Secs. 5 and 6 in [1]). The former is a measure of how well an analytical solution satisfies the discrete scheme while the latter represents a measure of how well a solution to the discrete scheme approximates the corresponding analytical solution. Thus the numerical results presented in Sec. 4 do not contradict the conclusion reached here.*

(b) Because (i) each of the two decoupled subsystems in each of Eqs. (2.14) and (2.15) is PT invariant by itself, and (ii) the system Eq. (2.37) is also PT invariant if α and β are parameters independent of (j,n), by the definition of PT invariance one can easily see that the basic form of the $a(3)$ scheme is PT invariant.

(c) Let $\vec{q}(j,n) = \vec{q}_o(j,n)$, $(j,n) \in \Omega$, be a solution to the basic form. Then, by substituting $\vec{q}(j,n) = \vec{q}_o(j,n)$ into the basic form, one obtains a system of identities involving $\vec{q}_o(j,n)$, $(j,n) \in \Omega$. Due to the PT invariance of the basic form, the above system of identities is equivalent to that obtained by substituting $\vec{q}(j,n) = U\vec{q}_o(-j,-n)$ into the basic form. As such $\vec{q}(j,n) = \vec{q}_o(j,n)$, $(j,n) \in \Omega$, represent a solution to the basic form $\Leftrightarrow$ $\vec{q}(j,n) = U\vec{q}_o(-j,-n)$, $(j,n) \in \Omega$, represent another solution to the basic form. In other words, *the PT image of a solution to the basic form is also a solution and vice versa*. Obviously this conclusion is valid for other PT invariant forms of the $a(3)$ scheme.

Next, the forward marching form Eqs. (2.61)–(2.62) will be cast into a matrix form. Let

$$\vec{c}_0(\nu) \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ -\nu \\ (1+2\nu^2)/3 \end{pmatrix}, \quad \vec{c}_+(\nu) \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ -(1+\nu) \\ (2/3)(1+\nu+\nu^2) \end{pmatrix}, \quad \vec{c}_-(\nu) \stackrel{\text{def}}{=} \begin{pmatrix} 1 \\ 1-\nu \\ (2/3)(1-\nu+\nu^2) \end{pmatrix} \tag{2.72}$$

$$\vec{d}_0(\nu) \stackrel{\text{def}}{=} \begin{pmatrix} 2 \\ 2\nu \\ -3 \end{pmatrix}, \quad \vec{d}_+(\nu) \stackrel{\text{def}}{=} \begin{pmatrix} -(1+\nu)/2 \\ (1-2\nu)/2 \\ (3/2) \end{pmatrix}, \quad \vec{d}_-(\nu) \stackrel{\text{def}}{=} \begin{pmatrix} -(1-\nu)/2 \\ -(1+2\nu)/2 \\ (3/2) \end{pmatrix} \tag{2.73}$$

$$Q_0(\nu) \stackrel{\text{def}}{=} \vec{d}_0(\nu)\,[\vec{c}_0(\nu)]^t = \begin{pmatrix} 2 & -2\nu & (2/3)(1+2\nu^2) \\ 2\nu & -2\nu^2 & (2/3)\nu(1+2\nu^2) \\ -3 & 3\nu & -(1+2\nu^2) \end{pmatrix} \tag{2.74}$$

$$Q_+(\nu) \stackrel{\text{def}}{=} \vec{d}_+(\nu)\,[\vec{c}_+(\nu)]^t = \begin{pmatrix} -(1+\nu)/2 & (1+\nu)^2/2 & -(1+\nu)(1+\nu+\nu^2)/3 \\ (1-2\nu)/2 & -(1-2\nu)(1+\nu)/2 & (1-2\nu)(1+\nu+\nu^2)/3 \\ 3/2 & -(3/2)(1+\nu) & 1+\nu+\nu^2 \end{pmatrix} \tag{2.75}$$

and

$$Q_-(\nu) \stackrel{\text{def}}{=} \vec{d}_-(\nu)\,[\vec{c}_-(\nu)]^t = \begin{pmatrix} -(1-\nu)/2 & -(1-\nu)^2/2 & -(1-\nu)(1-\nu+\nu^2)/3 \\ -(1+2\nu)/2 & -(1+2\nu)(1-\nu)/2 & -(1+2\nu)(1-\nu+\nu^2)/3 \\ 3/2 & (3/2)(1-\nu) & 1-\nu+\nu^2 \end{pmatrix} \tag{2.76}$$

Hereafter $\vec{c}^{\,t}$ denote the transpose of any column or row matrix $\vec{c}$. By using Eqs. (2.31) and (2.74)–(2.76), the forward marching form can be cast into the matrix form:

$$\vec{q}(j,n) = Q_0(\nu)\vec{q}(j,n-1) + Q_+(\nu)\vec{q}(j+1,n-1) + Q_-(\nu)\vec{q}(j-1,n-1), \qquad (j,n) \in \Omega \tag{2.77}$$

Here the reader is warned that the notations $Q_+(\nu)$ and $Q_-(\nu)$ used in earlier CESE papers are now replaced by $Q_-(\nu)$ and $Q_+(\nu)$, respectively. As such, *the terms $Q_-(\nu)\vec{q}(j+1,n-1)$ and $Q_+(\nu)(j-1,n-1)$ in Eq. (3.48) of [71] appear here as $Q_+(\nu)\vec{q}(j+1,n-1)$ and $Q_-(\nu)(j-1,n-1)$, respectively*. Also note that each of $Q_0(\nu)$, $Q_+(\nu)$, and $Q_-(\nu)$ is in the form of $\vec{d}\,\vec{c}^{\,t}$ where $\vec{c}$ and $\vec{d}$ are 3×1 column matrix. Thus each is a matrix of rank one (see pp. 80-82 in [74]). Rank-one matrices are singular and have many interesting properties. As an example, the eigenvalues of $Q_0(\nu)$ are 0, 0, and $[\vec{c}_0(\nu)]^t\,\vec{d}_0(\nu)$ with $\vec{d}_0(\nu)$ being the eigenvector of the last eigenvalue.

To facilitate the proof of the PT invariance of the forward marching form, first we will introduce some basic concept. Note that, *for any set of variables x_ℓ, y_ℓ, $\ell = 1,2$*, the conditions

$$x_1 + y_1 = x_2 - y_2 \quad \text{and} \quad x_1 - y_1 = x_2 + y_2 \tag{2.78}$$

$\Leftrightarrow$

$$x_1 = x_2 \quad \text{and} \quad y_1 = -y_2 \tag{2.79}$$

Thus, the image of Eq. (2.78) under *any* one-to-one mapping

$$(x_\ell, y_\ell) \leftrightarrow (x'_\ell, y'_\ell), \qquad \ell = 1,2 \tag{2.80}$$

i.e.,

$$x_1' + y_1' = x_2' - y_2' \quad \text{and} \quad x_1' - y_1' = x_2' + y_2' \tag{2.81}$$

$\Leftrightarrow$ the image of Eq. (2.79) under the same mapping, i.e.,

$$x_1' = x_2' \quad \text{and} \quad y_1' = -y_2' \tag{2.82}$$

where the variables x_ℓ' and y_ℓ', $\ell = 1, 2$, may or may not be related to x_ℓ, y_ℓ, $\ell = 1, 2$. Moreover, in case that these two sets of variables are related, the condition Eq. (2.78) (or its equivalent Eq. (2.79)) may or may not be equivalent to the condition Eq. (2.81) (or its equivalent Eq. (2.82)). If the mapping Eq. (2.80) is such that Eq. (2.78) $\Leftrightarrow$ the image under this mapping (i.e., Eq. (2.81)), then Eq. (2.79) (the equivalent of Eq. (2.78)) $\Leftrightarrow$ Eq. (2.82) (the equivalent of Eq. (2.81)). Eq. (2.80) with $x_\ell' = x_\ell$ and $y_\ell' = y_\ell$, $\ell = 1, 2$, is an example of such mapping while Eq. (2.80) with $x_\ell' = y_\ell$ and $y_\ell' = x_\ell$, $\ell = 1, 2$, is not.

To prove the PT invariance of the forward marching form, Note that: (i) the basic form of the $a(3)$ scheme $\Leftrightarrow$ its forward marching form for *any choice* of $\vec{q}(j,n)$, $(j,n) \in \Omega$; and (ii) the PT images of the basic and forward marching forms, respectively, are obtained from the basic and forward marching forms through the mapping Eq. (2.30), i.e., through replacing $\vec{q}(j,n)$ in the basic form and the forward marching form with $U\vec{q}(-j,-n)$, $(j,n) \in \Omega$. From the above observations and the illustration given in the last paragraph, one concludes that the PT image of the basic form $\Leftrightarrow$ that of the forward marching form. Because the basic form is PT invariant, i.e., the PT image of the basic form $\Leftrightarrow$ the basic form itself, Now we arrive at the conclusion that the forward marching form $\Leftrightarrow$ the basic form $\Leftrightarrow$ the PT image of the basic form $\Leftrightarrow$ the PT image of the forward marching form. Thus the forward marching form $\Leftrightarrow$ its PT image, i.e., the forward marching form is PT invariant. QED.

With the above preliminaries, the backward marching form of the $a(3)$ scheme will be developed in Sec. 2.6.

2.6. The backward marching forms of the $a(3)$ scheme

The PT invariance of the forward marching form of the $a(3)$ scheme implies that Eq. (2.77) $\Leftrightarrow$ its PT image, i.e.,

$$U\vec{q}(-j,-n) = Q_0(\nu)U\vec{q}(-j,-n+1) + Q_+(\nu)U\vec{q}(-j-1,-n+1) + Q_-(\nu)U\vec{q}(-j+1,-n+1), \quad (j,n) \in \Omega \tag{2.83}$$

Moreover, by multiplying Eq. (2.83) from left using the matrix U and using Eq. (2.33), one concludes that Eq. (2.83) $\Leftrightarrow$

$$\vec{q}(-j,-n) = \hat{Q}_0(\nu)\vec{q}(-j,-n+1) + \hat{Q}_-(\nu)\vec{q}(-j-1,-n+1) + \hat{Q}_+(\nu)\vec{q}(-j+1,-n+1), \qquad (j,n) \in \Omega \tag{2.84}$$

where

$$\hat{Q}_0(\nu) \overset{\text{def}}{=} UQ_0(\nu)U = \begin{pmatrix} 2 & 2\nu & (2/3)(1+2\nu^2) \\ -2\nu & -2\nu^2 & -(2/3)\nu(1+2\nu^2) \\ -3 & -3\nu & -(1+2\nu^2) \end{pmatrix} \tag{2.85}$$

$$\hat{Q}_-(\nu) \overset{\text{def}}{=} UQ_+(\nu)U = \begin{pmatrix} -(1+\nu)/2 & -(1+\nu)^2/2 & -(1+\nu)(1+\nu+\nu^2)/3 \\ -(1-2\nu)/2 & -(1-2\nu)(1+\nu)/2 & -(1-2\nu)(1+\nu+\nu^2)/3 \\ 3/2 & (3/2)(1+\nu) & 1+\nu+\nu^2 \end{pmatrix} \tag{2.86}$$

and

$$\hat{Q}_+(\nu) \overset{\text{def}}{=} UQ_-(\nu)U = \begin{pmatrix} -(1-\nu)/2 & (1-\nu)^2/2 & -(1-\nu)(1-\nu+\nu^2)/3 \\ (1+2\nu)/2 & -(1+2\nu)(1-\nu)/2 & (1+2\nu)(1-\nu+\nu^2)/3 \\ 3/2 & -(3/2)(1-\nu) & 1-\nu+\nu^2 \end{pmatrix} \tag{2.87}$$

By replacing the "dummy" indices $-j$ and $-n$ everywhere in Eq. (2.84) with j and n, respectively, one can see that the system Eq. (2.84) is identical to the system

$$\vec{q}(j,n) = \hat{Q}_0(\nu)\vec{q}(j,n+1) + \hat{Q}_+(\nu)\vec{q}(j+1,n+1) + \hat{Q}_-(\nu)\vec{q}(j-1,n+1), \quad (j,n) \in \Omega \tag{2.88}$$

Because the mesh variables at (j, n) can be determined in terms of those at $(j-1, n+1)$, $(j, n+1)$, and $(j+1, n+1)$ using Eq. (2.88), hereafter Eq. (2.88) (which is equivalent to other forms of the $a(3)$ scheme) will be referred to as the backward marching form of the $a(3)$ scheme.

According to Eqs. (2.74) and (2.85), $\hat{Q}_0(\nu) = U\vec{d}_0(\nu)\,[\vec{c}_0(\nu)]^t\,U$. Because $U\vec{d}_0(\nu)$ and $[\vec{c}_0(\nu)]^t\,U$ are 3×1 column matrix and 1×3 row matrix, respectively, $\hat{Q}_0(\nu)$ is a rank-one matrix. Similarly, $\hat{Q}_-(\nu)$ and $\hat{Q}_+(\nu)$ are also rank-one matrices.

Eq. (2.88) was derived using the PT invariance of the forward marching form of the $a(3)$ scheme. Alternatively, it can also be derived from the basic form. To proceed, note that: (i) by replacing the indices j and n everywhere in Eq. (2.14) with $j+1$ and $n+1$ and using the fact that $(j,n)\in\Omega \Leftrightarrow (j-1,n-1)\in\Omega \Leftrightarrow (j+1,n+1)\in\Omega$, one can see that the system Eq. (2.14) is identical to the system

$$\left[u + (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_j^n = \left[u - (1-\nu)u_{\bar{x}} + \frac{2(1-\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j+1}^{n+1}, \quad (j,n)\in\Omega \tag{2.89}$$

(ii) by replacing the indices j and n everywhere in Eq. (2.15) with $j-1$ and $n+1$ and using the fact that $(j,n)\in\Omega \Leftrightarrow (j+1,n-1)\in\Omega \Leftrightarrow (j-1,n+1)\in\Omega$, one can see that the system Eq. (2.15) is identical to the system

$$\left[u - (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_j^n = \left[u + (1+\nu)u_{\bar{x}} + \frac{2(1+\nu+\nu^2)}{3}u_{\bar{x}\bar{x}}\right]_{j-1}^{n+1}, \quad (j,n)\in\Omega \tag{2.90}$$

and (iii) by replacing the index n everywhere in Eq. (2.59) with $n+1$ and using the fact that $(j,n)\in\Omega \Leftrightarrow (j,n-1)\in\Omega \Leftrightarrow (j,n+1)\in\Omega$, one can see that the system Eq. (2.59) is identical to the system

$$\left[u - \nu u_{\bar{x}} + \frac{1+2\nu^2}{3}u_{\bar{x}\bar{x}}\right]_j^n = \left[u + \nu u_{\bar{x}} + \frac{1+2\nu^2}{3}u_{\bar{x}\bar{x}}\right]_j^{n+1} \qquad (j,n)\in\Omega \tag{2.91}$$

As such the system Eqs. (2.89)–(2.91) are identical to Eqs. (2.14), (2.15), (2.59), respectively.

For *each* $(j,n)\in\Omega$, Eqs. (2.14), (2.15), and (2.59) form a linear system of three equations for the three mesh variables u_j^n, $(u_{\bar{x}})_j^n$, and $(u_{\bar{x}\bar{x}})_j^n$. Eqs. (2.89)–(2.91) form another system. Moreover, one can see that, under the mesh variable mapping

$$\begin{aligned} &\vec{q}(j,n) \leftrightarrow U\vec{q}(j,n), \quad \vec{q}(j,n-1) \leftrightarrow \vec{q}(j,n+1), \\ &\vec{q}(j+1,n-1) \leftrightarrow U\vec{q}(j-1,n+1), \quad \text{and} \quad \vec{q}(j-1,n-1) \leftrightarrow \vec{q}(j+1,n+1) \end{aligned} \tag{2.92}$$

Eqs. (2.89)–(2.91), respectively, are the images of Eqs. (2.14), (2.15), and (2.59) and vice versa. By using the concept introduced earlier in a discussion involving Eqs. (2.78)–(2.82), one concludes that the solution to Eqs. (2.89)–(2.91) must be the image of Eq. (2.77) (i.e., the solution to Eqs. (2.14), (2.15) and (2.59)) under the same mapping. In other words, the solution to Eqs. (2.89)–(2.91) is

$$U\vec{q}(j,n) = Q_0(\nu)U\vec{q}(j,n+1) + Q_+(\nu)U\vec{q}(j-1,n+1) + Q_-(\nu)U\vec{q}(j+1,n+1), \qquad (j,n)\in\Omega \tag{2.93}$$

By multiplying Eq. (2.93) from left using the matrix U and using Eqs. (2.33) and (2.85)–(2.87), one has Eq. (2.88). QED.

As a preliminary for the developments in Sec. 3, in the following, important algebraic relations involving $Q_0(\nu)$, $Q_+(\nu)$, $Q_-(\nu)$, $\hat{Q}_0(\nu)$, $\hat{Q}_+(\nu)$, and $\hat{Q}_-(\nu)$ will be extracted from the PT invariance of the $a(3)$ scheme.

2.7. Algebraic relations associated with PT invariance

Let $(j_o,n_o)\in\Omega$ be any given fixed mesh point. Let $\vec{q}(j_o,n_o)$, $\vec{q}(j_o\pm 1,n_o)$, and $\vec{q}(j_o\pm 2,n_o)$, respectively, be the arbitrary initial data specified at (j_o,n_o), $(j_o\pm 1,n_o)$, and $(j_o\pm 2,n_o)$, respectively. Let $\vec{q}(j_o,n_o+1)$, and $\vec{q}(j_o \pm 1, n_o+1)$ be specified in terms of the mesh variables at the n_oth time level using the forward marching form Eq. (2.77), i.e.,

$$\vec{q}(j_o, n_o+1) = Q_0(\nu)\,\vec{q}(j_o, n_o) + Q_+(\nu)\,\vec{q}(j_o+1, n_o) + Q_-(\nu)\,\vec{q}(j_o-1, n_o) \tag{2.94}$$

$$\vec{q}(j_o+1, n_o+1) = Q_0(\nu)\,\vec{q}(j_o+1, n_o) + Q_+(\nu)\,\vec{q}(j_o+2, n_o) + Q_-(\nu)\,\vec{q}(j_o, n_o) \tag{2.95}$$

and

$$\vec{q}(j_o-1, n_o+1) = Q_0(\nu)\,\vec{q}(j_o-1, n_o) + Q_+(\nu)\,\vec{q}(j_o, n_o) + Q_-(\nu)\,\vec{q}(j_o-2, n_o) \tag{2.96}$$

On the other hand, because Eq. (2.77) ⇔ Eq. (2.88), $\vec{q}(j_o, n_o+1)$, $\vec{q}(j_o \pm 1, n_o+1)$, and $\vec{q}(j_o, n_o)$ must also be linked by Eq. (2.88), i.e.,

$$\vec{q}(j_o, n_o) = \hat{Q}_0(\nu)\;\vec{q}(j_o, n_o+1) + \hat{Q}_+(\nu)\,\vec{q}(j_o+1, n_o+1) + \hat{Q}_-(\nu)\,\vec{q}(j_o-1, n_o+1) \tag{2.97}$$

Substituting Eqs. (2.94)–(2.96) into (2.97), one has

$$\begin{aligned}
&\big[\hat{Q}_0(\nu)Q_0(\nu) + \hat{Q}_+(\nu)Q_-(\nu) + \hat{Q}_-(\nu)Q_+(\nu) - I\big]\vec{q}(j_o, n_o)\\
&+ \big[\hat{Q}_0(\nu)Q_+(\nu) + \hat{Q}_+(\nu)Q_0(\nu)\big]\vec{q}(j_o+1, n_o) + \big[\hat{Q}_0(\nu)Q_-(\nu) + \hat{Q}_-(\nu)Q_0(\nu)\big]\vec{q}(j_o-1, n_o)\\
&+ \hat{Q}_+(\nu)Q_+(\nu)\,\vec{q}(j_o+2, n_o) + \hat{Q}_-(\nu)Q_-(\nu)\,\vec{q}(j_o-2, n_o) = \vec{0}
\end{aligned} \tag{2.98}$$

where I is the 3×3 identity matrix and $\vec{0}$ is the 3×1 null column matrix.

Because Eq. (2.98) must be valid for any choice of $\vec{q}(j_o, n_o)$, $\vec{q}(j_o \pm 1, n_o)$, and $\vec{q}(j_o \pm 2, n_o)$, the coefficients matrices in front of these column matrices must be null identically, i.e.,

$$\hat{Q}_0(\nu)Q_0(\nu) + \hat{Q}_+(\nu)Q_-(\nu) + \hat{Q}_-(\nu)Q_+(\nu) = I \tag{2.99}$$

$$\hat{Q}_0(\nu)Q_+(\nu) + \hat{Q}_+(\nu)Q_0(\nu) = \mathbf{0} \tag{2.100}$$

$$\hat{Q}_0(\nu)Q_-(\nu) + \hat{Q}_-(\nu)Q_0(\nu) = \mathbf{0} \tag{2.101}$$

$$\hat{Q}_+(\nu)Q_+(\nu) = \mathbf{0} \tag{2.102}$$

and

$$\hat{Q}_-(\nu)Q_-(\nu) = \mathbf{0} \tag{2.103}$$

where $\mathbf{0}$ is the 3×3 null matrix. As an example, one can prove Eq. (2.99) by substituting into Eq. (2.98) each of the following sets of the initial data: (i) $\vec{q}(j_o \pm 1, n_o) = \vec{q}(j_o \pm 2, n_o) = \vec{0}$ and $\vec{q}(j_o, n_o) = (1,0,0)^t$, (ii) $\vec{q}(j_o \pm 1, n_o) = \vec{q}(j_o \pm 2, n_o) = \vec{0}$ and $\vec{q}(j_o, n_o) = (0,1,0)^t$, and (iii) $\vec{q}(j_o \pm 1, n_o) = \vec{q}(j_o \pm 2, n_o) = \vec{0}$ and $\vec{q}(j_o, n_o) = (0,0,1)^t$.

Similarly, by substituting the backward marching relations

$$\vec{q}(j_o, n_o-1) = \hat{Q}_0(\nu)\,\vec{q}(j_o, n_o) + \hat{Q}_+(\nu)\,\vec{q}(j_o+1, n_o) + \hat{Q}_-(\nu)\,\vec{q}(j_o-1, n_o) \tag{2.104}$$

$$\vec{q}(j_o+1, n_o-1) = \hat{Q}_0(\nu)\,\vec{q}(j_o+1, n_o) + \hat{Q}_+(\nu)\,\vec{q}(j_o+2, n_o) + \hat{Q}_-(\nu)\,\vec{q}(j_o, n_o) \tag{2.105}$$

and

$$\vec{q}(j_o-1, n_o-1) = \hat{Q}_0(\nu)\,\vec{q}(j_o-1, n_o) + \hat{Q}_+(\nu)\,\vec{q}(j_o, n_o) + \hat{Q}_-(\nu)\,\vec{q}(j_o-2, n_o) \tag{2.106}$$

into the forward marching relation

$$\vec{q}(j_o, n_o) = Q_0(\nu)\;\vec{q}(j_o, n_o-1) + Q_+(\nu)\,\vec{q}(j_o+1, n_o-1) + Q_-(\nu)\,\vec{q}(j_o-1, n_o-1) \tag{2.107}$$

one has

$$\begin{aligned}&\big[Q_0(\nu)\hat{Q}_0(\nu)+Q_+(\nu)\hat{Q}_-(\nu)+Q_-(\nu)\hat{Q}_+(\nu)-I\big]\vec{q}(j_o,n_o)\\&+\big[Q_0(\nu)\hat{Q}_+(\nu)+Q_+(\nu)\hat{Q}_0(\nu)\big]\vec{q}(j_o+1,n_o)+\big[Q_0(\nu)\hat{Q}_-(\nu)+Q_-(\nu)\hat{Q}_0(\nu)\big]\vec{q}(j_o-1,n_o)\\&+Q_+(\nu)\hat{Q}_+(\nu)\,\vec{q}(j_o+2,n_o)+Q_-(\nu)\hat{Q}_-(\nu)\,\vec{q}(j_o-2,n_o)=\vec{0}\end{aligned} \tag{2.108}$$

Because Eq. (2.107) must be valid for any choice of $\vec{q}(j_o,n_o)$, $\vec{q}(j_o\pm 1,n_o)$, and $\vec{q}(j_o\pm 2,n_o)$, one concludes that

$$Q_0(\nu)\hat{Q}_0(\nu)+Q_+(\nu)\hat{Q}_-(\nu)+Q_-(\nu)\hat{Q}_+(\nu)=I \tag{2.109}$$

$$Q_0(\nu)\hat{Q}_+(\nu)+Q_+(\nu)\hat{Q}_0(\nu)=\mathbf{0} \tag{2.110}$$

$$Q_0(\nu)\hat{Q}_-(\nu)+Q_-(\nu)\hat{Q}_0(\nu)=\mathbf{0} \tag{2.111}$$

$$Q_+(\nu)\hat{Q}_+(\nu)=\mathbf{0} \tag{2.112}$$

and

$$Q_-(\nu)\hat{Q}_-(\nu)=\mathbf{0} \tag{2.113}$$

By using Eqs. (2.32) and (2.85)–(2.87), it can be shown that: (i) Eq. (2.99) ⇔ Eq. (2.109) ⇔

$$Q_0(\nu)UQ_0(\nu)+Q_-(\nu)UQ_-(\nu)+Q_+(\nu)UQ_+(\nu)=U \tag{2.114}$$

(ii) Eq. (2.100) ⇔ Eq. (2.111) ⇔

$$Q_0(\nu)UQ_+(\nu)+Q_-(\nu)UQ_0(\nu)=\mathbf{0} \tag{2.115}$$

(iii) Eq. (2.101) ⇔ Eq. (2.110) ⇔

$$Q_0(\nu)UQ_-(\nu)+Q_+(\nu)UQ_0(\nu)=\mathbf{0} \tag{2.116}$$

(iv) Eq. (2.102) ⇔ Eq. (2.113) ⇔

$$Q_-(\nu)UQ_+(\nu)=\mathbf{0} \tag{2.117}$$

and (v) Eq. (2.103) ⇔ Eq. (2.112) ⇔

$$Q_+(\nu)UQ_-(\nu)=\mathbf{0} \tag{2.118}$$

2.8. Other invariant properties and related algebraic relations

By using Eqs. (2.32) and (2.74)–(2.76), one can show that

$$Q_0(-\nu)=UQ_0(\nu)U,\quad Q_-(-\nu)=UQ_+(\nu)U,\quad\text{and}\quad Q_+(-\nu)=UQ_-(\nu)U \tag{2.119}$$

By using Eqs. (2.85)–(2.87), one can also show that Eq. (2.119) ⇔

$$\hat{Q}_0(-\nu)=U\hat{Q}_0(\nu)U,\quad \hat{Q}_-(-\nu)=U\hat{Q}_+(\nu)U,\quad\text{and}\quad \hat{Q}_+(-\nu)=U\hat{Q}_-(\nu)U \tag{2.120}$$

As will be shown, the above relations are linked with other invariant properties of the $a(3)$ scheme.

Let the advection speed a in Eq. (1.1) be considered as a variable parameter. Let $u=u(x,t;a)$ be a solution to Eq. (1.1), in the domain $-\infty<x,t,a<+\infty$, i.e.,

$$\frac{\partial u(x,t;a)}{\partial t}+a\frac{\partial u(x,t;a)}{\partial x}\equiv 0,\qquad -\infty<x,t,a<+\infty \tag{2.121}$$

Let

$$x'\overset{\text{def}}{=}-x,\quad t'\overset{\text{def}}{=}t,\quad\text{and}\quad a'=-a,\qquad -\infty<x,t,a<+\infty \tag{2.122}$$

and

$$\hat{u}(x,t;a) \overset{\text{def}}{=} u(-x,t;-a) \tag{2.123}$$

Then (i) Eq. (2.121) $\Leftrightarrow$

$$\frac{\partial u(x',t';a')}{\partial t'} + a'\frac{\partial u(x',t';a')}{\partial x'} \equiv 0, \qquad -\infty < x', t', a' < +\infty \tag{2.124}$$

and (ii)

$$\frac{\partial}{\partial t'} = \frac{\partial}{\partial t} \quad \text{and} \quad \frac{\partial}{\partial x'} = -\frac{\partial}{\partial x} \tag{2.125}$$

Thus one concludes that Eq. (2.121) $\Leftrightarrow$

$$\frac{\partial \hat{u}(x,t;a)}{\partial t} + a\frac{\partial \hat{u}(x,t;a)}{\partial x} \equiv 0, \qquad -\infty < x, t < +\infty \tag{2.126}$$

In other words, if $u = u(x,t;a)$ is a solution to Eq. (1.1), so must be $u = \hat{u}(x,t;a)$ and vice versa. Because the one-to-one mapping

$$(x,t,a) \leftrightarrow (-x,t,-a), \qquad -\infty < x, t, a < +\infty \tag{2.127}$$

represents a combined spatial-reflection (parity) and advection direction reversal (ADR) operation, hereafter (i) a pair of functions such as u and $\hat{u}$ will be referred to as the PADR images of each other; and (ii) a PDE such as Eq. (1.1) is said to be PADR invariant if the PADR image of a solution is also a solution and vice versa.

Because $\nu = a\Delta t/\Delta x$, the numerical analogue of Eq. (2.127) is

$$(j,n) \leftrightarrow (-j,n) \quad \text{and} \quad \nu \leftrightarrow -\nu \tag{2.128}$$

Motivated by an argument similar to that leads to Eq. (2.30) for PT mapping, the PADR mapping for the $a(3)$ scheme is defined by

$$\vec{q}(j,n) \leftrightarrow U\vec{q}(-j,n) \quad \text{and} \quad \nu \leftrightarrow -\nu, \qquad (j,n) \in \Omega \tag{2.129}$$

Thus the PADR image Eq. (2.77) is

$$U\vec{q}(-j,n) = Q_0(-\nu)U\vec{q}(-j,n-1) + Q_+(-\nu)U\vec{q}(-j-1,n-1) + Q_-(-\nu)U\vec{q}(-j+1,n-1), \qquad (j,n) \in \Omega \tag{2.130}$$

By using Eqs. (2.32) and (2.119), it can be shown that Eq. (2.130) $\Leftrightarrow$

$$\vec{q}(-j,n) = Q_0(\nu)\vec{q}(-j,n-1) + Q_-(\nu)\vec{q}(-j-1,n-1) + Q_+(\nu)\vec{q}(-j+1,n-1), \qquad (j,n) \in \Omega \tag{2.131}$$

By replacing the dummy index $-j$ with j everywhere in Eq. (2.131) and using the fact that $(-j,n) \in \Omega \Leftrightarrow (j,n) \in \Omega$, one concludes that Eq. (2.131) $\Leftrightarrow$ Eq. (2.77). Thus Eq. (2.77) is PADR invariant, i.e., it is equivalent to its PADR image.

By exchanging the roles of x and t, one can define invariance under a combined time reversal and advection direction reversal operation. Because (i) this operation is equivalent to a PT operation followed by a PADR operation or vice versa, and (ii) Eq. (1.1) and the $a(3)$ scheme are invariant under both PT and PADR operations, one concludes that Eq. (1.1) and the $a(3)$ scheme are also invariant under the new operation. In fact, invariance of the $a(3)$ scheme under this new operation can be proved using Eq. (2.120) (which is equivalent to Eq. (2.119)).

3. von Neumann analysis

Let $G(\nu,\theta)$ be a 3×3 nonsingular complex matrix function of ν and the phase angle θ such that

$$\vec{q}(j,n) = e^{ij\theta}\left[G(\nu,\theta)\right]^n \vec{b}, \qquad (j,n)\in\Omega;\ -\infty<\nu<+\infty;\ -\pi<\theta\le\pi \qquad (i\equiv\sqrt{-1}) \tag{3.1}$$

is a solution to Eq. (2.77) for all possible complex constant 3×1 column matrices $\vec{b}$. Note that: (i) without any loss of generality, hereafter the domain of θ is limited to $-\pi<\theta\le\pi$ and, *unless specified otherwise*, this domain will be assumed implicitly; and (ii) because $[G(\nu,\theta)]^n \overset{\text{def}}{=} \left\{[G(\nu,\theta)]^{-1}\right\}^{|n|}$ for an integer $n<0$, $[G(\nu,\theta)]^n$ is not defined if $n<0$ unless $[G(\nu,\theta)]^{-1}$ exists, i.e., $G(\nu,\theta)$ is nonsingular. By substituting Eq. (3.1) into Eq. (2.77), one has

$$\left[G(\nu,\theta)-Q_0(\nu)-e^{i\theta}Q_+(\nu)-e^{-i\theta}Q_-(\nu)\right]\left[G(\nu,\theta)\right]^n\vec{b}=0, \qquad n=0,\pm1,\pm2,\ldots \tag{3.2}$$

Because (i) $[G(\nu,\theta)]^0 = I$, and (ii) $\vec{b}$ can be any complex constant 3×1 column matrix, Eq. (3.2) $\Leftrightarrow$

$$G(\nu,\theta)=Q_0(\nu)+e^{i\theta}Q_+(\nu)+e^{-i\theta}Q_-(\nu) \tag{3.3}$$

By definition, $G(\nu,\theta)$ is the amplification matrix of the forward marching form of the $a(3)$ scheme. Because $Q_0(\nu)$, $Q_+(\nu)$, and $Q_-(\nu)$ are real matrices, Eq. (3.3) implies that

$$G(\nu,-\theta)=\overline{G(\nu,\theta)} \tag{3.4}$$

Hereafter $\overline{M}$ denotes the complex conjugate of any matrix M. Also, with the aid of Eq. (2.119) and the relation $U=U^{-1}$, one has

$$G(-\nu,\theta)=UG(\nu,-\theta)U=UG(\nu,-\theta)U^{-1} \tag{3.5}$$

At this juncture, some comments on the dual roles played by the amplification matrix $G(\nu,\theta)$ in determining the accuracy and the stability of the $a(3)$ scheme are in order. Note that the von Neumann analysis represents essentially a rigorous discrete Fourier analysis performed for a Fourier mode of a solution to a linear marching scheme such as Eq. (2.77) *assuming periodic spatial boundary conditions* (see Sec. 4 in [1]). The only difference between them is that the parameter θ (which specifies a Fourier mode) in the von Neumann analysis can assume any value in the domain $-\pi<\theta\le\pi$ while that in the discrete Fourier analysis can only assume a set of K uniformly distributed discrete values within the domain $-\pi<\theta\le\pi$ if the spatial domain is divided into K uniform mesh intervals. As such, the time evolution and therefore the accuracy of a Fourier mode of a solution to a linear scheme assuming periodic spatial boundary conditions can be determined using the corresponding amplification matrix (see Sec. 5 in [1]). Moreover, because a linear combination of solutions to a linear marching scheme is also a solution by itself, the time evolution of any Fourier mode of the *round-off errors* originally introduced during any marching step is also governed by the linear scheme and therefore it can also be determined using the amplification factor. As a result of this consideration and the facts that: (i) a scheme is stable if and only if these round-off errors will not be amplified without bound after many marching steps, and (ii) the spectrum of round-off errors generally covers all possible Fourier modes, i.e., all possible discrete values of θ, one concludes that, *assuming periodic spatial boundary conditions, the $a(3)$ scheme is stable for a given ν if and only if, for all possible K discrete values of θ in the domain $-\pi<\theta\le\pi$, every element of the matrix $[G(\nu,\theta)]^m$ remains bounded as the positive integer $m\to+\infty$.* Because the distribution of the allowed discrete values of θ becomes very dense in the domain $-\pi<\theta\le\pi$ for a large K, for simplicity, the K discrete values of θ referred to in the above stability definition is replaced by all values of θ in the domain $-\pi<\theta\le\pi$ in Definition 1 of Sec. 3.9.

In the following, we will show that the $a(3)$ scheme must be neutrally stable when it is stable.

3.1. Neutral stability of the $a(3)$ scheme

By using Eqs. (2.114)–(2.118), one can show easily that

$$\begin{aligned}&U\left[Q_0(\nu)+e^{i\theta}Q_-(\nu)+e^{-i\theta}Q_+(\nu)\right]U\left[Q_0(\nu)+e^{i\theta}Q_+(\nu)+e^{-i\theta}Q_-(\nu)\right]\\&=\left[Q_0(\nu)+e^{i\theta}Q_+(\nu)+e^{-i\theta}Q_-(\nu)\right]U\left[Q_0(\nu)+e^{i\theta}Q_-(\nu)+e^{-i\theta}Q_+(\nu)\right]U=I\end{aligned}\tag{3.6}$$

Thus $G(\nu,\theta)$ defined in Eq. (3.3) is nonsingular and its inverse is

$$[G(\nu,\theta)]^{-1}=U\left[Q_0(\nu)+e^{i\theta}Q_-(\nu)+e^{-i\theta}Q_+(\nu)\right]U\tag{3.7}$$

Indeed, with the aid of Eq. (2.85)–(2.87), Eq. (3.7) is what one obtains after substituting Eq. (3.1) into the backward marching form Eq. (2.88). Moreover, by using Eqs. (2.32), (3.3), (3.4), and (3.7), one has

$$[G(\nu,\theta)]^{-1}=U\overline{G(\nu,\theta)}U^{-1}\tag{3.8}$$

For each (ν,θ), let the three eigenvalues of $G(\nu,\theta)$ be denoted as $\sigma_\ell(\nu,\theta)$, $\ell=1,2,3$, respectively. They will be referred to as the amplification factors of the $a(3)$ scheme. Because $G(\nu,\theta)$ is nonsingular,

$$\sigma_\ell(\nu,\theta)\neq 0,\qquad \ell=1,2,3\tag{3.9}$$

(see part (i) of Theorem 1 given below). Also, as will be shown, $\sigma_\ell(\nu,\theta)$, $\ell=1,2,3$, satisfy the following set condition:

$$\left\{\frac{1}{\sigma_1(\nu,\theta)},\,\frac{1}{\sigma_2(\nu,\theta)},\,\frac{1}{\sigma_3(\nu,\theta)}\right\}=\left\{\overline{\sigma_1(\nu,\theta)},\,\overline{\sigma_2(\nu,\theta)},\,\overline{\sigma_3(\nu,\theta)}\right\}\tag{3.10}$$

Hereafter $\overline{z}$ denotes the complex conjugate of any complex number z.

As a preliminary, first we introduce the following well-known matrix theorems:

Theorem 1. Let A be a nonsingular $N\times N$ matrix with the eigenvalues λ_ℓ, $\ell=1,2,\ldots,N$. Then (i) $\lambda_\ell\neq 0$, $\ell=1,2,\ldots,N$; and (ii) the eigenvalues of A^{-1} are $1/\lambda_\ell$, $\ell=1,2,\ldots,N$.

Theorem 2. Let A be a $N\times N$ matrix with the eigenvalues λ_ℓ, $\ell=1,2,\ldots,N$. Then the eigenvalues of $\overline{A}$, the complex conjugate of A, are $\overline{\lambda_\ell}$, $\ell=1,2,\ldots,N$.

Theorem 3. Let A and B be two similar $N\times N$ matrices, i.e., there exists a nonsingular $N\times N$ matrix S so that $B=S^{-1}AS$. Then A and B have the same eigenvalues, counting multiplicity.

Theorems 1 and 2 are proved in Appendix A, while Theorem 3 is proved on p. 45 of [76].

To prove Eq. (3.10), note that part (ii) of Theorem 1 implies that, for any (ν,θ), the eigenvalues of $[G(\nu,\theta)]^{-1}$ are $1/\sigma_\ell(\nu,\theta)$, $\ell=1,2,3$. Next, by using Theorems 2 and 3, and the fact that $(U^{-1})^{-1}=U$, one can see that the eigenvalues of the matrix on the right side of Eq. (3.8) are $\overline{\sigma_\ell(\nu,\theta)}$, $\ell=1,2,3$. Thus Eq. (3.10) now is an immediate result of Eq. (3.8). QED.

An immediate result of Eq. (3.10) is

$$\frac{1}{\sigma_1(\nu,\theta)}\cdot\frac{1}{\sigma_2(\nu,\theta)}\cdot\frac{1}{\sigma_3(\nu,\theta)}=\overline{\sigma_1(\nu,\theta)}\cdot\overline{\sigma_2(\nu,\theta)}\cdot\overline{\sigma_3(\nu,\theta)}$$

i.e.,

$$|\sigma_1(\nu,\theta)|\cdot|\sigma_2(\nu,\theta)|\cdot|\sigma_3(\nu,\theta)|=1\tag{3.11}$$

As will be shown in Sec. 3.9, for any given ν, a necessary condition for the stability of the $a(3)$ scheme is

$$|\sigma_\ell(\nu,\theta)|\leq 1,\qquad \ell=1,2,3\tag{3.12}$$

Thus Eq. (3.11) implies that, for any given ν, the $a(3)$ scheme must be neutrally stable, i.e.,

$$|\sigma_\ell(\nu,\theta)|=1,\qquad \ell=1,2,3\tag{3.13}$$

if it is stable. As such, *Eq. (3.8) does not imply neutral stability of the $a(3)$ scheme*. However, it does imply that the scheme can only be neutrally stable (i.e., non-dissipative) if it is stable. Here we have reached this conclusion without using the explicit form of $\sigma_\ell(\nu,\theta)$, $\ell = 1, 2, 3$.

At this juncture, note that one can obtain

$$\sigma_\ell(-\nu,\theta) = \sigma_\ell(\nu,-\theta) = \overline{\sigma_\ell(\nu,\theta)}, \qquad \ell = 1,2,3 \tag{3.14}$$

by using Eqs. (3.4) and (3.5) along with Theorems 2 and 3.

Eq. (3.10) and (3.14) are the fundamental relations governing the eigenvalues of $G(\nu,\theta)$. In the following, we explore other properties of these eigenvalues.

3.2. Characteristic equation of $G(\nu,\theta)$

By using Eqs. (2.74)–(2.76) and (3.3), one has

$$G(\nu,\theta) =$$

$$\begin{pmatrix} 2-\cos\theta - i\nu\sin\theta & 2\nu(\cos\theta-1)+i(1+\nu^2)\sin\theta & \dfrac{2(1+2\nu^2)}{3}(1-\cos\theta) - \dfrac{2i\nu(2+\nu^2)}{3}\sin\theta \\ 2\nu(1-\cos\theta)+i\sin\theta & (2\nu^2-1)\cos\theta - 2\nu^2 + i\nu\sin\theta & \dfrac{2\nu(1+2\nu^2)}{3}(1-\cos\theta) + \dfrac{2i(1-\nu^2)}{3}\sin\theta \\ 3(\cos\theta-1) & 3\nu(1-\cos\theta) - 3\,i\sin\theta & 2(1+\nu^2)\cos\theta - 1 - 2\nu^2 + 2\,i\nu\sin\theta \end{pmatrix}$$

$$-\infty < \nu < +\infty;\ -\pi < \theta \le \pi \tag{3.15}$$

It follows from Eq. (3.15) that (i)

$$\det[G(\nu,\theta)] = -1 \tag{3.16}$$

and (ii) any eigenvalue σ of $G(\nu,\theta)$ must be a root of the characteristic equation:

$$\det[\sigma I - G(\nu,\theta)] \equiv \sigma^3 + h(\nu,\theta)\sigma^2 + \overline{h(\nu,\theta)}\sigma + 1 = 0 \tag{3.17}$$

where

$$h(\nu,\theta) \overset{\text{def}}{=} -1 + 4\nu^2(1-\cos\theta) - 2i\nu\sin\theta \tag{3.18}$$

The reader may be surprised by the simple result Eq. (3.16). However, by using Eq. (3.3) and the fact that each of $Q_0(\nu)$, $Q_+(\nu)$, and $Q_-(\nu)$ has the form $\vec{d}\,\vec{c}^{\,t}$ with $\vec{c}$ and $\vec{d}$ being 3×1 column vectors, an application of the fundamental definition of determinant (in which the Levi-Civita antisymmetric symbol is used) leads to the conclusion that $\det[G(\nu,\theta)]$ must be independent of θ, i.e., $\det[G(\nu,\theta)] = \det[G(\nu,0)]$. As such, Eq. (3.16) now follows from the fact that $G(\nu,0) = U$ (see Eqs. (3.15) and (2.32)) and $\det(U) = -1$. Hereafter, for simplicity, the arguments ν and θ may be omitted if no confusion would arise.

Because σ_1, σ_2, and σ_3 are the eigenvalues of G, Eq. (3.17) implies that

$$\sigma^3 + h\sigma^2 + \overline{h}\sigma + 1 \equiv (\sigma-\sigma_1)(\sigma-\sigma_2)(\sigma-\sigma_3) \tag{3.19}$$

for any complex variable σ. On the other hand, because Eq. (3.10) $\Leftrightarrow$

$$\{\sigma_1(\nu,\theta),\, \sigma_2(\nu,\theta),\, \sigma_3(\nu,\theta)\} = \left\{ \frac{1}{\overline{\sigma_1(\nu,\theta)}},\ \frac{1}{\overline{\sigma_2(\nu,\theta)}},\ \frac{1}{\overline{\sigma_3(\nu,\theta)}} \right\} \tag{3.20}$$

$1/\overline{\sigma_1}$, $1/\overline{\sigma_2}$, and $1/\overline{\sigma_3}$ must also be the eigenvalues. Thus

$$\sigma^3 + h\sigma^2 + \overline{h}\sigma + 1 \equiv \left(\sigma - \frac{1}{\overline{\sigma_1}}\right)\left(\sigma - \frac{1}{\overline{\sigma_2}}\right)\left(\sigma - \frac{1}{\overline{\sigma_3}}\right) \tag{3.21}$$

for any complex variable σ. In the following, Eq. (3.21) will be derived directly from Eq. (3.19) without using any other assumption.

Proof. Eqs. (3.19) $\Leftrightarrow$

$$\sigma_1\,\sigma_2\,\sigma_3 = -1, \quad \sigma_1\,\sigma_2 + \sigma_2\,\sigma_3 + \sigma_3\,\sigma_1 = \overline{h}, \quad \text{and} \quad \sigma_1 + \sigma_2 + \sigma_3 = -h \tag{3.22}$$

Eq. (3.9) follows from the relation $\sigma_1\sigma_2\sigma_3 = -1$. Also $\sigma_1\sigma_2\sigma_3 = -1 \Leftrightarrow$ Eq. (3.21) is valid if $\sigma = 0$.

Let $\sigma \neq 0$. Then, by replacing σ with $1/\overline{\sigma}$ in Eq. (3.19), one has

$$\frac{1}{\overline{\sigma}^3} + \frac{h}{\overline{\sigma}^2} + \frac{\overline{h}}{\overline{\sigma}} + 1 = \left(\frac{1}{\overline{\sigma}} - \sigma_1\right)\left(\frac{1}{\overline{\sigma}} - \sigma_2\right)\left(\frac{1}{\overline{\sigma}} - \sigma_3\right) \tag{3.23}$$

Also, by using the relation $\sigma_1\sigma_2\sigma_3 = -1$, one has

$$\overline{\sigma}^3 = (-\overline{\sigma}/\sigma_1)(-\overline{\sigma}/\sigma_2)(-\overline{\sigma}/\sigma_3) \tag{3.24}$$

Because the product of the expressions on the left sides of Eq. (3.23) and (3.24) equals to that on the right sides, we have

$$\overline{\sigma}^3 + \overline{h}\,\overline{\sigma}^2 + h\overline{\sigma} + 1 = \left(\overline{\sigma} - \frac{1}{\sigma_1}\right)\left(\overline{\sigma} - \frac{1}{\sigma_2}\right)\left(\overline{\sigma} - \frac{1}{\sigma_3}\right) \tag{3.25}$$

Eq. (3.21) is the complex conjugate form of Eq. (3.25). QED.

Moreover, according to Eq. (3.18),

$$h(-\nu,\theta) = h(\nu,-\theta) = \overline{h(\nu,\theta)}, \qquad -\infty < \nu < +\infty;\ -\pi < \theta \leq \pi \tag{3.26}$$

Thus Eq. (3.14) can also be derived directly from Eq. (3.19).

In this section, we will prove the following proposition:

Proposition 1. $|\sigma_\ell(\nu,\theta)| = 1$ for all ℓ and θ, $\ell = 1,2,3$, and $-\pi < \theta \leq \pi$, if and only if $|\nu| \leq 1/2$.

Proposition 1 can be divided into two parts, i.e.,

Proposition 1(a). $|\sigma_\ell(\nu,\theta)| = 1$ for all ℓ and θ, $\ell = 1,2,3$, and $-\pi < \theta \leq \pi$, if $|\nu| \leq 1/2$.

and

Proposition 1(b). For any ν with $|\nu| > 1/2$, there is a pair of ℓ_o and θ_o such that

$$\ell_o = 1,2,3, \quad -\pi < \theta_o \leq \pi, \quad \text{and} \quad |\sigma_{\ell_o}(\nu,\theta_o)| \neq 1 \tag{3.27}$$

A simple proof for Proposition 1(a) will be given in Sec. 3.3. Based on more exhausted developments, another proof for Proposition 1(a) and a proof for Proposition 1(b) will be given in Sec. 3.7.

3.3. A proof for Proposition 1(a)

First we introduce the following well-established algebraic theorem:

Theorem 4. Let $\sigma_1, \sigma_2, \ldots, \sigma_{N'}$ be the distinct roots of the Nth-order algebraic equation

$$\sigma^N + a_1\sigma^{N-1} + a_2\sigma^{N-2} + \ldots + a_{N-1}\sigma + a_N = 0 \tag{3.28}$$

where $a_1, a_2, \ldots, a_N$ are complex constant coefficients and σ is a complex variable. For each $\ell = 1, 2, \ldots, N'$, let $m_\ell \geq 1$ denote the multiplicity of the root σ_ℓ. Then

$$\sigma^N + a_1\sigma^{N-1} + a_2\sigma^{N-2} + \ldots + a_{N-1}\sigma + a_N \equiv \prod_{\ell=1}^{N'}(\sigma - \sigma_\ell)^{m_\ell} \qquad \text{and} \qquad \sum_{\ell=1}^{N'} m_\ell = N \tag{3.29}$$

According to the above theorem, for any given (ν, θ), the roots of the cubic equation Eq. (3.17) must fall into one of the following three mutually exclusive cases: (a) there is one triple root (multiplicity = 3); (b) there are one double root (multiplicity = 2) and one simple root (multiplicity = 1) and (c) there are three simple roots.

Consider case (a). Then $\sigma_1 = \sigma_2 = \sigma_3$. Let σ_o denote the common value of σ_1, σ_2, and σ_3. Then Eqs. (3.9) and (3.20) imply that (i) $\sigma_o \neq 0$ and (ii) $1/\overline{\sigma_o}$ must also be a triple root of Eq. (3.17). Thus the only choice that will not contradict Theorem 4 is that $\sigma_o = 1/\overline{\sigma_o}$, i.e., $|\sigma_o| = |\sigma_1| = |\sigma_2| = |\sigma_3| = 1$.

Consider case (b). Without any loss of generality, one can assume $\sigma_1 = \sigma_2 \neq \sigma_3$. Again let σ_o denote the common value of σ_1 and σ_2. Then Eqs. (3.9) and (3.20) imply that (i) $\sigma_o \neq 0$; (ii) $\sigma_3 \neq 0$; and (iii) $1/\overline{\sigma_o}$ and $1/\overline{\sigma_3}$ must also be a double root and a simple root of Eq. (3.17), respectively. Thus the only choice that will not contradict Theorem 4 is that $\sigma_o = 1/\overline{\sigma_o}$ and $\sigma_3 = 1/\overline{\sigma_3}$, i.e., $|\sigma_o| = |\sigma_1| = |\sigma_2| = |\sigma_3| = 1$.

The conclusions reached above imply the following lemma:

Lemma 1. For any given (ν, θ), the roots of Eq. (3.17) must all be of unit magnitude if any one of them is a multiple root.

Thus, to prove Proposition 1(a), we need only to consider case (c), i.e., the case with

$$\sigma_1 \neq \sigma_2, \quad \sigma_1 \neq \sigma_3, \quad \text{and} \quad \sigma_2 \neq \sigma_3 \tag{3.30}$$

To proceed, each $\sigma_\ell(\nu, \theta)$ is expressed in its polar form, i.e.,

$$\sigma_\ell(\nu, \theta) = r_\ell(\nu, \theta) e^{i\phi_\ell(\nu, \theta)}, \qquad \ell = 1, 2, 3;\ -\infty < \nu < +\infty;\ -\pi < \theta \leq \pi \tag{3.31}$$

where, because of Eq. (3.9)

$$r_\ell(\nu, \theta) \overset{\text{def}}{=} |\sigma_\ell(\nu, \theta)| > 0, \qquad \ell = 1, 2, 3;\ -\infty < \nu < +\infty;\ -\pi < \theta \leq \pi \tag{3.32}$$

Moreover, for each $\sigma_\ell(\nu, \theta)$, the corresponding phase angle $\phi_\ell(\nu, \theta)$ is uniquely defined by Eq. (3.31) and

$$-\pi < \phi_\ell(\nu, \theta) \leq \pi, \qquad \ell = 1, 2, 3;\ -\infty < \nu < +\infty;\ -\pi < \theta \leq \pi \tag{3.33}$$

Hereafter, the arguments ν and θ may be dropped from $r_\ell(\nu, \theta)$ and $\phi_\ell(\nu, \theta)$ if no confusion would arise. It follows from Eqs. (3.31) and (3.32) that

$$1/\overline{\sigma_\ell} = (1/r_\ell) e^{i\phi_\ell}, \qquad \ell = 1, 2, 3 \tag{3.34}$$

Also, by using Eqs. (3.31)–(3.33), Eqs. (3.30) can be expressed as the following ordered pair inequalities:

$$(r_1, \phi_1) \neq (r_2, \phi_2), \quad (r_1, \phi_1) \neq (r_3, \phi_3), \quad \text{and} \quad (r_2, \phi_2) \neq (r_3, \phi_3) \tag{3.35}$$

The distribution of ϕ_1, ϕ_2, and ϕ_3 must fall into one of the following mutually exclusive cases: (c1) all have distinct values; (c2) two of them have the same value while the third assumes a different value; and (c3) all have the same values. In the following, these sub-cases will be discussed separately.

Consider case (c1) where

$$\phi_1 \neq \phi_2, \quad \phi_1 \neq \phi_3, \quad \text{and} \quad \phi_2 \neq \phi_3 \tag{3.36}$$

Because Eqs. (3.20) and (3.34) imply that $(1/r_\ell)e^{i\phi_\ell}$, $\ell = 1, 2, 3$, must also be roots of Eq. (3.17), Eq. (3.36) implies that the only choice that will not contradict Theorem 4 is that $r_\ell = 1/r_\ell$, i.e., $r_\ell = 1$, $\ell = 1, 2, 3$. Thus, *for case (c1), again we have* $|\sigma_1| = |\sigma_2| = |\sigma_3| = 1$.

Consider case (c2) where, without any loss of generality, one can assume that

$$\phi_1 = \phi_2 \neq \phi_3 \tag{3.37}$$

Because of (3.35), Eq. (3.37) implies that

$$r_1 \neq r_2 \tag{3.38}$$

By using Eqs. (3.37) and (3.38) along with the fact that $(1/r_\ell)e^{i\phi_\ell}$, $\ell = 1, 2, 3$, must also be roots of Eq. (3.17), one concludes that the only choice that will not contradict Theorem 4 is that $r_1 r_2 = 1$, $r_1 \neq 1$, $r_2 \neq 1$, and $r_3 = 1$. Thus, *for case (c2), (i) one of the roots is of unit magnitude while the other two are not; and (ii) the product of the magnitudes of the two roots which are not of unit magnitude is one*.

Consider case (c3) where

$$\phi_1 = \phi_2 = \phi_3 \tag{3.39}$$

Because of Eq. (3.35), Eq. (3.39) implies that

$$r_1 \neq r_2, \quad r_1 \neq r_3, \quad \text{and} \quad r_2 \neq r_3 \tag{3.40}$$

By using an argument similar to that invoked in the discussion of case (c2), one concludes that, *for case (c3), again (i) one of the roots is of unit magnitude while the other two are not; and (ii) the product of the magnitudes of the two roots which are not of unit magnitude is one*.

As a result of the above discussions, we have the following lemma:

Lemma 2. For any given (ν, θ), the case with at least one of the roots of Eq. (3.17) not being of unit magnitude may occur only if it meets the following conditions: (i) one and only one of r_1, r_2, and r_3 is of unit magnitude; and (ii) the two roots that are not of unit magnitude share the same phase angle and the product of their magnitudes is one.

Consider any case that meets the conditions referred to in Lemma 2. Then, without any loss of generality, one may assume that

$$r_1 r_2 = 1, \quad r_1 \neq 1, \quad r_3 = 1, \quad \text{and} \quad \phi_1 = \phi_2 = \phi \tag{3.41}$$

where ϕ denotes the common value of ϕ_1 and ϕ_2. Moreover, by using Eq. (3.31), Eqs. (3.18) and (3.22) imply that

$$r_1 r_2 r_3 e^{i(\phi_1+\phi_2+\phi_3)} = -1 \tag{3.42}$$

$$r_1 r_2 e^{i(\phi_1+\phi_2)} + r_1 r_3 e^{i(\phi_1+\phi_3)} + r_2 r_3 e^{i(\phi_2+\phi_3)} = -1 + 4\nu^2(1-\cos\theta) + 2i\nu\sin\theta \tag{3.43}$$

and

$$r_1 e^{i\phi_1} + r_2 e^{i\phi_2} + r_3 e^{i\phi_3} = 1 - 4\nu^2(1-\cos\theta) + 2i\nu\sin\theta \tag{3.44}$$

Because of Eq. (3.32), Eq. (3.42) $\Leftrightarrow$

$$r_1 r_2 r_3 = 1 \tag{3.45}$$

and

$$e^{i(\phi_1+\phi_2+\phi_3)} = -1 \tag{3.46}$$

By using Eqs. (3.45) and (3.46), Eq.(3.43) $\Leftrightarrow$

$$\frac{1}{r_1}e^{i\phi_1} + \frac{1}{r_2}e^{i\phi_2} + \frac{1}{r_3}e^{i\phi_3} = 1 - 4\nu^2(1-\cos\theta) + 2i\nu\sin\theta \tag{3.47}$$

Thus Eqs. (3.44)–(3.47) represent all the independent constraints imposed on r_ℓ and ϕ_ℓ, $\ell = 1, 2, 3$.

Note that: (i) Eq. (3.41) implies Eq. (3.45); and (ii) Eqs. (3.41) and (3.46) imply that

$$e^{i\phi_1} = e^{i\phi_2} = e^{i\phi} \quad \text{and} \quad e^{i\phi_3} = -e^{-i(\phi_1+\phi_2)} = -e^{-2i\phi} \tag{3.48}$$

Let

$$\rho \overset{\text{def}}{=} r_1 \tag{3.49}$$

Then, with the aid of Eqs. (3.41) and (3.48), both Eqs. (3.44) and (3.47) reduce to

$$f(\rho)e^{i\phi} - e^{-2i\phi} = 1 - 4\nu^2(1-\cos\theta) + 2i\nu\sin\theta, \qquad -\pi < \phi \leq \pi;\ \rho > 0 \text{ and } \rho \neq 1 \tag{3.50}$$

where

$$f(\rho) \overset{\text{def}}{=} \rho + \frac{1}{\rho}, \qquad \rho > 0 \text{ and } \rho \neq 1 \tag{3.51}$$

Eq. (3.50) $\Leftrightarrow$

$$f(\rho)\cos\phi - \cos(2\phi) = 1 - 4\nu^2(1-\cos\theta), \qquad -\pi < \phi \leq \pi;\ \rho > 0 \text{ and } \rho \neq 1 \tag{3.52}$$

and

$$f(\rho)\sin\phi + \sin(2\phi) = 2\nu\sin\theta, \qquad -\pi < \phi \leq \pi;\ \rho > 0 \text{ and } \rho \neq 1 \tag{3.53}$$

Thus, given any (ν, θ), Eqs. (3.52) and (3.53) must admit a solution for ρ and ϕ in the specified domain if the case Eq. (3.41) indeed exists.

To explore Eqs. (3.52) and (3.53), note that (i)

$$[f(\rho)\cos\phi - \cos(2\phi)]^2 + [f(\rho)\sin\phi + \sin(2\phi)]^2 = \left[1 - 4\nu^2(1-\cos\theta)\right]^2 + [2\nu\sin\theta]^2 \tag{3.54}$$

is a direct result of Eqs. (3.52) and (3.53); (ii)

$$[f(\rho)\cos\phi - \cos(2\phi)]^2 + [f(\rho)\sin\phi + \sin(2\phi)]^2 \equiv [f(\rho) - 1]^2 + 2f(\rho)\,[1 - \cos(3\phi)] \tag{3.55}$$

and (iii)

$$\left[1 - 4\nu^2(1-\cos\theta)\right]^2 + [2\nu\sin\theta]^2 \equiv 1 - 4\nu^2(1-4\nu^2)(1-\cos\theta)^2 \tag{3.56}$$

Next, because (i) the minimum of $f(\rho)$ in the domain $\rho > 0$ occurs at $\rho = 1$ and (ii) $f(1) = 2$, we have

$$f(\rho) > 2 \quad \text{if} \quad \rho > 0 \text{ and } \rho \neq 1 \tag{3.57}$$

Combining Eqs. (3.55) and (3.57), and using the fact that $1 - \cos(3\phi) \geq 0$ for all ϕ, one has

$$[f(\rho)\cos\phi - \cos(2\phi)]^2 + [f(\rho)\sin\phi + \sin(2\phi)]^2 > 1 \quad \text{if} \quad \rho > 0 \text{ and } \rho \neq 1 \tag{3.58}$$

On the other hand, because $1 - 4\nu^2 \geq 0$ if $|\nu| \leq 1/2$, Eq. (3.56) implies that

$$\left[1 - 4\nu^2(1-\cos\theta)\right]^2 + [2\nu\sin\theta]^2 \leq 1 \quad \text{if} \quad |\nu| \leq 1/2 \tag{3.59}$$

Combining Eqs. (3.58) and (3.59), one arrives at the conclusion that, for all θ,

$$[f(\rho)\cos\phi - \cos(2\phi)]^2 + [f(\rho)\sin\phi + \sin(2\phi)]^2 > \left[1 - 4\nu^2(1-\cos\theta)\right]^2 + [2\nu\sin\theta]^2 \tag{3.60}$$

i.e., Eq. (3.54) cannot be satisfied, if (i) $|\nu| \leq 1/2$; and (ii) $\rho > 0$ and $\rho \neq 1$. Because Eq. (3.54) is a direct result of Eqs. (3.52) and (3.53), this implies that, for any θ, Eqs. (3.52) and (3.53) admit no solution for ρ and ϕ in the specified domain, i.e., the case Eq. (3.41) does not exist, if $|\nu| \leq 1/2$. In turn, this implies that, for all θ, the roots of Eq. (3.17) are all of unit magnitude if $|\nu| \leq 1/2$, i.e., Proposition 1(a) has been proved. QED.

Note that, for a reason to be given in Sec. 3.9, by itself Proposition 1(a) does not imply that $a(3)$ scheme is stable when $|\nu| \leq 1/2$. Next, as a preliminary for later developments, several special cases will be discussed in Secs. 3.4 and 3.5.

3.4. The $|\nu| = 1/2$ case

Let $\nu = 1/2$. Then Eq. (3.17) reduces to

$$\sigma^3 - e^{i\theta}\sigma^2 - e^{-i\theta}\sigma + 1 \equiv \left(\sigma - e^{i\theta}\right)\left(\sigma - e^{-i\theta/2}\right)\left(\sigma + e^{-i\theta/2}\right) = 0, \quad -\pi < \theta \leq \pi \quad (\nu = 1/2) \tag{3.61}$$

Thus the roots of Eq. (3.17) are

$$\sigma = \sigma_0(\theta) \overset{\text{def}}{=} e^{i\theta} \quad \text{and} \quad \sigma = \sigma_\pm(\theta) \overset{\text{def}}{=} \pm e^{-i\theta/2}, \qquad -\pi < \theta \le \pi \quad (\nu = 1/2) \tag{3.62}$$

On the other hand, by using Eqs. (3.14) and (3.62), one concludes that the roots for the case $\nu = -1/2$ are

$$\sigma = \overline{\sigma_0(\theta)} = e^{-i\theta} \quad \text{and} \quad \sigma = \overline{\sigma_\pm(\theta)} = \pm e^{i\theta/2}, \qquad -\pi < \theta \le \pi \quad (\nu = -1/2) \tag{3.63}$$

For each of the above two cases, Eqs. (3.62) and (3.63) imply that the these roots are distinct if

$$\theta \neq 0 \quad \text{and} \quad |\theta| \neq \frac{2\pi}{3} \qquad\qquad (|\nu| = 1/2) \tag{3.64}$$

In fact, there are one double root and one simple root if $\theta = 0$ or $|\theta| = 2\pi/3$. Also, because the analytical amplification factor is $e^{-i\nu\theta}$ for any (ν, θ) (see p.4 of [61]), for the case $\nu = 1/2$ ($\nu = -1/2$), one of the roots of Eq. (3.17), i.e., $\sigma_+(\theta)$ ($\overline{\sigma_+(\theta)}$), is identical to the analytical amplification factor.

Consider the plane wave solution

$$u(x, t) = e^{ik(x-at)} \qquad\qquad (ka \neq 0) \tag{3.65}$$

The period associated with this solution is

$$T = \frac{2\pi}{|ka|} \tag{3.66}$$

Let n, the number of total marching steps, and Δt be chosen such that

$$n\Delta t = NT, \qquad\qquad N = 1, 2, 3, \ldots \tag{3.67}$$

Then, by using Eq. (3.66), one has

$$n = \frac{2\pi N}{|\theta||\nu|} \tag{3.68}$$

where

$$\theta = k\Delta x \neq 0 \tag{3.69}$$

is the variation of phase angle over the interval Δx. For the case $|\nu| = 1/2$, Eq. (3.68) reduces to

$$n = \frac{4\pi N}{|\theta|} \qquad\qquad (n\Delta t = NT;\ |\nu| = 1/2)) \tag{3.70}$$

Eqs. (3.62), (3.63), and (3.70) imply that

$$[\sigma_\pm(\theta)]^n = \left[\overline{\sigma_\pm(\theta)}\right]^n = (\pm 1)^n \tag{3.71}$$

and

$$[\sigma_0(\theta)]^n = \left[\overline{\sigma_0(\theta)}\right]^n = 1 \tag{3.72}$$

Thus

$$[\sigma_\pm(\theta)]^n = \left[\overline{\sigma_\pm(\theta)}\right]^n = 1, \qquad \text{if } n \text{ is even} \tag{3.73}$$

On the other hand, Eq. (3.68) implies that

$$\left(e^{-i\nu\theta}\right)^n = 1 \tag{3.74}$$

By using an analytical procedure similar to that used in Sec. 5 of [1], one can show that Eqs. (3.72)–(3.74) and the fact that $e^{-i\nu\theta}$ is the analytical amplification factor lead to the conclusion that, for the case $|\nu| = 1/2$, the numerical solution generated by the $a(3)$ scheme in a simulation involving a periodic boundary condition, aside from round-off errors, should be identical to the exact solution if (i) n and Δt are chosen according to Eq. (3.67), and n is even; and (ii) the phase angles of the Fourier components involved in the simulation observe the condition Eq. (3.64) (i.e., the three eigenvalues associated with each Fourier component are distinct). This prediction has been verified numerically (see Sec. 4).

Next, a brief discussion on the roots of Eq. (3.17) for the three special cases: (a) $\nu = 0$; (b) $\theta = 0$; and (c) $\theta = \pi$ will be given in Sec. 3.5.

3.5. Three other special cases

Let $\nu = 0$ or $\theta = 0$. Then Eqs. (3.17) and (3.18) imply that

$$\sigma^3 - \sigma^2 - \sigma + 1 \equiv (\sigma - 1)^2(\sigma + 1) = 0 \tag{3.75}$$

i.e., the roots of Eq. (3.17) are 1, 1, and -1. According to Eqs. (3.31)–(3.33), (i) $r_1 = r_2 = r_3 = 1$, and (ii) one can assume that $\phi_1 = \phi_2 = 0$ and $\phi_3 = \pi$.

Let $\theta = \pi$. Then Eqs. (3.17) and (3.18) imply that

$$\sigma^3 + (8\nu^2 - 1)\sigma^2 + (8\nu^2 - 1)\sigma + 1 \equiv (\sigma + 1)\left[\sigma^2 + 2(4\nu^2 - 1)\sigma + 1\right] = 0 \tag{3.76}$$

i.e., the roots of Eq. (3.17) are $\sigma = -1$ (i.e., $r_\ell = 1$ and $\phi_\ell = \pi$ for a value of ℓ), and

$$\sigma = 1 - 4\nu^2 \pm \sqrt{8\nu^2(2\nu^2 - 1)} \tag{3.77}$$

We have

$$1 - 4\nu^2 - \sqrt{8\nu^2(2\nu^2 - 1)} < 1 - 2 = -1 \quad \text{if} \quad 2\nu^2 > 1$$

i.e.,

$$\left|1 - 4\nu^2 - \sqrt{8\nu^2(2\nu^2 - 1)}\right| > 1 \quad \text{if} \quad 2\nu^2 > 1 \tag{3.78}$$

Thus the magnitude of at least one root of Eq. (3.17) is greater than one if $\theta = \pi$ and $|\nu| > 1/\sqrt{2}$.

On the other hand,

$$\begin{aligned} &\left|1 - 4\nu^2 \pm \sqrt{8\nu^2(2\nu^2 - 1)}\right| = \left|1 - 4\nu^2 \pm i\,2\sqrt{2}\,|\nu|\sqrt{1 - 2\nu^2}\right| \\ &= \sqrt{(1 - 4\nu^2)^2 + 8\nu^2(1 - 2\nu^2)} = 1 \quad \text{if} \quad 2\nu^2 \le 1 \end{aligned} \tag{3.79}$$

Thus, for the case $\theta = \pi$ and $|\nu| \le 1/\sqrt{2}$, $r_1 = r_2 = r_3 = 1$. Moreover, for the special case

$$\theta = \pi \quad \text{and} \quad |\nu| = 1/\sqrt{2} \tag{3.80}$$

Eq. (3.76) reduces to $(\sigma + 1)^3 = 0$, i.e., -1 is the triple root of Eq. (3.17). In fact, it will be shown in Sec. 3.6 that *the only possible triple root of unit magnitude for Eq. (3.17) is* -1 *and it has this root only for the case Eq. (3.80).*

Eq. (3.17) has three roots for any given (ν, θ). In Sec. 3.6, we derive a set of equations governing the phase angles of these roots *when they all are of unit magnitude*. To pave the way, let

$$\Psi \overset{\text{def}}{=} \{(\nu, \theta) | -\infty < \nu < +\infty;\ -\pi < \theta \le \pi,\ \text{and}\ r_1(\nu, \theta) = r_2(\nu, \theta) = r_3(\nu, \theta) = 1\} \tag{3.81}$$

and

$$\Psi_o \overset{\text{def}}{=} \{(\nu, \theta) | \nu \ne 0,\ 0 < |\theta| < \pi,\ \text{and}\ r_1(\nu, \theta) = r_2(\nu, \theta) = r_3(\nu, \theta) = 1\} \tag{3.82}$$

According to Proposition 1(a) (which has been proved), $(\nu,\theta)\in\Psi$ if $|\nu|\le 1/2$ and $-\pi<\theta\le\pi$.

3.6. Phase angle equations for $(\nu,\theta)\in\Psi$

With the aid of Eq. (3.22), and (3.31)–(3.33), one concludes that the condition

$$r_1(\nu,\theta)=r_2(\nu,\theta)=r_3(\nu,\theta)=1 \tag{3.83}$$

$\Leftrightarrow$ the real phase angles $\phi_\ell(\nu,\theta)$, $\ell=1,2,3$, satisfy

$$e^{i(\phi_1+\phi_2+\phi_3)}=-1 \tag{3.84}$$

$$e^{i(\phi_1+\phi_2)}+e^{i(\phi_1+\phi_3)}+e^{i(\phi_2+\phi_3)}=-1+4\nu^2(1-\cos\theta)+2i\nu\sin\theta,\quad -\infty<\nu<+\infty;\ -\pi<\theta\le\pi \tag{3.85}$$

and

$$e^{i\phi_1}+e^{i\phi_2}+e^{i\phi_3}=1-4\nu^2(1-\cos\theta)+2i\nu\sin\theta,\qquad -\infty<\nu<+\infty;\ -\pi<\theta\le\pi \tag{3.86}$$

Because (i) Eq. (3.84) implies that

$$e^{i(\phi_1+\phi_2)}=-e^{-i\phi_3},\quad e^{i(\phi_1+\phi_3)}=-e^{-i\phi_2},\quad e^{i(\phi_2+\phi_3)}=-e^{-i\phi_1} \tag{3.87}$$

and (ii) the complex conjugate of Eq. (3.86) is

$$e^{-i\phi_1}+e^{-i\phi_2}+e^{-i\phi_3}=1-4\nu^2(1-\cos\theta)-2i\nu\sin\theta \tag{3.88}$$

one can see easily that Eq. (3.85) is a result of Eqs. (3.84) and (3.86). Thus Eq. (3.83) $\Leftrightarrow$ Eqs. (3.84) and (3.86).

At this juncture, we will prove that the only possible triple root of unit magnitude for Eq. (3.17) is -1 and it has this root only for the case Eq. (3.80).

Proof. Let Eq. (3.17) have a triple root of unit magnitude. Then $(\nu,\theta)\in\Psi$ and $\phi_1=\phi_2=\phi_3$. With the aid of Eqs. (3.33), in turn Eq. (3.84) implies that either (a)

$$\phi_1=\phi_2=\phi_3=\pi \tag{3.89}$$

or (b)

$$\phi_1=\phi_2=\phi_3=\pm\pi/3 \tag{3.90}$$

For case (a), by substituting Eq. (3.89) into Eq. (3.86), one has

$$\nu^2(1-\cos\theta)=1\quad\text{and}\quad\nu\sin\theta=0 \tag{3.91}$$

which $\Leftrightarrow$ $\cos\theta=-1$ and $|\nu|=1/\sqrt{2}$. Because $-\pi<\theta\le\pi$, in turn Eq. (3.91) $\Leftrightarrow$ Eq. (3.80), i.e., case (a) is the same case defined by Eq. (3.80).

For case (b), by Substituting Eq. (3.90) into Eq. (3.86), one has

$$(3/2)(1\pm\sqrt{3}\,i)=1-4\nu^2(1-\cos\theta)+2i\nu\sin\theta \tag{3.92}$$

By taking the real part of Eq. (3.92), one arrives at the result

$$\nu^2(1-\cos\theta)=-1/8 \tag{3.93}$$

Because $\nu^2(1-\cos\theta)\ge 0$ for any (ν,θ) and $-1/8<0$, Eq. (3.93) cannot be true and therefore case (b) does not exist. QED.

Let $(\nu,\theta)\in\Psi$. Then Eqs. (3.84) and (3.86) are valid. By eliminating $e^{i\phi_3}$ from Eqs. (3.84) and (3.86), one has

$$e^{i\phi_1}+e^{i\phi_2}-e^{-i(\phi_1+\phi_2)}=1-4\nu^2(1-\cos\theta)+2i\nu\sin\theta \tag{3.94}$$

By separating the real and imaginary parts of Eq. (3.94), we have

$$\cos\phi_1 + \cos\phi_2 - \cos(\phi_1+\phi_2) - 1 = -4\nu^2(1-\cos\theta) \tag{3.95}$$

and

$$\sin\phi_1 + \sin\phi_2 + \sin(\phi_1+\phi_2) = 2\nu\sin\theta \tag{3.96}$$

Similarly, one can show that

$$\cos\phi_2 + \cos\phi_3 - \cos(\phi_2+\phi_3) - 1 = -4\nu^2(1-\cos\theta) \tag{3.97}$$

$$\sin\phi_2 + \sin\phi_3 + \sin(\phi_2+\phi_3) = 2\nu\sin\theta \tag{3.98}$$

$$\cos\phi_3 + \cos\phi_1 - \cos(\phi_3+\phi_1) - 1 = -4\nu^2(1-\cos\theta) \tag{3.99}$$

and

$$\sin\phi_3 + \sin\phi_1 + \sin(\phi_3+\phi_1) = 2\nu\sin\theta \tag{3.100}$$

At this juncture, note that Eqs. (3.95), (3.97), and (3.99) imply that

$$\nu^2(1-\cos\theta) = 0 \quad \text{if} \quad \phi_\ell = 0 \text{ for any } \ell = 1,2,3 \tag{3.101}$$

Because $-\pi < \theta \le \pi$, Eq. (3.101) implies that at least one of the two cases (a) $\nu = 0$ and (b) $\theta = 0$ must occur if $\phi_\ell = 0$ for any $\ell = 1,2,3$. It is shown in Sec. 3.5 that, for $\nu = 0$ or $\theta = 0$, indeed $\phi_\ell = 0$ for two different values of ℓ.

On the other hand, Eqs. (3.96), (3.98), and (3.100) imply that

$$\nu\sin\theta = 0 \quad \text{if} \quad \phi_\ell = \pi \text{ for any } \ell = 1,2,3 \tag{3.102}$$

Because $-\pi < \theta \le \pi$, Eq. (3.102) implies that at least one of the three cases (a) $\nu = 0$, (b) $\theta = 0$, (c) $\theta = \pi$ must occur if $\phi_\ell = \pi$ for any $\ell = 1,2,3$. It is also shown in Sec. 3.5 that, for any of above three cases, indeed $\phi_\ell = \pi$ for a value of ℓ.

Using the above results along with Eqs. (3.33) and (3.82), one arrives at the important conclusion that

$$0 < |\phi_\ell(\nu,\theta)| < \pi, \quad \ell = 1,2,3, \quad \text{if} \quad (\nu,\theta) \in \Psi_o \tag{3.103}$$

To eliminate ϕ_2, let (i) Eq. (3.95) be multiplied by $(1+\cos\phi_1)$; and (ii) Eq. (3.96) be multiplied by $\sin\phi_1$. After subtracting the resulting equations from each other, a rearrangement using elementary trigonometry yields

$$\sin^2\phi_1 - 2\nu^2(1-\cos\theta)(1+\cos\phi_1) - \nu\sin\theta\sin\phi_1 = 0 \tag{3.104}$$

By applying similar manipulations over Eqs. (3.97), (3.98), (3.99), and (3.100), one can show that Eq. (3.104) remains valid if ϕ_1 is replaced by ϕ_2 or ϕ_3. Thus, for any $(\nu,\theta) \in \Psi$, we have

$$F(\nu,\theta,\phi_\ell) = 0, \qquad\qquad \ell = 1,2,3 \tag{3.105}$$

where

$$\begin{aligned} F(\nu,\theta,\phi) &\overset{\text{def}}{=} \sin^2\phi - 2\nu^2(1-\cos\theta)(1+\cos\phi) - \nu\sin\theta\sin\phi \\ -\infty < \nu &< +\infty;\ -\pi < \theta \le \pi;\ -\pi < \phi \le \pi \end{aligned} \tag{3.106}$$

Eq. (3.106) implies that, for all (ν,θ) with $-\infty < \nu < +\infty$ and $-\pi < \theta \le \pi$, we have

$$F(-\nu,\theta,\phi) \equiv F(\nu,-\theta,\phi) \equiv F(\nu,\theta,-\phi) \tag{3.107}$$

$$F(\pm 1/2,\theta,\pm\theta) \equiv F(\pm 1/2,\theta,\mp\theta/2) \equiv F(\pm 1/2,\theta,\pi\mp\theta/2) \equiv 0 \tag{3.108}$$

and

$$F(\nu,\theta,\pi) \equiv 0 \tag{3.109}$$

Eqs. (3.105) and (3.107) are consistent with Eqs. (3.14) and (3.31) while Eq. (3.108) is consistent with the special results Eqs. (3.62) and (3.63). On the other hand, because (i) $\sin\pi = 1+\cos\pi = 0$, and (ii) Eq. (3.104) is obtained from a subtraction of two expressions which result from multiplying Eqs. (3.95) and (3.96) with $(1+\cos\phi_1)$ and $\sin\phi_1$, respectively, the fact that Eq. (3.109) is true for all (ν,θ), $-\infty < \nu < +\infty;\ -\pi < \theta \le \pi$. is an artificial result accidently introduced in the derivation of Eq. (3.105).

According to the above discussions, given any $(\nu,\theta) \in \Psi$, the phase angle ϕ of any root of Eq. (3.17) must satisfy

$$F(\nu,\theta,\phi) = 0, \qquad -\infty < \nu < +\infty;\ -\pi < \theta \le \pi;\ -\pi < \phi \le \pi \tag{3.110}$$

Recall that the analytical amplification factor is given by $e^{-i\nu\theta}$. Thus it is expected that $\phi = -\nu\theta$ should be a good approximated solution to Eq. (3.110) when $|\theta|$ is small (i.e., when the solution Eq. (3.65) has a very small variation over the spatial interval Δx and thus it is closely approximated by a discrete solution). In fact, with the aid of Eq. (3.106) and the Taylor's expansions

$$\sin x = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + O(x^9) \quad \text{and} \quad \cos x = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + O(x^8) \tag{3.111}$$

one has

$$F(\nu,\theta,-\nu\theta) = \sin^2(\nu\theta) - 2\nu^2(1-\cos\theta)[1+\cos(\nu\theta)] + \nu\sin\theta\sin(\nu\theta) = (4\nu^2-1)(\nu^2-1)\frac{\nu^2\theta^6}{360} + O(\theta^8) \tag{3.112}$$

Because $4\nu^2 - 1 = 0 \Leftrightarrow |\nu| = 1/2$, the above result is consistent with the fact that $F(\pm 1/2, \theta, \mp\theta/2) \equiv 0$, which was presented as part of Eq. (3.108).

Given any (ν,θ), Newton's iterative procedure for obtaining a root ϕ of Eq. (3.110) is defined by

$$\phi^{n+1} = \phi^n - \frac{F(\nu,\theta,\phi^n)}{F_\phi(\nu,\theta,\phi^n)}, \qquad n = 0,1,2,3,\ldots \tag{3.113}$$

where (i) ϕ^n is the nth iterative value of ϕ and (ii)

$$F_\phi(\nu,\theta,\phi) \overset{\text{def}}{=} \frac{\partial F(\nu,\theta,\phi)}{\partial\phi} = \sin(2\phi) + 2\nu^2(1-\cos\theta)\sin\phi - \nu\sin\theta\cos\phi \tag{3.114}$$

For a given $(\nu,\theta) \in \Psi$, the phase angle ϕ of any $\sigma_\ell(\nu,\theta)$ must satisfy Eq. (3.110). Moreover, according to Eq. (3.103), $0 < |\phi| < \pi$ if $\nu \ne 0$ and $0 < |\theta| < \pi$. As a preliminary to the proof to be given in Sec. 3.7, the equation $F(\nu,\theta,\phi) = 0$ will be cast into a cubic equation assuming

$$\nu \ne 0 \quad \text{and} \quad 0 < |\theta|, |\phi| < \pi \tag{3.115}$$

As a result of Eq. (3.115), we have

$$1 + \cos\phi \ne 0 \quad \text{and} \quad \nu\sin\theta \ne 0 \tag{3.116}$$

With the aid of Eqs. (3.106) and (3.116), Eq. (3.110) implies that

$$\frac{1-\cos\phi - 2\nu^2(1-\cos\theta)}{\nu\sin\theta} - \frac{\sin\phi}{1+\cos\phi} = 0, \qquad \nu \ne 0;\ 0 < |\theta|, |\phi| < \pi \tag{3.117}$$

Because

$$1 - \cos\phi \equiv \frac{2\tan^2(\phi/2)}{1+\tan^2(\phi/2)}, \quad \frac{\sin\phi}{1+\cos\phi} \equiv \tan(\phi/2), \quad \text{and} \quad \frac{1-\cos\theta}{\sin\theta} \equiv \tan(\theta/2), \quad 0 < |\theta|, |\phi| < \pi \tag{3.118}$$

Eq. (3.117) $\Leftrightarrow$

$$\tau^3 + 2\left[\nu \tan(\theta/2) - \frac{1}{\nu \sin\theta}\right]\tau^2 + \tau + 2\nu\tan(\theta/2) = 0, \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.119}$$

where τ is related to ϕ through the *one-to-one* relation

$$\tau \overset{\text{def}}{=} \tan(\phi/2), \qquad 0 < |\phi| < \pi \tag{3.120}$$

It has been shown that, for any $(\nu,\theta) \in \Psi_o$, each real phase angle $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$ must satisfy Eq. (3.119) through the one-to-one relation $\tau = \tan[\phi_\ell(\nu,\theta)/2]$. As such, assuming (i) $(\nu,\theta) \in \Psi_o$ and (ii) $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$ are distinct, Theorem 4 implies that the roots of Eq. (3.119) are also real and distinct, and can be indexed such that

$$\tau_\ell(\nu,\theta) = \tan[\phi_\ell(\nu,\theta)/2], \qquad \ell = 1,2,3 \tag{3.121}$$

However, for the case in which $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$, are not distinct, two or three of the phase angles that share a common value may be linked with a real root of Eq. (3.119) through a relation in the form of Eq. (3.120). As such there is a possibility that one or two roots of Eq. (3.119) may not be linked with any $\phi_\ell(\nu,\theta)$ in the form of Eq. (3.120) or any way whatsoever. In the following, this possibility will be ruled out. In fact, we will prove the following proposition:

Proposition 2. Let (i) $\nu \neq 0$ and $0 < |\theta| < \pi$, and (ii) $r_\ell(\nu,\theta)$ and $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$, be defined using Eqs. (3.31)–(3.33). Then $(\nu,\theta) \in \Psi_o$, i.e., Eq. (3.83) is true, if and only if the roots of Eq. (3.119) are all real. Moreover, these real roots can be indexed such that they and the real phase angles $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$, are related through Eq. (3.121).

Proof. Let $\tau_\ell(\nu,\theta)$, $\ell = 1,2,3$, denote the roots of Eq. (3.119) where (ν,θ) is only subjected to the condition $\nu \neq 0$ and $0 < |\theta| < \pi$. Then, *no matter how these roots are assigned the indices* $\ell = 1,2,3$, we have

$$\tau^3 + 2\left[\nu \tan(\theta/2) - \frac{1}{\nu \sin\theta}\right]\tau^2 + \tau + 2\nu\tan(\theta/2) \equiv (\tau - \tau_1)(\tau - \tau_2)(\tau - \tau_3), \quad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.122}$$

Hereafter, for simplicity, the arguments ν and θ may be dropped from $\tau_\ell(\nu,\theta)$ and $\tan[\phi_\ell(\nu,\theta)/2]$, $\ell = 1,2,3$. Eq. (3.122) $\Leftrightarrow$

$$\tau_1\tau_2\tau_3 = -2\nu\tan(\theta/2), \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.123}$$

$$\tau_1\tau_2 + \tau_2\tau_3 + \tau_3\tau_1 = 1, \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.124}$$

and

$$\tau_1 + \tau_2 + \tau_3 = 2\left[\frac{1}{\nu\sin\theta} - \nu\tan(\theta/2)\right], \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.125}$$

Because $\nu\tan(\theta/2) \neq 0$ if $\nu \neq 0$ and $0 < |\theta| < \pi$, Eq. (3.123) implies that

$$\tau_1 \neq 0, \quad \tau_2 \neq 0, \quad \text{and} \quad \tau_3 \neq 0, \tag{3.126}$$

Moreover, it follows from Eqs. (3.123)–(3.125) that the roots of Eq. (3.119) are all real and can be indexed such that they are related to $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$, through Eq. (3.121), if and only if

$$\tan(\phi_1/2)\tan(\phi_2/2)\tan(\phi_3/2) = -2\nu\tan(\theta/2), \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.127}$$

$$\tan(\phi_1/2)\tan(\phi_2/2) + \tan(\phi_2/2)\tan(\phi_3/2) + \tan(\phi_3/2)\tan(\phi_1/2) = 1, \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.128}$$

and

$$\tan(\phi_1/2) + \tan(\phi_2/2) + \tan(\phi_3/2) = 2\left[\frac{1}{\nu\sin\theta} - \nu\tan(\theta/2)\right], \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.129}$$

In the following, first we will show that Eq. (3.119) has only real roots and they can be specified by Eq. (3.121), i.e., Eqs. (3.127)–(3.129) are true, if $(\nu, \theta) \in \Psi_o$.

As a preliminary, first note that Eq. (3.83) ⇔ Eqs. (3.84) and (3.86) (a conclusion reached following Eq. (3.88)). Next, because of Eq. (3.33), Eq. (3.84) ⇔ either (i) Eq. (3.89), or (ii)

$$\phi_1 + \phi_2 + \phi_3 = \pm\pi \tag{3.130}$$

It was shown earlier that Eq. (3.89) can occur only for the case Eq. (3.80). Because this case is ruled out by the condition $0 < |\theta| < \pi$, Eq. (3.84) ⇔ Eq. (3.130) if $0 < |\theta| < \pi$. Moreover, assuming Eq. (3.130), one can see easily that Eq. (3.86) ⇔ Eq. (3.94) ⇔ Eqs. (3.95) and (3.96). Thus one concludes that *Eq. (3.83) ⇔ Eqs. (3.95), (3.96), and (3.130) if* $0 < |\theta| < \pi$. Note that Eqs. (3.97)–(3.100) are trivial results of Eqs. (3.95), (3.96), and (3.130).

Let $(\nu, \theta) \in \Psi_o$. Then according to Eqs. (3.82) and (3.103), and the above discussions, we have (i) Eqs. (3.95), and (3.96), and (ii)

$$\nu \neq 0, \qquad 0 < |\theta| < \pi, \qquad 0 < |\phi_1|, |\phi_2|, |\phi_3| < \pi, \qquad \text{and} \qquad \phi_1 + \phi_2 + \phi_3 = \pm\pi \tag{3.131}$$

To prove Eq. (3.128), note that the last two conditions given in Eq. (3.131) imply that

$$\tan(\phi_3/2) = \tan\left(\pm\frac{\pi}{2} - \frac{\phi_1 + \phi_2}{2}\right) \tag{3.132}$$

(Note: $\tan(\phi_3/2)$ is undefined when $\phi_3 = \pm\pi, \pm 3\pi, \pm 5\pi, \ldots$. However these undefined cases are ruled out by the condition $0 < |\phi_3| < \pi$.) Eq. (3.128) follows immediately from Eq. (3.132) and the relation

$$\tan\left(\pm\frac{\pi}{2} - \frac{\phi_1 + \phi_2}{2}\right) \equiv \cot\left(\frac{\phi_1 + \phi_2}{2}\right) \equiv \frac{1 - \tan(\phi_1/2)\tan(\phi_2/2)}{\tan(\phi_1/2) + \tan(\phi_2/2)} \tag{3.133}$$

(Note: Because of the last two conditions given in Eq. (3.131), all terms and expressions which appear in Eq. (3.133) is well defined.)

To prove Eqs. (3.127) and (3.129), note that, because of Eq. (3.131), the term $2\nu\sin\theta$ on the right side of Eq. (3.96) is nonzero. Thus the expression on the left side is also nonzero, i.e.,

$$\sin\phi_1 + \sin\phi_2 + \sin(\phi_1 + \phi_2) \neq 0 \tag{3.134}$$

As such Eq. (3.96) ⇔

$$\frac{4}{\sin\phi_1 + \sin\phi_2 + \sin(\phi_1 + \phi_2)} = \frac{2}{\nu\sin\theta} \tag{3.135}$$

if Eq. (3.131) is assumed. Also by dividing Eq. (3.95) over (3.96), and using the last identity presented in Eq. (3.118), one has

$$\frac{\cos\phi_1 + \cos\phi_2 - \cos(\phi_1 + \phi_2) - 1}{\sin\phi_1 + \sin\phi_2 + \sin(\phi_1 + \phi_2)} = -2\nu\tan(\theta/2) \tag{3.136}$$

Adding Eq. (3.135) to Eq. (3.136), we have

$$\frac{\cos\phi_1 + \cos\phi_2 - \cos(\phi_1 + \phi_2) + 3}{\sin\phi_1 + \sin\phi_2 + \sin(\phi_1 + \phi_2)} = 2\left[\frac{1}{\nu\sin\theta} - \nu\tan(\theta/2)\right] \tag{3.137}$$

Furthermore, by assuming Eqs. (3.130) and (3.134), it is shown in Appendix B that

$$\frac{\cos\phi_1 + \cos\phi_2 - \cos(\phi_1+\phi_2) - 1}{\sin\phi_1 + \sin\phi_2 + \sin(\phi_1+\phi_2)} = \tan(\phi_1/2)\tan(\phi_2/2)\tan(\phi_3/2) \tag{3.138}$$

and

$$\frac{\cos\phi_1 + \cos\phi_2 - \cos(\phi_1+\phi_2) + 3}{\sin\phi_1 + \sin\phi_2 + \sin(\phi_1+\phi_2)} = \tan(\phi_1/2) + \tan(\phi_2/2) + \tan(\phi_3/2) \tag{3.139}$$

Eqs. (3.127) and (3.129) now follow from Eqs. (3.136)–(3.139). Thus we have shown that the roots of Eq. (3.119) are the real roots specified by Eq. (3.121), if $(\nu,\theta)\Psi_o$.

Next consider any (ν,θ) such that (i) $\nu \neq 0$ and $0 < |\theta| < \pi$; and (ii) the roots of Eq. (3.119) are all real. In the following we will complete the proof by showing that, for such a (ν,θ), (i) Eq. (3.83) is true, i.e., $(\nu,\theta) \in \Psi_o$; and (ii) the real roots of Eq. (3.119) can be indexed such that they and the real phase angles $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$, are related by Eq. (3.121).

Let the real roots be denoted by $\tau_\ell(\nu,\theta)$, $\ell = 1,2,3$. Then we have Eqs. (3.123)–(3.126). Because of Eq. (3.126), for each $\tau_\ell(\nu,\theta)$, there exists one and only one $\varphi_\ell(\nu,\theta)$ such that

$$0 < |\varphi_1(\nu,\theta)|, |\varphi_2(\nu,\theta)|, |\varphi_3(\nu,\theta)| < \pi \tag{3.140}$$

and

$$\tau_\ell(\nu,\theta) = \tan\left[\varphi_\ell(\nu,\theta)/2\right], \qquad \ell = 1,2,3 \tag{3.141}$$

Substituting Eq. (3.141) into Eq. (3.123)–(3.125) and dropping the arguments ν and θ, we have

$$\tan(\varphi_1/2)\tan(\varphi_2/2)\tan(\varphi_3/2) = -2\nu\tan(\theta/2), \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.142}$$

$$\tan(\varphi_1/2)\tan(\varphi_2/2) + \tan(\varphi_2/2)\tan(\varphi_3/2) + \tan(\varphi_3/2)\tan(\varphi_1/2) = 1, \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.143}$$

and

$$\tan(\varphi_1/2) + \tan(\varphi_2/2) + \tan(\varphi_3/2) = 2\left[\frac{1}{\nu\sin\theta} - \nu\tan(\theta/2)\right], \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.144}$$

Because of Eq. (3.140), at least two of φ_1, φ_2, and φ_3 must both be positive or negative. Let φ_1 and φ_2 be both positive or negative, i.e., $\varphi_1\varphi_2 > 0$. Then Eq. (3.140) also implies that $\tan(\varphi_1/2)\tan(\varphi_2/2) > 0$. In turn, we have

$$\varphi_1 + \varphi_2 \neq 0 \tag{3.145}$$

and

$$\tan(\varphi_1/2) + \tan(\varphi_2/2) \neq 0 \tag{3.146}$$

(Note: Recall that $(a+b)^2 \equiv a^2 + b^2 + 2ab$. Thus $(a+b)^2 > 0$, i.e., $a+b \neq 0$, if $ab > 0$.). Next, by combining Eqs. (3.140) and (3.145), one has

$$0 < \left|\frac{\varphi_1+\varphi_2}{2}\right| < \pi \tag{3.147}$$

and therefore

$$\sin(\frac{\varphi_1+\varphi_2}{2}) \neq 0 \tag{3.148}$$

By using Eq. (3.140) and (3.148), one can show easily that

$$\cot\left(\frac{\varphi_1+\varphi_2}{2}\right) = \frac{1 - \tan(\varphi_1/2)\tan(\varphi_2/2)}{\tan(\varphi_1/2) + \tan(\varphi_2/2)} \tag{3.149}$$

Moreover, with the aid of Eq. (3.146), Eq. (3.143) implies that

$$\tan(\varphi_3/2) = \frac{1 - \tan(\varphi_1/2)\tan(\varphi_2/2)}{\tan(\varphi_1/2) + \tan(\varphi_2/2)} \tag{3.150}$$

Combining Eqs. (3.149) and (3.150), one arrives at the conclusion that

$$\tan(\varphi_3/2) = \cot\left(\frac{\varphi_1 + \varphi_2}{2}\right) \tag{3.151}$$

Eq. (3.151) implies that φ_1, φ_2, and φ_3 must satisfy one of the following conditions:

$$\varphi_1 + \varphi_2 + \varphi_3 = m\pi, \qquad m = \pm 1, \pm 3, \pm 5, \ldots \tag{3.152}$$

Because $0 < |\varphi_1 + \varphi_2 + \varphi_3| < 3\pi$ is required by Eq. (3.140), Eq. (3.152) now implies that

$$\varphi_1 + \varphi_2 + \varphi_3 = \pm\pi \tag{3.153}$$

Note that, using similar arguments, we will arrive at the same conclusion Eq. (3.153) if, instead of assuming $\varphi_1\varphi_2 > 0$ at the beginning, we assume $\varphi_2\varphi_3 > 0$ or $\varphi_3\varphi_1 > 0$.

To proceed, note that: (i) by combining Eqs. (3.140) and (3.153) with the assumptions $\nu \neq 0$ and $0 < |\theta| < \pi$, we have

$$\nu \neq 0, \qquad 0 < |\theta| < \pi, \qquad 0 < |\varphi_1|, |\varphi_2|, |\varphi_3| < \pi, \qquad \text{and} \qquad \varphi_1 + \varphi_2 + \varphi_3 = \pm\pi \tag{3.154}$$

and (ii)

$$\sin\varphi_1 + \sin\varphi_2 + \sin(\varphi_1 + \varphi_2) = \pm 4\cos(\varphi_1/2)\cos(\varphi_2/2)\cos(\varphi_3/2) \tag{3.155}$$

is a result of Eq. (3.153) (see the proof given in Appendix B). By combining Eq. (3.155) with a result of Eq. (3.140), i.e.,

$$\cos(\varphi_\ell/2) > 0, \qquad \ell = 1, 2, 3 \tag{3.156}$$

one concludes that

$$\sin\varphi_1 + \sin\varphi_2 + \sin(\varphi_1 + \varphi_2) \neq 0 \tag{3.157}$$

is a result of Eq. (3.140) and (3.153). Using arguments similar to those used in the proof of Eqs. (3.138) and (3.139) (see Appendix B), one can prove that

$$\frac{\cos\varphi_1 + \cos\varphi_2 - \cos(\varphi_1 + \varphi_2) - 1}{\sin\varphi_1 + \sin\varphi_2 + \sin(\varphi_1 + \varphi_2)} = \tan(\varphi_1/2)\tan(\varphi_2/2)\tan(\varphi_3/2) \tag{3.158}$$

and

$$\frac{\cos\varphi_1 + \cos\varphi_2 - \cos(\varphi_1 + \varphi_2) + 3}{\sin\varphi_1 + \sin\varphi_2 + \sin(\varphi_1 + \varphi_2)} = \tan(\varphi_1/2) + \tan(\varphi_2/2) + \tan(\varphi_3/2) \tag{3.159}$$

follows from Eqs. (3.153) and (3.157).

Eqs. (3.142), (3.144), (3.158), and (3.159) now imply that

$$\frac{\cos\varphi_1 + \cos\varphi_2 - \cos(\varphi_1 + \varphi_2) - 1}{\sin\varphi_1 + \sin\varphi_2 + \sin(\varphi_1 + \varphi_2)} = -2\nu\tan(\theta/2) \tag{3.160}$$

and

$$\frac{\cos\varphi_1 + \cos\varphi_2 - \cos(\varphi_1 + \varphi_2) + 3}{\sin\varphi_1 + \sin\varphi_2 + \sin(\varphi_1 + \varphi_2)} = 2\left[\frac{1}{\nu\sin\theta} - \nu\tan(\theta/2)\right] \tag{3.161}$$

By subtracting Eq. (3.160) from Eq. (3.161), and then taking the reciprocals of the expressions on the both sides of the resulting equation, one has

$$\sin\varphi_1 + \sin\varphi_2 + \sin(\varphi_1+\varphi_2) = 2\nu\sin\theta \tag{3.162}$$

Moreover, by substituting Eq. (3.162) into Eq. (3.160) and using the last identity presented in Eq. (3.118), one has

$$\cos\varphi_1 + \cos\varphi_2 - \cos(\varphi_1+\varphi_2) - 1 = -4\nu^2(1-\cos\theta) \tag{3.163}$$

At this juncture, note that Eqs. (3.153), (3.154), (3.162), (3.163) will become Eqs. (3.130), (3.131), (3.96), and (3.95), respectively, if the symbols φ_1, φ_2, and φ_3 in the former equations are replaced by ϕ_1, ϕ_2, and ϕ_3, respectively.

Next, by using Eq. (3.153), (3.162), and (3.163), one has

$$e^{i(\varphi_1+\varphi_2+\varphi_3)} = -1 \tag{3.164}$$

and

$$e^{i\varphi_1} + e^{i\varphi_2} + e^{i\varphi_3} = 1 - 4\nu^2(1-\cos\theta) + 2i\nu\sin\theta \tag{3.165}$$

Because Eq. (3.85) is a result of Eqs. (3.84) and (3.86), one can see that

$$e^{i(\varphi_1+\varphi_2)} + e^{i(\varphi_1+\varphi_3)} + e^{i(\varphi_2+\varphi_3)} = -1 + 4\nu^2(1-\cos\theta) + 2i\nu\sin\theta \tag{3.166}$$

is a result of Eqs. (3.164) and (3.165). By comparing Eqs. (3.164)–(3.166) with Eq. (3.22), one concludes that the roots of Eq. (3.17) are $e^{i\varphi_\ell}$, $\ell = 1,2,3$. Thus, according to Eqs. (3.31)–(3.33) and (3.140), Eq. (3.83) is true, and one can choose $\phi_\ell \overset{\text{def}}{=} \varphi_\ell$, $\ell = 1,2,3$. As such, it has been shown that, for any (ν,θ) such that (i) $\nu \neq 0$ and $0 < |\theta| < \pi$; and (ii) the roots of Eq. (3.119) are all real, Eq. (3.83) is true and, through a proper indexing, the real roots $\tau_\ell(\nu,\theta)$, $\ell = 1,2,3$, of Eq. (3.119) and the real phase angle $\phi_\ell(\nu,\theta)$, $\ell = 1,2,3$, of the roots to Eq. (3.17) are related through Eq. (3.121). Thus the proof for Proposition 2 is completed. QED.

According to Eqs. (3.62) and (3.63), for the special cases $\nu = \pm 1/2$, (i) Eq. (3.83) is true in the domain $-\pi < \theta \le \pi$; and (ii) in the domain $0 < |\theta| < \pi$, the phase angles ϕ_ℓ, $\ell = 1,2,3$ (which are subjected to the condition Eq. (3.33)), can be chosen as

$$\phi_1 = \pm\theta, \quad \phi_2 = \mp\theta/2, \quad \text{and} \quad \phi_3 = \begin{cases} \pm(\pi-\theta/2) & \text{if } \pi > \theta > 0 \\ \mp(\pi+\theta/2) & \text{if } 0 > \theta > -\pi \end{cases}, \quad \nu = \pm 1/2;\ 0 < |\theta| < \pi \tag{3.167}$$

(Note: Hereafter, for Eq. (3.167) and similar equations associated with the special cases $\nu = \pm 1/2$, each equation is valid when the upper (lower) signs are taken uniformly.) According to Eq. (3.121) and (3.167), the roots to Eq. (3.119) for the current special cases are

$$\tau_1 = \pm\tan(\theta/2), \quad \tau_2 = \mp\tan(\theta/4), \quad \text{and} \quad \tau_3 = \pm\cot(\theta/4) \qquad \nu = \pm 1/2;\ 0 < |\theta| < \pi \tag{3.168}$$

In fact, by using the relations

$$-\tan(\theta/2) + \tan(\theta/4) - \cot(\theta/4) = \tan(\theta/2) - \frac{4}{\sin\theta}, \qquad 0 < |\theta| < \pi \tag{3.169}$$

and

$$\tan(\theta/2)\left[\cot(\theta/4) - \tan(\theta/4)\right] = 2, \qquad 0 < |\theta| < \pi \tag{3.170}$$

(which are proved in Appendix B), Eq. (3.168) implies that

$$\tau_1+\tau_2+\tau_3 = \pm\left[\frac{4}{\sin\theta} - \tan(\theta/2)\right], \quad \tau_1\tau_2+\tau_2\tau_3+\tau_3\tau_1 = 1, \quad \text{and} \quad \tau_1\tau_2\tau_3 = \mp\tan(\theta/2) \tag{3.171}$$

In other words, τ_ℓ, $\ell = 1, 2, 3$, given in Eq. (3.168) indeed satisfy Eqs. (3.123)–(3.125) for the special cases $\nu = \pm 1/2$.

According to Eq. (3.168), we have (i)

$$\tau_2 \neq \tau_1 \quad \text{and} \quad \tau_2 \neq \tau_3, \qquad\qquad \nu = \pm 1/2;\ 0 < |\theta| < \pi \tag{3.172}$$

(ii)

$$\tau_1 \neq \tau_3 \quad \text{if} \quad \nu = \pm 1/2,\ 0 < |\theta| < \pi,\ \text{and}\ |\theta| \neq 2\pi/3 \tag{3.173}$$

and (iii)

$$\tau_1 = \tau_3 \quad \text{if} \quad \nu = \pm 1/2 \text{ and } |\theta| = 2\pi/3 \tag{3.174}$$

As a result, for the special cases $\nu = \pm 1/2$, Eq. (3.119) has: (i) three distinct real roots if $0 < |\theta| < \pi$ and $|\theta| \neq 2\pi/3$; and (ii) one doubt real root (i.e., $\tau_1 = \tau_3$) and one simple real root (i.e., τ_2) if $|\theta| = 2\pi/3$.

3.7. Proof for Proposition 1

As a preliminary, we introduce the following well-known theorem:

Theorem 5. Consider the cubic equation

$$\tau^3 + a_2\tau^2 + a_1\tau + a_0 = 0 \tag{3.175}$$

where a_0, a_1, and a_2 are real coefficients. Let

$$q \overset{\text{def}}{=} \frac{a_1}{3} - \frac{(a_2)^2}{9}, \quad r \overset{\text{def}}{=} \frac{a_1 a_2 - 3a_0}{6} - \frac{(a_2)^3}{27}, \quad \text{and} \quad D \overset{\text{def}}{=} q^3 + r^2 \tag{3.176}$$

Then Eq. (3.175) has: (i) one real root and a pair of complex conjugate roots if $D > 0$; (ii) three real roots and at least two are equal if $D = 0$; and (iii) three distinct real roots if $D < 0$.

For each (ν, θ), Eq. (3.119) is a special case of Eq. (3.175) with

$$a_0 \overset{\text{def}}{=} 2\nu\tan(\theta/2), \quad a_1 \overset{\text{def}}{=} 1, \quad \text{and} \quad a_2 \overset{\text{def}}{=} 2\left[\nu\tan(\theta/2) - \frac{1}{\nu\sin\theta}\right], \quad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.177}$$

Thus, for each (ν, θ), the discriminant D associated with Eq. (3.119) has the form

$$\begin{aligned}
D(\nu,\theta) &= \left\{\frac{1}{3} - \frac{4}{9}\left[\nu\tan(\theta/2) - \frac{1}{\nu\sin\theta}\right]^2\right\}^3 + \left\{\frac{1}{3}\left[2\nu\tan(\theta/2) + \frac{1}{\nu\sin\theta}\right] + \frac{8}{27}\left[\nu\tan(\theta/2) - \frac{1}{\nu\sin\theta}\right]^3\right\}^2 \\
&= \frac{1}{27} - \frac{4}{27}\left[\nu\tan(\theta/2) - \frac{1}{\nu\sin\theta}\right]^2 + \frac{16}{81}\left[\nu\tan(\theta/2) - \frac{1}{\nu\sin\theta}\right]^4 \\
&\quad + \frac{1}{9}\left[2\nu\tan(\theta/2) + \frac{1}{\nu\sin\theta}\right]^2 + \frac{16}{81}\left[2\nu\tan(\theta/2) + \frac{1}{\nu\sin\theta}\right]\left[\nu\tan(\theta/2) - \frac{1}{\nu\sin\theta}\right]^3 \\
&= -\frac{1}{27\nu^2\sin^2\theta}\left[1 + \frac{16\tan(\theta/2)}{\sin\theta}\right] + \frac{1}{27}\left[1 + \frac{48\tan^2(\theta/2)}{\sin^2\theta} + \frac{20\tan(\theta/2)}{\sin\theta}\right] \\
&\quad + \frac{8\nu^2\tan^2(\theta/2)}{27}\left[1 - \frac{6\tan(\theta/2)}{\sin\theta}\right] + \frac{16\nu^4}{27}\tan^4(\theta/2), \qquad \nu \neq 0;\ 0 < |\theta| < \pi
\end{aligned} \tag{3.178}$$

Let

$$s \overset{\text{def}}{=} \nu^2 \tag{3.179}$$

Then Eq. (3.178) and the relation

$$\frac{\tan(\theta/2)}{\sin\theta} = \frac{\sin(\theta/2)/\cos(\theta/2)}{2\sin(\theta/2)\cos(\theta/2)} = \frac{1}{2}\sec^2(\theta/2), \qquad 0 < |\theta| < \pi \tag{3.180}$$

imply that

$$\begin{aligned}\eta(s,\theta) \overset{\text{def}}{=} \frac{27sD(\nu,\theta)\sin^2\theta}{16} &= \big[\tan^4(\theta/2)\sin^2\theta\big]s^3 + \frac{1}{2}\tan^2(\theta/2)\sin^2\theta\big[1 - 3\sec^2(\theta/2)\big]s^2 \\ &+ \frac{1}{16}\sin^2\theta\big[1 + 12\sec^4(\theta/2) + 10\sec^2(\theta/2)\big]s - \frac{1}{16}\big[1 + 8\sec^2(\theta/2)\big], \quad s > 0;\ 0 < |\theta| < \pi\end{aligned} \tag{3.181}$$

For the special cases $\nu = \pm 1/2$, i.e., $s = 1/4$, $\eta(s,\theta)$ reduces to

$$\begin{aligned}\eta(1/4,\theta) = \frac{1}{64}\Big\{&\tan^4(\theta/2)\sin^2\theta + 2\tan^2(\theta/2)\sin^2\theta\big[1 - 3\sec^2(\theta/2)\big] \\ &+ \sin^2\theta\big[1 + 12\sec^4(\theta/2) + 10\sec^2(\theta/2)\big] - 4\big[1 + 8\sec^2(\theta/2)\big]\Big\}, \quad 0 < |\theta| < \pi\end{aligned} \tag{3.182}$$

By using trigonometric relations such as

$$\tan^2(\theta/2) = \sec^2(\theta/2) - 1 \quad \text{and} \quad \sec^2(\theta/2)\sin^2\theta = 4\sin^2(\theta/2) \tag{3.183}$$

Eq. (3.182) can be simplified as

$$\eta(1/4,\theta) = -\frac{1}{16}\left[\frac{4\cos^2(\theta/2) - 1}{\cos(\theta/2)}\right]^2, \qquad 0 < |\theta| < \pi \tag{3.184}$$

Because (i) $|\theta| < \pi \Leftrightarrow |\theta/2| < \pi/2$; (ii) $4\cos^2(\theta/2) = 1 \Leftrightarrow \cos(\theta/2) = 1/2$ if $|\theta/2| < \pi/2$; and (iii) $\cos(\theta/2) = 1/2 \Leftrightarrow |\theta/2| = \pi/3$ if $|\theta/2| < \pi/2$, Eq.(3.184) implies that

$$\eta(1/4,\theta) = \begin{cases} < 0 & \text{if } 0 < |\theta| < \pi \text{ and } |\theta| \neq 2\pi/3 \\ = 0 & \text{if } |\theta| = 2\pi/3 \end{cases} \tag{3.185}$$

With the aid of Eq. (3.179) and the definition of $\eta(s,\theta)$ given in Eq. (3.181), Eq. (3.185) and Theorem 5 imply that, for the case with $\nu = \pm 1/2$, Eq. (3.119) has: (i) three distinct real roots if $0 < |\theta| < \pi$ and $|\theta| \neq 2\pi/3$; and (ii) three real roots and at least two are equal if $|\theta| = 2\pi/3$. This result is consistent with the conclusion reached following Eq. (3.174).

Let

$$A(\theta) \overset{\text{def}}{=} 3\tan^4(\theta/2)\sin^2\theta, \qquad 0 < |\theta| < \pi \tag{3.186}$$

$$B(\theta) \overset{\text{def}}{=} \tan^2(\theta/2)\sin^2\theta\left[1 - 3\sec^2(\theta/2)\right], \qquad 0 < |\theta| < \pi \tag{3.187}$$

and

$$C(\theta) \overset{\text{def}}{=} \frac{1}{16}\sin^2\theta\left[1 + 12\sec^4(\theta/2) + 10\sec^2(\theta/2)\right], \qquad 0 < |\theta| < \pi \tag{3.188}$$

Then (i)

$$A(\theta) > 0, \quad B(\theta) < 0, \quad \text{and} \quad C(\theta) > 0, \qquad 0 < |\theta| < \pi \tag{3.189}$$

(ii)

$$\frac{\partial\eta(s,\theta)}{\partial s} = A(\theta)s^2 + B(\theta)s + C(\theta) = A(\theta)\left\{\left[s + \frac{B(\theta)}{2A(\theta)}\right]^2 + \frac{4A(\theta)C(\theta) - [B(\theta)]^2}{4[A(\theta)]^2}\right\}, \ s > 0;\ 0 < |\theta| < \pi \tag{3.190}$$

and (iii) because $\sec^2(\theta/2) > 1$, $0 < |\theta| < \pi$,

$$4A(\theta)C(\theta) - [B(\theta)]^2 = \tan^4(\theta/2)\sin^4\theta\left[(27/2)\sec^2(\theta/2) - (1/4)\right] > 0, \qquad 0 < |\theta| < \pi \tag{3.191}$$

Eqs. (3.189)–(3.191) now imply that

$$\frac{\partial\eta(s,\theta)}{\partial s} > 0, \qquad s > 0;\ 0 < |\theta| < \pi \tag{3.192}$$

Combining Eqs. (3.185) and (3.192), one arrives at the conclusion that (i)

$$\eta(s,\theta) < 0 \qquad \text{if} \quad 0 < s < 1/4 \text{ and } 0 < |\theta| < \pi \tag{3.193}$$

and (ii)

$$\eta(s,\theta) > 0 \qquad \text{if} \quad s > 1/4 \text{ and } |\theta| = 2\pi/3 \tag{3.194}$$

By using Eqs. (3.185), (3.193), (3.179), and (3.181), one concludes that

$$D(\nu,\theta)\begin{cases} < 0 & \text{if } 0 < |\nu| < 1/2 \text{ and } 0 < |\theta| < \pi \\ < 0 & \text{if } |\nu| = 1/2,\ 0 < |\theta| < \pi, \text{ and } |\theta| \neq 2\pi/3 \\ = 0 & \text{if } |\nu| = 1/2 \text{ and } |\theta| = 2\pi/3 \end{cases} \tag{3.195}$$

With the aid of Theorem 5, Eq. (3.195) infers that the roots of Eq. (3.119) are real and distinct if (i) $0 < |\nu| < 1/2$ and $0 < |\theta| < \pi$, and also if (ii) $|\nu| = 1/2$, $0 < |\theta| < \pi$, and $|\theta| \neq 2\pi/3$. Moreover, it infers that these roots are real and at least two of them are equal if $|\nu| = 1/2$ and $|\theta| = 2\pi/3$. By using Proposition 2, in turn, one concludes that $\sigma_\ell(\nu,\theta)$, $\ell = 1,2,3$, are distinct and of unit magnitude if (i) $0 < |\nu| < 1/2$ and $0 < |\theta| < \pi$, and also if (ii) $|\nu| = 1/2$, $0 < |\theta| < \pi$, and $|\theta| \neq 2\pi/3$. Moreover, one concludes that $\sigma_\ell(\nu,\theta)$, $\ell = 1,2,3$, are of unit magnitude and at least two of them are equal if $|\nu| = 1/2$ and $|\theta| = 2\pi/3$.

Proposition 1(a) now follows from the above conclusions and several results obtained in Sec. 3.5, i.e., (i) the roots of Eq. (3.17) are 1, 1, and -1 if $\nu = 0$ or $\theta = 0$; and (ii) the roots of Eq. (3.17) are -1 and two distinct complex conjugate numbers of unit magnitude if $0 < |\nu| < 1/\sqrt{2}$ and $\theta = \pi$ (see Eq. (3.77) and (3.79)). On the other hand, Proposition 1(b) is a trivial result of (i) Eqs. (3.194), (3.181), and (3.179); (ii) Theorem 5; and (iii) Proposition 2. Thus the proof for Proposition 1 is completed. QED.

In Sec. 3.8, we introduce and prove Proposition 3, which along with the results presented in Sec. 3.5, defines the sets of (ν,θ) for which Eq. (3.17), respectively, has (i) three distinct roots of unit magnitude, (ii) one double root of unit magnitude and one simple root of unit magnitude, (iii) one triple root of unit magnitude, and (iv) at least one root not of unit magnitude.

3.8. Proposition 3

Note that, for any given θ with $0 < |\theta| < \pi$, (i) Eqs. (3.192) and (3.193) imply that $\eta(s,\theta)$ is a strictly monotonically increasing function of s in the domain $s > 0$ and it becomes negative uniformly in the domain $0 < s < 1/4$; and (ii) the coefficient $\tan^4(\theta/2)\sin^2\theta$ of the third-order term in s in the expression on the right side of Eq. (3.181) is positive and thus $\eta(s,\theta) \to +\infty$ as $s \to +\infty$. The above observations coupled with Eq. (3.179) imply that, for a given θ with $0 < |\theta| < \pi$, (i) there is one and only one positive value $\nu^*(\theta)$ such that

$$\eta(\nu^2,\theta)\begin{cases} < 0 & \text{if } |\nu| < \nu^*(\theta) \\ = 0 & \text{if } |\nu| = \nu^*(\theta) \\ > 0 & \text{if } |\nu| > \nu^*(\theta) \end{cases} \qquad \nu \neq 0;\ 0 < |\theta| < \pi \tag{3.196}$$

and (ii)

$$\nu^*(\theta)\begin{cases} = 1/2 & \text{if } |\theta| = 2\pi/3 \\ > 1/2 & \text{if } 0 < |\theta| < \pi \text{ and } |\theta| \neq 2\pi/3 \end{cases} \tag{3.197}$$

By using (i) Eqs. (3.196) and (3.197), (ii) Eqs. (3.179) and (3.181), (iii) Theorem 5, (iv) Proposition 2, (v) the relation $\sigma_1(\nu,\theta)\sigma_2(\nu,\theta)\sigma_3(\nu,\theta) = -1$ (see Eq. (3.22)), and (vi) the fact that the only possible triple root of unit magnitude for Eq. (3.17) is -1 and it occurs only for the case Eq. (3.80) (proved in Sec. 3.6), one now arrives at Proposition 3:

Proposition 3. For a given θ with $0 < |\theta| < \pi$, Eq. (3.17) has: (i) three distinct roots of unit magnitude if $0 < |\nu| < \nu^*(\theta)$; (ii) one double root of unit magnitude and a simple root of unit magnitude if $|\nu| = \nu^*(\theta)$; and (iii) at least one root with its magnitude > 1 if $|\nu| > \nu^*(\theta)$.

An immediate result of Eq. (3.197) and Proposition 3 is the following corollary:

Corollary to Proposition 3. For a given θ with $0 < |\theta| < \pi$, Eq. (3.17) has: (i) three distinct roots of unit magnitude if (a) $0 < |\nu| < 1/2$, and also if (b) $|\nu| = 1/2$ and $|\theta| \neq 2\pi/3$; (ii) one double root of unit magnitude and a simple root of unit magnitude if $|\nu| = 1/2$ and $|\theta| = 2\pi/3$; and (iii) at least one root with its magnitude > 1 if $|\nu| > 1/2$ and $|\theta| = 2\pi/3$.

With the above preliminaries, a rigorous study of the stability conditions of the $a(3)$ scheme will be given in Sec. 3.9.

3.9. Stability condition of the $a(3)$ scheme

Because of Eq. (3.1) and a reason presented right before Sec. 3.1, we have the following definition:

Definition 1. The $a(3)$ scheme is said to be stable with respect to a given ν if, for every θ, $-\pi < \theta \leq \pi$, every element of the matrix $[G(\nu,\theta)]^m$ remains bounded as the positive integer $m \to +\infty$. On the other hand, the scheme is said to be unstable with respect to a given ν if, for any θ, $-\pi < \theta \leq \pi$, at least one element of the matrix $[G(\nu,\theta)]^m$ becomes unbounded as $m \to +\infty$.

To study stability of the $a(3)$ scheme, in the following we introduce needed matrix preliminaries. First note that an $N \times N$ matrix has at least one eigenvector and at most N linearly independent eigenvectors [76]. Related to this subject, we have Definition 2 [76]:

Definition 2. An $N \times N$ matrix A is said to be nondefective if it admits N linearly independent eigenvectors. On the other hand, A is defective if it admits less than N linearly independent eigenvectors.

Another definition we need later is Definition 3:

Definition 3. Let λ_ℓ, $\ell = 1, 2, 3, \ldots, N$, be the eigenvalues (which may or may not coincide with one another) of an $N \times N$ matrix A. Then the spectral radius of A is

$$\rho(A) \overset{\text{def}}{=} \max_{\ell=1}^{N}\{|\lambda_\ell|\} \tag{3.198}$$

Next we have the following well-established theorem [76]:

Theorem 6. For any $N \times N$ matrix, (i) each distinct eigenvalue of multiplicity m is associated with at least one eigenvector and at most m linear independent eigenvectors; and (ii) two eigenvectors associated with two different eigenvalues are linearly independent.

By using Definition 2 along with Theorems 4 and 6, we have Theorem 7, i.e.,

Theorem 7. An $N \times N$ matrix A is defective if and only if A has at least one eigenvalue which satisfies the following properties: (i) its multiplicity m is greater than one; and (ii) the number of linearly independent eigenvectors associated with this eigenvalue is less than m.

Moreover, with the aid of Theorems 2 and 3, one can easily prove Theorem 8:

Theorem 8. Let (i) A be an $N \times N$ matrix, (ii) $\overline{A}$ be the complex conjugate of A, and (iii) B be a matrix similar to A (i.e., there exist a nonsingular $N \times N$ matrix S such that $B = S^{-1}AS$). Then A is defective (nondefective) $\Leftrightarrow$ $\overline{A}$ is defective (nondefective) $\Leftrightarrow$ B is defective (nondefective).

An immediate result of Theorem 7 is Lemma 3, i.e.,

Lemma 3. An $N \times N$ diagonal matrix is nondefective.

Next we will prove Theorem 9, i.e.,

Theorem 9. Let A be a 3×3 matrix. Then every element of A^m remains bounded as the positive integer $m \to +\infty$ if and only if

$$\rho(A) \begin{cases} \le 1 & \text{if } A \text{ is nondefective} \\ < 1 & \text{if } A \text{ is defective} \end{cases} \tag{3.199}$$

Proof. Let A be nondefective. Then the Jordan canonical form theorem [76] implies that there is a nonsingular 3×3 matrix S so that $A = S\Lambda_0 S^{-1}$ where

$$\Lambda_0 \overset{\text{def}}{=} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \tag{3.200}$$

with λ_1, λ_2, and λ_3 being the eigenvalues of A. Because $A = S\Lambda_0 S^{-1}$ implies that $A^m = S(\Lambda_0)^m S^{-1}$, every element of A remains bounded as $m \to +\infty \Leftrightarrow$ every element of Λ_0 remain bounded as $m \to +\infty$. As such, for the nondefective case, Theorem 9 now follow from Eq. (3.198) and the fact that (i)

$$(\Lambda_0)^m = \begin{pmatrix} (\lambda_1)^m & 0 & 0 \\ 0 & (\lambda_2)^m & 0 \\ 0 & 0 & (\lambda_3)^m \end{pmatrix} \qquad m = 1, 2, 3 \ldots \tag{3.201}$$

and (ii) for a complex number c

$$\lim_{m \to +\infty} |c^m| \begin{cases} \le 1 & \text{if } |c| \le 1 \\ = +\infty & \text{if } |c| > 1 \end{cases} \tag{3.202}$$

Next let A be defective. Then, according to Theorems 4, 6, and 7, it must belong to one of the following mutually exclusive cases: (a) it has an eigenvalue λ of multiplicity $m = 3$ and it admits one and only one linearly independent eigenvector; (b) it has an eigenvalue λ of $m = 3$ and it admits two and only two linearly independent eigenvectors; and (c) it has an eigenvalue λ_1 of $m = 1$ and another eigenvalue λ_2 of $m = 2$ such that there is one and only one linearly independent eigenvectors associated with the eigenvalue λ_2. According to the Jordan canonical form theorem, for each case, there exists a nonsingular 3×3 matrix S such that $A = S\Lambda S^{-1}$ where (i) for case (a),

$$\Lambda = \Lambda_1 \overset{\text{def}}{=} \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \tag{3.203}$$

(ii) for case (b),

$$\Lambda = \Lambda_2 \overset{\text{def}}{=} \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \tag{3.204}$$

and (iii) for case (c),

$$\Lambda = \Lambda_3 \overset{\text{def}}{=} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 1 \\ 0 & 0 & \lambda_2 \end{pmatrix} \tag{3.205}$$

By induction, one can prove easily the following relations:

$$(\Lambda_1)^m = \begin{pmatrix} \lambda^m & m\lambda^{m-1} & [m(m-1)/2]\lambda^{m-2} \\ 0 & \lambda^m & m\lambda^{m-1} \\ 0 & 0 & \lambda^m \end{pmatrix} \qquad m = 1, 2, 3, \ldots \tag{3.206}$$

$$(\Lambda_2)^m = \begin{pmatrix} \lambda^m & 0 & 0 \\ 0 & \lambda^m & m\lambda^{m-1} \\ 0 & 0 & \lambda^m \end{pmatrix} \qquad m = 1, 2, 3, \ldots \tag{3.207}$$

and

$$(\Lambda_3)^m = \begin{pmatrix} (\lambda_1)^m & 0 & 0 \\ 0 & (\lambda_2)^m & m(\lambda_2)^{m-1} \\ 0 & 0 & (\lambda_2)^m \end{pmatrix} \qquad m = 1, 2, 3, \ldots \tag{3.208}$$

Because $A = S\Lambda S^{-1}$ implies that $A^m = S(\Lambda)^m S^{-1}$, every element of A remains bounded as $m \to +\infty \Leftrightarrow$ every element of Λ remain bounded as $m \to +\infty$. As such, for the defective case, Theorem 9 now follows from (i) Eqs. (3.198) and (3.203)–(3.208), and (ii) the fact that Eq. (3.202) and the relations

$$\lim_{m\to+\infty} |mc^{m-1}| = \begin{cases} 0 & \text{if } |c| < 1 \\ +\infty & \text{if } |c| \geq 1 \end{cases} \tag{3.209}$$

and

$$\lim_{m\to+\infty} |[m(m-1)/2]c^{m-2}| = \begin{cases} 0 & \text{if } |c| < 1 \\ +\infty & \text{if } |c| \geq 1 \end{cases} \tag{3.210}$$

are true for any complex number c. QED.

An immediate result of Definition 1 and Theorem 9 is Lemma 4:

Lemma 4 The $a(3)$ scheme is stable with respect to a given ν if and only if, for this ν and every θ, $-\pi < \theta \leq \pi$,

$$\rho\left(G(\nu,\theta)\right) \begin{cases} \leq 1 & \text{if } G(\nu,\theta) \text{ is nondefective} \\ < 1 & \text{if } G(\nu,\theta) \text{ is defective} \end{cases} \qquad -\pi < \theta \leq \pi \tag{3.211}$$

Because $\sigma_\ell(\nu,\theta)$, $\ell = 1, 2, 3$, are the eigenvalues of $G(\nu,\theta)$, with the aid of Proposition 1(b) (or part (iii) of Corollary to Proposition 3) and Eq. (3.198), Lemma 4 implies that the $a(3)$ scheme is unstable if $|\nu| > 1/2$. In the following, we will prove Proposition 4, i.e.,

Proposition 4. The $a(3)$ scheme (i) is stable if and only if

$$|\nu| < 1/2 \tag{3.212}$$

(ii) is neutrally stable for any ν satisfying Eq. (3.212); and (iii) is *linearly unstable* (in a sense to be defined) if $|\nu| = 1/2$.

As a preliminary, first we will study defectiveness of $G(\nu,\theta)$ over several domains of (ν,θ). We begin with Lemma 5:

Lemma 5. (i) $G(\nu, 0)$ is nondefective for any ν; (ii) $G(\nu,\pi)$ is defective if and only if $|\nu| = 1/\sqrt{2}$; and (iii) $G(0,\theta)$, $-\pi < \theta \leq \pi$, is nondefective.

Proof. Because

$$G(\nu, 0) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.213}$$

Part (i) follows from Lemma 3 and Eq. (3.213) immediately.

Next, by using Eq. (3.76), one concludes that $G(\nu,\pi)$ has an eigenvalue with multiplicity $m > 1$ if and only if either (i) the two eigenvalues given in Eq. (3.77) are equal, i.e.,

$$\nu^2(2\nu^2 - 1) = 0 \tag{3.214}$$

or (ii) the eigenvalue -1 is equal to one of those given in Eq. (3.77), i.e.,

$$1 - 2\nu^2 = \sqrt{2\nu^2(2\nu^2-1)} \quad \text{or} \quad 1 - 2\nu^2 = -\sqrt{2\nu^2(2\nu^2-1)} \tag{3.215}$$

Eq. (3.215) implies that $(1-2\nu^2)^2 = 2\nu^2(2\nu^2-1)$ which $\Leftrightarrow$

$$2\nu^2 = 1 \tag{3.216}$$

Combining Eqs. (3.76), (3.214) and (3.216), one concludes that: (i) $G(\nu, \pi)$ has an eigenvalue with $m > 1$ if and only if either $\nu = 0$ or $\nu = \pm 1/\sqrt{2}$; (ii) the eigenvalue of $G(0, \pi)$ with $m = 2$ is 1; and (iii) the eigenvalue of $G(\pm 1/\sqrt{2}, \pi)$ with $m = 3$ is -1. Also it can be show easily that (iv)

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -2 \\ 0 \\ 3 \end{pmatrix} \tag{3.217}$$

are two linearly independent eigenvectors of

$$G(0, \pi) = \begin{pmatrix} 3 & 0 & 4/3 \\ 0 & 1 & 0 \\ -6 & 0 & -3 \end{pmatrix} \tag{3.218}$$

associated with the eigenvalue 1. By using the above results (i), (ii), and (iv), Theorem 7 implies that $G(\nu, \pi)$ is nondefective if $|\nu| \neq 1/\sqrt{2}$. To complete the proof of part (ii) of Lemma 5, we need only to show that $G(\pm 1/\sqrt{2}, \pi)$ is defective.

To proceed, note that

$$G(\pm 1/\sqrt{2}, \pi) = \begin{pmatrix} 3 & \mp 2\sqrt{2} & 8/3 \\ \pm 2\sqrt{2} & -1 & \pm 4\sqrt{2}/3 \\ -6 & \pm 3\sqrt{2} & -5 \end{pmatrix} \tag{3.219}$$

Let $\vec{x} = (x_1, x_2, x_3)^t$ be an eigenvector of $G(\pm 1/\sqrt{2}, \pi)$ with the eigenvalue -1, i.e., (i) $\vec{x} \neq \vec{0}$ and (ii)

$$G(\pm 1/\sqrt{2}, \pi)\vec{x} = -\vec{x} \tag{3.220}$$

By using Eq. (3.219), Eq. (3.220) $\Leftrightarrow$

$$x_2 = 0 \quad \text{and} \quad 3x_1 + 2x_3 = 0 \tag{3.221}$$

Thus any eigenvector of $G(\pm 1/\sqrt{2}, \pi)$ associated with the eigenvalue -1 must be in the form

$$c \begin{pmatrix} 2 \\ 0 \\ -3 \end{pmatrix} \tag{3.222}$$

where c is a complex number $\neq 0$. In other words, there is one and only one linearly independent eigenvector of $G(\pm 1/\sqrt{2}, \pi)$ associated with the eigenvalue -1. Because $m = 3$ for this eigenvalue, Theorem 7 implies that $G(\pm 1/\sqrt{2}, \pi)$ is defective. Thus the proof of part (ii) is completed.

Next, according to Eq. (3.75), for the matrix

$$G(0, \theta) = \begin{pmatrix} 2 - \cos\theta & i\sin\theta & (2/3)(1 - \cos\theta) \\ i\sin\theta & -\cos\theta & (2/3)i\sin\theta \\ 3(\cos\theta - 1) & -3i\sin\theta & 2\cos\theta - 1 \end{pmatrix} \qquad -\pi < \theta \leq \pi \tag{3.223}$$

the eigenvalue with $m = 2$ is 1 while the eigenvalue with $m = 1$ is -1. On the other hand,

$$\begin{pmatrix} 1 \\ \\ \dfrac{i(1 - \cos\theta)}{\sin\theta} \\ \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ \\ \dfrac{2i(1 - \cos\theta)}{\sin\theta} \\ \\ 3 \end{pmatrix} \qquad 0 < |\theta| < \pi \tag{3.224}$$

are two linearly independent eigenvectors of $G(0,\theta)$ associated with the eigenvalue 1 if $0 < |\theta| < \pi$. By using the above results, part (iii) of Lemma 5 now follows from Theorem 7 and parts (i) and (ii) of Lemma 5. QED

Next, defectiveness of the matrix $G(\nu,\theta)$ when $|\nu| = 1/2$ and $|\theta| = 2\pi/3$ will be established in Lemma 6, i.e.,

Lemma 6. When $|\nu| = 1/2$ and $|\theta| = 2\pi/3$, $G(\nu,\theta)$ is defective, and it has an eigenvalue with multiplicity $m = 2$ and another eigenvalue with $m = 1$.

Proof. Consider the case $\nu = 1/2$ and $\theta = 2\pi/3$. Eq. (3.62) implies that

$$\sigma_0(2\pi/3) = \sigma_-(2\pi/3) = -(1 - i\sqrt{3})/2 \quad \text{and} \quad \sigma_+(2\pi/3) = (1 - i\sqrt{3})/2 \tag{3.225}$$

i.e., the eigenvalue of $G(1/2, 2\pi/3)$ with $m = 2$ is $-(1-i\sqrt{3})/2$ while the eigenvalue with $m = 1$ is $(1-i\sqrt{3})/2$. Moreover, let $\vec{x} = (x_1, x_2, x_3)^t$ be an eigenvector of

$$G(1/2, 2\pi/3) = \begin{pmatrix} (10 - i\sqrt{3})/4 & (-12 + i5\sqrt{3})/8 & (12 - i3\sqrt{3})/8 \\ (3 + i\sqrt{3})/2 & (-1 + i\sqrt{3})/4 & (3 + i\sqrt{3})/4 \\ -9/2 & (9 - i6\sqrt{3})/4 & (-11 + i2\sqrt{3})/4 \end{pmatrix} \tag{3.226}$$

with the eigenvalue $-(1 - i\sqrt{3})/2$, i.e., (i) $\vec{x} \neq \vec{0}$ and (ii)

$$G(1/2, 2\pi/3)\vec{x} = -\left[(1 - i\sqrt{3})/2\right]\vec{x} \tag{3.227}$$

By using Eq. (3.226), Eq. (3.227) $\Leftrightarrow$

$$\begin{pmatrix} 12 - i3\sqrt{3} & -12 + i5\sqrt{3} \\ 3 + i\sqrt{3} & 1 - i\sqrt{3} \\ 3 & -3 + i2\sqrt{3} \end{pmatrix} \begin{pmatrix} 2x_1 + x_3 \\ x_2 \end{pmatrix} = 0 \tag{3.228}$$

Because the 3×2 coefficient matrix in Eq. (3.228) is formed by two linearly independent 3×1 column matrices, Eq. (3.228) $\Leftrightarrow$

$$2x_1 + x_3 = 0 \quad \text{and} \quad x_2 = 0 \tag{3.229}$$

Thus any eigenvector of $G(1/2, 2\pi/3)$ associated with the eigenvalue $-(1 - i\sqrt{3})/2$ must be in the form

$$c\begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \tag{3.230}$$

where c is a complex number $\neq 0$. In other words, there is one and only one linearly independent eigenvector of $G(1/2, 2\pi/3)$ associated with the eigenvalue $-(1 - i\sqrt{3})/2$. Because $m = 2$ for this eigenvalue, Theorem 7 implies that $G(1/2, 2\pi/3)$ is defective.

For each of the matrices $G(1/2, -2\pi/3)$, $G(-1/2, 2\pi/3)$, and $G(-1/2, -2\pi/3)$, its defectiveness can be proved by using (i) defectiveness of $G(1/2, 2\pi/3)$, (ii) Eqs. (3.4) and (3.5), and (iii) Theorem 8. Also the fact that each has an eigenvalue with $m = 2$ and another eigenvalue with $m = 1$ can be proved by using Eqs. (3.62) and (3.63) directly, or by using Eq. (3.14) along with the proved similar property of $G(1/2, 2\pi/3)$. QED.

Next, an immediate result of Theorem 7 and part (i) of Corollary to Proposition 3 is Lemma 7:

Lemma 7. For a given θ with $0 < |\theta| < \pi$, $G(\nu,\theta)$ is nondefective if (a) $0 < |\nu| < 1/2$, and also if (b) $|\nu| = 1/2$ and $|\theta| \neq 2\pi/3$.

Moreover, by combining Lemmas 5 and 7, we have Lemma 8:

Lemma 8. $G(\nu,\theta)$ is nondefective if (a) $|\nu| < 1/2$ and $-\pi < \theta \le \pi$, and also if (b) $|\nu| = 1/2$, $-\pi < \theta \le \pi$, and $|\theta| \ne 2\pi/3$.

To prove Proposition 4, note that it has been shown earlier that the $a(3)$ scheme is unstable if $|\nu| > 1/2$. On the other hand, by using Proposition 1(a) and Definition 3, one has

$$|\rho\left(G(\nu,\theta)\right)| = |\sigma_1(\nu,\theta)| = |\sigma_2(\nu,\theta)| = |\sigma_3(\nu,\theta)| = 1, \quad -\pi < \theta \le \pi, \quad \text{if} \quad |\nu| \le 1/2 \tag{3.231}$$

In turn, with the aid of Eq. (3.231) and part (a) of Lemma 8, Lemma 4 implies that the $a(3)$ scheme is *neutrally* stable if $|\nu| < 1/2$. Thus, to complete the proof, one needs only to prove that the $a(3)$ scheme is linearly unstable if $|\nu| = 1/2$.

By using Proposition 1(a), part (b) of Lemma 8, and Theorem 9, one concludes that every element of $[G(\nu,\theta)]^m$ remains bounded as $m \to +\infty$ for any (ν,θ) with $|\nu| = 1/2$, $-\pi < \theta \le \pi$, and $|\theta| \ne 2\pi/3$. Thus, Definition 1 implies that the $a(3)$ scheme is unstable at $\nu = \pm 1/2$ if and only if at least one element of $[G(\pm 1/2,\theta)]^m$ becomes unbounded as $m \to +\infty$ when $\theta = 2\pi/3$ or $\theta = -2\pi/3$.

Consider any (ν,θ) with $|\nu| = 1/2$ and $|\theta| = 2\pi/3$. Then, by using Eq. (3.231) and Lemma 6, Theorem 9 implies that at least one element of $[G(\nu,\theta)]^m$ becomes unbounded as $m \to +\infty$, i.e., we have proved that the $a(3)$ scheme indeed is unstable if $|\nu| = 1/2$. Moreover, according to Lemma 6, $G(\nu,\theta)$ is defective and has an eigenvalue (denoted by $\sigma_1(\nu,\theta)$) with multiplicity = 1 and another eigenvalue (denoted by $\sigma_2(\nu,\theta)$) with multiplicity = 2. As such the Jordan canonical form theorem implies that there exists a nonsingular 3×3 matrix S such that $G(\nu,\theta) = S\Lambda_3 S^{-1}$ where Λ_3 is defined in Eq. (3.205) with

$$\lambda_1 = \sigma_1(\nu,\theta) \quad \text{and} \quad \lambda_2 = \sigma_2(\nu,\theta) \qquad\qquad (|\nu| = 1/2;\ |\theta| = 2\pi/3) \tag{3.232}$$

Because $G(\nu,\theta) = S\Lambda_3 S^{-1}$ implies that $[G(\nu,\theta)]^m = S(\Lambda_3)^m S^{-1}$, the behavior of the elements of $[G(\nu,\theta)]^m$ as $m \to +\infty$ is completely determined by that of $(\Lambda_3)^m$ as $m \to +\infty$. Moreover, because $|\lambda_1| = |\lambda_2| = 1$ (which follows from Eqs. (3.231) and (3.232)), Eq. (3.208) implies that the only element of $(\Lambda_3)^m$ that will become unbounded as $m \to +\infty$ is the element $m(\lambda_2)^{m-1}$ and that the magnitude of this element grows only linearly with m. As a result of the above considerations, one concludes that *any element of* $[G(\nu,\theta)]^m$ *at most can only grow linearly with* m *for any case with* $|\nu| = 1/2$ *and* $|\theta| = 2\pi/3$. Because of this reason and the fact that the time evolution of the round-off errors originally introduced during any marching step is also governed by the $a(3)$ scheme, the round-off errors originally introduced during a single marching step at most can only grow linearly with the marching-step number if $|\nu| = 1/2$. It is in this sense that the $a(3)$ scheme is said to be linearly unstable when $|\nu| = 1/2$. QED.

Note that the total round-off errors observed at the start of any marching step is the sum of the "offsprings" of the round-off errors originally introduced during all the previous marching steps. Because of the intrinsic random nature of round-off-error generation and the accompanied effect of (in-phase and out-of-phase) interferences, evaluating the sum referred to above is much more complex than evaluating the offspring of the round-off errors introduced during a single marching step. Nevertheless, because the growth rate of the magnitude of the term $m(\lambda_2)^{m-1}$ for the case $|\lambda_2| = 1$ is much lower than the exponential growth rate of a term associated with a case with $\rho(G(\nu,\theta)) > 1$, one still can infer that the instability of the $a(3)$ scheme when $|\nu| = 1/2$ is much milder than that for a case with $|\nu| > 1/2$. As will be shown in Sec. 4, this prediction is borne out by numerical results.

4. Numerical results

To assess the accuracy of the $a(3)$ scheme, consider the model problem with the PDE

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 \tag{4.1}$$

and the exact solution

$$u = u_e(x,t) \overset{\text{def}}{=} \sin\left[2\pi(x-t)\right] \equiv \frac{1}{2i}\left[e^{i2\pi(x-t)} - e^{-i2\pi(x-t)}\right] \tag{4.2}$$

We have

$$a = L = T = 1 \tag{4.3}$$

where L = wavelength and T = period. Moreover, $u_e(x,t)$ is a linear combination of two plane wave solutions $e^{k_+(x-t)}$ and $e^{k_-(x-t)}$ with

$$k_\pm = \pm 2\pi \tag{4.4}$$

Let (i)

$$u_{xe}(x,t) \overset{\text{def}}{=} \frac{\partial u_e(x,t)}{\partial x} \quad \text{and} \quad u_{xxe}(x,t) \overset{\text{def}}{=} \frac{\partial^2 u_e(x,t)}{\partial x^2} \tag{4.5}$$

and (ii) the spatial domain of unit length be divided into K uniform intervals. Then, with the aid of Eq. (4.3), one has

$$\Delta x = 1/K, \quad \Delta t = \nu\Delta x, \quad \text{and} \quad t = n\Delta t \tag{4.6}$$

where n = number of time steps, and t = total marching time. The computer code solving the model problem for various values of K and n using the $a(3)$ scheme is listed in Appendix C while that using the dual a scheme [71] is listed in Appendix D. Because the dual a scheme (which are defined over the set Ω) is formed by two completely decoupled a schemes (which are defined over the sets Ω_1 and Ω_2, respectively), the accuracy and stability conditions of the dual a scheme are identical to those of the a scheme.

Based on the von Neumann analysis, it was shown in Sec. 3 (Proposition 4) that the $a(3)$ scheme (i) is stable if and only if $|\nu| < 1/2$; (ii) is neutrally stable if $|\nu| < 1/2$; and (iii) is linearly unstable if $|\nu| = 1/2$. On the other hand, by using the amplification matrix given in Eq. (3.51) of [71] and a line of arguments similar to that presented in Sec. 3, one can show that the a scheme (and the dual a scheme) (i) is stable if and only if $|\nu| < 1$; (ii) is neutrally stable if $|\nu| < 1$; and (iii) is linearly instable if $|\nu| = 1$. These stability conditions have been verified numerically using the codes implementing the $a(3)$ scheme and the dual a scheme, which are listed in Appendices C and D, respectively.

In Tables 1–4, the numerical errors of several computations using the $a(3)$ scheme and the dual a scheme are presented in terms of the following error norms for the *non-normalized* independent mesh variables:

$$E(K,n,\nu) \overset{\text{def}}{=} \sqrt{\frac{1}{K}\sum_{j=0}^{K-1}\left[u_j^n - u_e(x_j,t^n)\right]^2} \tag{4.7}$$

$$E_x(K,n,\nu) \overset{\text{def}}{=} \sqrt{\frac{1}{K}\sum_{j=0}^{K-1}\left[(u_x)_j^n - u_{xe}(x_j,t^n)\right]^2} \tag{4.8}$$

and

$$E_{xx}(K,n,\nu) \overset{\text{def}}{=} \sqrt{\frac{1}{K}\sum_{j=0}^{K-1}\left[(u_{xx})_j^n - u_{xxe}(x_j,t^n)\right]^2} \tag{4.9}$$

Because, at each mesh point (j,n), the only non-normalized independent mesh variables associated with the dual a scheme are u_j^n and $(u_x)_j^n$, obviously the error norm $E_{xx}(K,n,\nu)$ is not applicable to the dual a scheme.

The numerical errors of several simulations with $\nu = 0.1$ and $t = 9.876$ are given in Table 1. For the dual a scheme, as the values of K and n become larger, the values of E and E_x are both reduced by a factor of about 4 as both K and n double their values, i.e., the scheme is 2nd order in accuracy for both u_j^n and $(u_x)_j^n$. On the other hand, for the $a(3)$ scheme, the values of E, E_x, and E_{xx} are reduced by the factors 16, 16, and 4, respectively. Thus the $a(3)$ scheme is 4th order in accuracy for both u_j^n and $(u_x)_j^n$ while only 2nd order in accuracy for $(u_{xx})_j^n$. From the results shown, one can see that the $a(3)$ scheme is much more accurate than the dual a scheme. As an example, for the case with $K = 25$ and $n = 2469$, the value of E for the dual a scheme is larger than that for the $a(3)$ scheme by a factor of 3450! Because the a scheme is only 2nd order in accuracy for u_j^n, it is estimated that the accuracy of u_j^n achieved by the $a(3)$ scheme with $K = 25$ and $n = 2469$ is identical to that achieved by the dual a scheme with $K = 25 \times \sqrt{3450} \approx 1468$ and $n = 2469 \times \sqrt{3450} \approx 145029$.

Here the reader is reminded that, for a reason given in Sec. 2.5, the conclusion reached above about the orders of accuracy of the $a(3)$ scheme does not contradict that reached in Sec. 2.5 about the orders of truncation error of the $a(3)$ scheme.

In Table 2, the cases considered have $\nu = 0.1$ and $t = 10.00 = 10T$. For these cases where t is an integer multiple of the period T, it is seen that the $a(3)$ scheme is 4th order in accuracy for u_j^n, $(u_x)_j^n$, *and* $(u_{xx})_j^n$.

In Table 3, the cases considered have $\nu = 0.5$ and $t = 49.38$, For these cases where the value of ν is right at the stability boundary of the $a(3)$ scheme, aside from round-off errors, the numerical values of $(u_x)_j^n$ generated using the $a(3)$ scheme are all identical to their exact solution values, respectively, if n are *even* integers.

In Table 4, the cases considered have $\nu = 0.5$ and $t = 50.00 = 50T$, i.e., the value of ν is right at the stability boundary of the $a(3)$ scheme and t is an integer multiple of T. It is seen that the numerical values of u_j^n, $(u_x)_j^n$, and $(u_{xx})_j^n$ generated using the $a(3)$ scheme, aside from round-off errors, are all identical to their exact solution values, respectively. Note that: (i) n and Δt are chosen according to Eq. (3.67) and n is even for each of these cases, and (ii) the exact solution are a linear combination of two plane wave solutions with $|\theta| = |k_\pm \Delta x| = |\pm 2\pi/K| < 2\pi/3$, $K = 25, 50, 100, 200$ (see Eq. (4.4)), i.e., θ observes the condition Eq. (3.64). Thus the numerical results of the $a(3)$ scheme shown in Table 4 confirm an accuracy prediction made in Sec. 3.4.

Moreover, the round-off errors associated with the $a(3)$ scheme shown in Table 4 also confirm a prediction made at the end of Sec. 3.9, i.e., the $a(3)$ scheme is only mildly unstable when $|\nu| = 1/2$.

According to the von Neumann analysis, (i) the dual a scheme has no dissipative errors (i.e., the magnitudes of all its amplification factors = 1 for all θ in the domain $-\pi < \theta \le \pi$) if $|\nu| \le 1$; and (ii) the $a(3)$ scheme has no dissipative errors if $|\nu| \le 1/2$. Thus, for a simulation with periodic boundary condition, aside from round-off errors, the numerical errors generated using the dual a scheme are contributed solely by the phase (dispersive) errors if $|\nu| \le 1$. On the other hand, those generated by using the $a(3)$ scheme are contributed solely by the phase errors if $|\nu| \le 1/2$. Thus the relative accuracy of the a scheme and the $a(3)$ scheme can also be evaluated by comparing their phase errors.

For both the $a(3)$ scheme and the dual a scheme, the phase error of the principal amplification factor associated with any (ν,θ) can be measured by

$$E_p(\nu,\theta) \overset{\text{def}}{=} 1 - \frac{\phi(\nu,\theta)}{\phi_e(\nu,\theta)} \qquad (\phi_e(\nu,\theta) \neq 0) \tag{4.10}$$

Here: (i) $\phi_e(\nu,\theta)$ is the phase angle of the exact amplification factor, i.e.,

$$\phi_e(\nu,\theta) = -\nu\theta \tag{4.11}$$

and (ii) $\phi(\nu,\theta)$ is the phase angle of the principal amplification factor. Note that, by using Eq. (3.14) and similar relations for the dual a scheme, one concludes that

$$\phi(-\nu,\theta) = \phi(\nu,-\theta) = -\phi(\nu,\theta) = -\phi(-\nu,-\theta) \tag{4.12}$$

An immediate result of Eqs. (4.10)–(4.12) is

$$E_p(\nu,\theta) = E_p(-\nu,\theta) = E_p(\nu,-\theta) = E_p(-\nu,-\theta) \tag{4.13}$$

Thus only the values of $E_p(\nu,\theta)$ with nonnegative ν and nonnegative θ need to be evaluated numerically.

In the code listed in Appendix E, for the dual a scheme, $\phi(\nu,\theta)$ is evaluated using the exact formula:

$$\phi(\nu,\theta) = \tan^{-1}\left(\frac{-\nu\sin\theta}{\sqrt{1-\nu^2\sin^2\theta}}\right), \qquad \nu^2 < 1 \; -\pi < \theta \leq \pi \qquad \text{(the dual } a \text{ scheme)} \tag{4.14}$$

(see Eq. (3.31) in [71]). On the other hand, for the $a(3)$ scheme, $\phi(\nu,\theta)$ is evaluated in the same code using the Newton's iterative procedure Eq. (3.113) and the assumption

$$\phi^0 = \phi_e(\nu,\theta) \tag{4.15}$$

After k iterations, ϕ^k is taken as the converged value of $\phi(\nu,\theta)$ if

$$|(\phi^k/\phi^{k-1}) - 1| < \epsilon \tag{4.16}$$

where $\epsilon > 0$ is a very small preset value. Note that the iterative procedure generally converges rapidly. Specifically, Eq. (4.16) with $\epsilon = 10^{-14}$ is satisfied after at most 5 iterations for all (ν,θ), $|\nu| < 1/2$ and $-\pi < \theta \leq \pi$. Moreover, as expected, convergence is reached after one iteration for all θ if $|\nu| = 1/2$.

The numerical values of $E_p(\nu,\theta)$ for the cases $\nu = 0.001, 0.01, 0.1, 0.5$ are plotted against θ (denoted by Z) in Fig. 3 for both the dual a scheme and the $a(3)$ scheme. The values of $E_p(\nu,\theta)$ for the dual a scheme are calibrated using the left-ordinate scale while those for the $a(3)$ scheme are calibrated using the right-ordinate scale. It can be seen that the values of $E_p(\nu,\theta)$ for the $a(3)$ scheme are uniformly much smaller than those for the dual a scheme. In fact, the numerical results indicate that, for the $a(3)$ scheme, (i) $\phi(\nu,\theta) = O\left[(\theta)^4\right]$ if $|\nu| < 1/2$; and (ii) aside from round-off errors, $\phi(\nu,\theta) = 0$ for all θ, if $|\nu| = 1/2$ (which is expected from Eqs. (3.62) and (3.63)—see a discussion given following Eq. (3.64)). On the other hand, for the dual a scheme, $\phi(\nu,\theta) = O\left[(\theta)^2\right]$ if $|\nu| < 1$.

5. Conclusions and discussions

A thorough and rigorous discussion of the $a(3)$ scheme, a new high order neutrally stable CESE solver of Eq. (1.1) has been presented. As in the case of other similar CESE neutrally stable solvers [1,5,11,72], the $a(3)$ scheme enforces conservation laws locally and globally, and it has the basic, forward marching, and backward marching forms. These forms are equivalent and satisfy the PT invariant property defined in Sec. 2.

Based on the concept of PT invariance, the algebraic relations Eqs. (2.114)–(2.118) are derived in Sec. 2. As it turns out, in the von Neumann analysis presented in Sec. 3, these relations can be used to prove that the $a(3)$ scheme is neutrally stable when it is stable. Another set of algebraic relations Eq. (2.119) which results from other invariant property are also discussed in Sec. 2.

In addition to establishing the neutral stability of $a(3)$ scheme, in Sec. 3, it is also proved rigorously that all three amplification factors of the $a(3)$ scheme are of unit magnitude for all phase angles θ if and only if $|\nu| \leq 1/2$ (Proposition 1). Moreover, it is proved that the $a(3)$ scheme is (i) stable if and only if $|\nu| < 1/2$; and (ii) linear unstable if $|\nu| = 1/2$ (Proposition 4). These theoretical results have been confirmed by numerical experiments.

It is shown in Sec. 4 that the $a(3)$ scheme generally is (i) 4th-order accurate for the mesh variables u_j^n and $(u_x)_j^n$; and (ii) 2nd-order accurate for $(u_{xx})_j^n$. However, in some exceptional cases, the scheme can achieve perfect accuracy aside from round-off errors. Moreover, it is shown that the phase errors of the principal amplification factor of the $a(3)$ scheme are $O(\theta^4)$ if $|\nu| < 1/2$, a sharp reduction from those of the dual a scheme which are $O(\theta^2)$ if $|\nu| < 1$.

Appendix A. Proof for Theorems 1 and 2

First we prove Theorem 1. According to Eq. (8.20) on p.265 of [75], the fact that λ_ℓ, $\ell = 1, 2, \ldots, N$ are the eigenvalues of the $N \times N$ matrix $A \Leftrightarrow$

$$\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)\cdots(\lambda_N - \lambda) \tag{A.1}$$

where λ is any complex variable and I is the $N \times N$ identity matrix. Let $\lambda = 0$. Eq. (A.1) implies that

$$\det(A) = \lambda_1 \lambda_2 \cdots \lambda_N \tag{A.2}$$

By definition, A is nonsingular $\Leftrightarrow \det(A) \neq 0$. Thus part (i) follows from Eq. (A.2).

According to Eq. (A.1), to prove part (ii) we need only to show that

$$\det\left(A^{-1} - \lambda I\right) = \left(\frac{1}{\lambda_1} - \lambda\right)\left(\frac{1}{\lambda_2} - \lambda\right)\cdots\left(\frac{1}{\lambda_N} - \lambda\right) \tag{A.3}$$

for any complex variable λ. Because $\det(BC) = \det(B)\det(C)$ for any two $N \times N$ matrices B and C, we have $\det(A)\det\left(A^{-1}\right) = \det\left(AA^{-1}\right) = \det(I) = 1$, i.e., $\det(A^{-1}) = 1/\det(A)$. Thus Eq. (A.2) implies that

$$\det\left(A^{-1}\right) = \frac{1}{\lambda_1 \lambda_2 \cdots \lambda_N} \tag{A.4}$$

By comparing Eqs. (A.3) and (A.4), one concludes that Eq. (A.3) is valid if $\lambda = 0$.

Let $\lambda \neq 0$. Then

$$A^{-1} - \lambda I = -\lambda I A^{-1}\left(A - \frac{1}{\lambda} I\right) \tag{A.5}$$

Thus

$$\det\left(A^{-1} - \lambda I\right) = \det(-\lambda I)\det\left(A^{-1}\right)\det\left(A - \frac{1}{\lambda} I\right) \tag{A.6}$$

With the aid of (i) Eqs. (A.1) and (A.4) and (ii) the fact that $\det(-\lambda I) = (-\lambda)^N$, Eq. (A.6) implies that

$$\begin{gathered}\det\left(A^{-1} - \lambda I\right) = \frac{(-\lambda)^N}{\lambda_1 \lambda_2 \cdots \lambda_N}\left(\lambda_1 - \frac{1}{\lambda}\right)\left(\lambda_2 - \frac{1}{\lambda}\right)\cdots\left(\lambda_N - \frac{1}{\lambda}\right) = \\ \left(\frac{-\lambda}{\lambda_1}\right)\left(\lambda_1 - \frac{1}{\lambda}\right)\left(\frac{-\lambda}{\lambda_2}\right)\left(\lambda_2 - \frac{1}{\lambda}\right)\cdots\left(\frac{-\lambda}{\lambda_N}\right)\left(\lambda_N - \frac{1}{\lambda}\right) = \left(\frac{1}{\lambda_1} - \lambda\right)\left(\frac{1}{\lambda_2} - \lambda\right)\cdots\left(\frac{1}{\lambda_N} - \lambda\right)\end{gathered} \tag{A.7}$$

i.e., Eq. (A.3) is also valid if $\lambda \neq 0$. Thus part (ii) of Theorem 1 has been proved. QED.

According to Eq. (A.1), to prove Theorem 2, we need only to show that

$$\det(\overline{A} - \lambda I) = (\overline{\lambda_1} - \lambda)(\overline{\lambda_2} - \lambda)\cdots(\overline{\lambda_N} - \lambda) \tag{A.8}$$

Because $\det(\overline{M}) = \overline{\det(M)}$, by using Eq. (A.1), we have

$$\begin{aligned}\det(\overline{A} - \lambda I) &= \det\overline{\left(A - \overline{\lambda} I\right)} = \overline{\det(A - \overline{\lambda} I)} \\ &= \overline{(\lambda_1 - \overline{\lambda})(\lambda_2 - \overline{\lambda})\cdots(\lambda_N - \overline{\lambda})} = (\overline{\lambda_1} - \lambda)(\overline{\lambda_2} - \lambda)\cdots(\overline{\lambda_N} - \lambda)\end{aligned} \tag{A.9}$$

i.e., Eq. (A.8) has been proved. QED.

Appendix B. Proof for Eqs. (3.138), (3.139), (3.155), (3.169), and (3.170)

Proof for Eq. (3.138). Assuming Eq. (3.134) and using elementary trigonometry, we have

$$\begin{aligned}
&\frac{\cos\phi_1+\cos\phi_2-\cos(\phi_1+\phi_2)-1}{\sin\phi_1+\sin\phi_2+\sin(\phi_1+\phi_2)} \equiv \frac{2\cos(\frac{\phi_1+\phi_2}{2})\cos(\frac{\phi_1-\phi_2}{2})-2\cos^2(\frac{\phi_1+\phi_2}{2})}{2\sin(\frac{\phi_1+\phi_2}{2})\cos(\frac{\phi_1-\phi_2}{2})+2\sin(\frac{\phi_1+\phi_2}{2})\cos(\frac{\phi_1+\phi_2}{2})}\\
&\equiv \frac{\cos(\frac{\phi_1+\phi_2}{2})\Big[\cos(\frac{\phi_1-\phi_2}{2})-\cos(\frac{\phi_1+\phi_2}{2})\Big]}{\sin(\frac{\phi_1+\phi_2}{2})\Big[\cos(\frac{\phi_1-\phi_2}{2})+\cos(\frac{\phi_1+\phi_2}{2})\Big]} \equiv \cot\left(\frac{\phi_1+\phi_2}{2}\right)\frac{2\sin(\phi_1/2)\sin(\phi_2/2)}{2\cos(\phi_1/2)\cos(\phi_2/2)}\\
&\equiv \cot\left(\frac{\phi_1+\phi_2}{2}\right)\tan(\phi_1/2)\tan(\phi_2/2) \equiv \tan(\phi_1/2)\tan(\phi_2/2)\tan\left(\pm\frac{\pi}{2}-\frac{\phi_1+\phi_2}{2}\right)
\end{aligned} \tag{B.1}$$

Eq. (3.138) follows from Eq. (B.1) and Eq. (3.130). QED.

Proof for Eq. (3.139). Assuming Eq. (3.134) and using elementary trigonometry, we have

$$\begin{aligned}
&\frac{\cos\phi_1+\cos\phi_2-\cos(\phi_1+\phi_2)+3}{\sin\phi_1+\sin\phi_2+\sin(\phi_1+\phi_2)} \equiv \frac{2\cos(\frac{\phi_1+\phi_2}{2})\cos(\frac{\phi_1-\phi_2}{2})+2\cos^2(\frac{\phi_1+\phi_2}{2})+4\sin^2(\frac{\phi_1+\phi_2}{2})}{2\sin(\frac{\phi_1+\phi_2}{2})\cos(\frac{\phi_1-\phi_2}{2})+2\sin(\frac{\phi_1+\phi_2}{2})\cos(\frac{\phi_1+\phi_2}{2})}\\
&\equiv \frac{\cos(\frac{\phi_1+\phi_2}{2})\Big[\cos(\frac{\phi_1-\phi_2}{2})+\cos(\frac{\phi_1+\phi_2}{2})\Big]+2\sin^2(\frac{\phi_1+\phi_2}{2})}{\sin(\frac{\phi_1+\phi_2}{2})\Big[\cos(\frac{\phi_1-\phi_2}{2})+\cos(\frac{\phi_1+\phi_2}{2})\Big]} \equiv \cot\left(\frac{\phi_1+\phi_2}{2}\right)+\frac{\sin(\frac{\phi_1+\phi_2}{2})}{\cos(\phi_1/2)\cos(\phi_2/2)}\\
&\equiv \tan\left(\pm\frac{\pi}{2}-\frac{\phi_1+\phi_2}{2}\right)+\frac{\sin(\phi_1/2)\cos(\phi_2/2)+\cos(\phi_1/2)\sin(\phi_2/2)}{\cos(\phi_1/2)\cos(\phi_2/2)}\\
&\equiv \tan\left(\pm\frac{\pi}{2}-\frac{\phi_1+\phi_2}{2}\right)+\tan(\phi_1/2)+\tan(\phi_2/2)
\end{aligned} \tag{B.2}$$

Eq. (3.139) follows from Eq. (B.2) and Eq. (3.130). QED

Proof for Eq. (3.155). Using elementary trigonometry, we have

$$\begin{aligned}
&\sin\varphi_1+\sin\varphi_2+\sin(\varphi_1+\varphi_2) \equiv 2\sin(\frac{\varphi_1+\varphi_2}{2})\cos(\frac{\varphi_1-\varphi_2}{2})+2\sin(\frac{\varphi_1+\varphi_2}{2})\cos(\frac{\varphi_1+\varphi_2}{2})\\
&\equiv 2\sin(\frac{\varphi_1+\varphi_2}{2})\Big[\cos(\frac{\varphi_1-\varphi_2}{2})+\cos(\frac{\varphi_1+\varphi_2}{2})\Big] \equiv 4\sin(\frac{\varphi_1+\varphi_2}{2})\cos(\varphi_1/2)\cos(\varphi_2/2)
\end{aligned} \tag{B.3}$$

Next, by using Eq. (3.153), we have

$$\sin(\frac{\varphi_1+\varphi_2}{2}) = \sin\left(\pm\frac{\pi}{2}-\frac{\varphi_3}{2}\right) \equiv \pm\cos(\phi_3/2) \tag{B.4}$$

Eq. (3.155) is a direct result of Eqs. (B.3) and (B.4). QED.

Proof for Eq. (3.169). Using elementary trigonometry, we have (i)

$$\tan(\theta/4)-\cot(\theta/4) = \tan(\theta/4)-\frac{1}{\tan(\theta/4)} = -\frac{2[1-\tan^2(\theta/4)]}{2\tan(\theta/4)} = -\frac{2}{\tan(\theta/2)}, \qquad 0<|\theta|<\pi \tag{B.5}$$

and (ii)

$$\begin{aligned}&-\tan(\theta/2)-\frac{2}{\tan(\theta/2)}=\tan(\theta/2)-\frac{2[1+\tan^2(\theta/2)]}{\tan(\theta/2)}=\tan(\theta/2)-\frac{2\sec^2(\theta/2)}{\tan(\theta/2)}\\&=\tan(\theta/2)-\frac{2}{\cos(\theta/2)\sin(\theta/2)}=\tan(\theta/2)-\frac{4}{\sin\theta}\end{aligned}\qquad 0<|\theta|<\pi \qquad (B.6)$$

Eq. (3.169) is a result of Eqs. (B.5) and (B.6). QED.

Proof for Eq. (3.170). Using elementary trigonometry, we have

$$\tan(\theta/2)[\cot(\theta/4)-\tan(\theta/4)]=\tan(\theta/2)\frac{1-\tan^2(\theta/4)}{\tan(\theta/4)}=\tan(\theta/2)\frac{2}{\tan(\theta/2)}=2,\qquad 0<|\theta|<\pi \qquad (B.7)$$

i.e., Eq. (3.170) has been proved. QED.

```
c                    Appendix C. Code "a3.for"
c
      implicit real*8(a-h,o-z)
      parameter (nxd=2000)
      dimension u(nxd), un(nxd), ux(nxd), uxn(nxd), uxx(nxd),
     *   uxxn(nxd)
c
c     Code "a3.for". This code implements a neutrally stable solver
c     for a pure advection equation with a sine-wave initial data and
c     periodic boundary conditions. The sine wave is of unit wavelength.
c
c     Theoretically, the solver is designed to have at least
c     third-order accuracy. However, it has been shown numerically
c     that the scheme is of 4th-order accuracy.
c
c     Let a, dx, and dt denote the advection speed, the spatial mesh
c     interval and the time step size, respectively. Let (i) the
c     Courant number cn = a*dt/dx; and (ii) theta denote the phase
c     angle of a Fourier mode. Then, according to Proposition 1(a),
c     for any pair of cn and theta with |cn| .le. 1/2 and -pi .lt.
c     theta .le. pi, the three amplification factors of the a3 scheme
c     are all of unit magnitude.
c
c     It can be shown analytically the scheme is (i) stable if
c     |cn| .lt. 1/2; (ii) linear unstable if |cn| = 1/2 ; and
c     (iii) unstable if |cn| .gt. 1/2.
c
c     There are three independent mesh variables (the analogues of the
c     dependent variable and its first and second spatial derivatives)
c     and three eauations per mesh points. The three equations are
c     obtained by imposing (i) two conservation conditions over CE- and
c     CE+ and (ii) a STI invariant condition that insures the solver has
c     third-order truncation errors.
c
c     ux, uxx, uxn, and uxxn represent the current and updated numerical
c     analogues of normalized spatial derivatives during time marching.
c     However, the output contains the non-normalized values.
c
c     it = no. of time steps.
c     k = mesh intervals per unit length.
c     cn = Courant number.
c     a = advection speed.
c     iop = output selector. Only the global L2 error norms eru2, erux2,
c         and eruxx2 (which correspond to u, ux, and uxx, respectively)
c         along with the problem defining parameters will be included in
c         the output if iop = 0. Otherwise, all local solution and error
c         values will also be included.
c
c     The computational domain is 0 .le. x .le. 1. In the output, (i)
c     ue, uxe and uxxe are local exact solution values, (ii) u, ux,
c     and uxx are local numerical solution values, (iii) eru = u - ue,
c     erux = ux - uxe, and eruxx = uxx - uxxe, and (iv) eru2, erux2,
c     and eruxx2 (which are given at the end) are the global L2 error
c     norms for u, ux, and uxx, respectively.
c
      it = 6000
      k = 120
      cn = 0.5d0
      a = 1.d0
      iop = 0
c
      dx = 1.d0/dfloat(k)
      dt = cn*dx/a
      dx1 = dx/2.d0
      dx2 = dx**2/4.d0
      t = dt*dfloat(it)
      k1 = k + 1
      k2 = k + 2
      pi = 3.1415926535897932d0
      tp = 2.d0*pi
      tps = tp**2
      tpat = tp*a*t
      pdx = pi*dx
      pdxs = pdx**2
      cns = cn**2
      cnp = 1.d0 + cn
      cnm = 1.d0 - cn
      a1 =  (1.d0 + 2.d0*cns)/3.d0
      a2 = 2.d0*(cnp + cns)/3.d0
      a3 = 2.d0*(cnm + cns)/3.d0
      bp = cnp/2.d0
      bm = cnm/2.d0
      cz = 2.d0*cn
      cm = 0.5d0 - cn
      cp = 0.5d0 + cn
c
      open (unit=8,file='a3.opt')
      write (8,10)
      write (8,15)
      write (8,20) it,k,dt
      write (8,25) a,cn
      write (8,30) t
      do 100 j = 1,k2
      tpx = tp*dfloat(j-1)*dx
      u(j) = dsin(tpx)
      ux(j) = pdx*dcos(tpx)
      uxx(j) = -pdxs*dsin(tpx)
100   continue
      do 400 i = 1,it
      do 200 j = 2,k1
      s = u(j) - cn*ux(j) + a1*uxx(j)
      sp = u(j+1) - cnp*ux(j+1) + a2*uxx(j+1)
      sm = u(j-1) + cnm*ux(j-1) + a3*uxx(j-1)
      un(j) = 2.d0*s - bp*sp - bm*sm
      uxn(j) = cz*s + cm*sp - cp*sm
      uxxn(j) = 1.5d0*(sp + sm) - 3.d0*s
200   continue
      do 300 j = 2,k1
      u(j) = un(j)
      ux(j) = uxn(j)
      uxx(j) = uxxn(j)
300   continue
      u(k2) = u(2)
      ux(k2) = ux(2)
      uxx(k2) = uxx(2)
      u(1) = u(k1)
      ux(1) = ux(k1)
      uxx(1) = uxx(k1)
400   continue
      eru2 = 0.d0
      erux2 = 0.d0
      eruxx2 = 0.d0
      do 500 j = 1,k
      x = dfloat(j-1)*dx
      tpx = tp*x
      ue = dsin(tpx-tpat)
      uxe = tp*dcos(tpx-tpat)
      uxxe = -tps*ue
      ux(j) = ux(j)/dx1
      uxx(j) = uxx(j)/dx2
      eru = u(j) - ue
      erux = ux(j) - uxe
      eruxx = uxx(j) - uxxe
      eru2 = eru2 + eru**2
      erux2 = erux2 + erux**2
      eruxx2 = eruxx2 + eruxx**2
      if (iop.eq.0) goto 500
      write (8,35) x
      write (8,40) ue,u(j),eru
      write (8,45) uxe,ux(j),erux
      write (8,50) uxxe,uxx(j),eruxx
500   continue
      eru2 = dsqrt(eru2/dfloat(k))
      erux2 = dsqrt(erux2/dfloat(k))
      eruxx2 = dsqrt(eruxx2/dfloat(k))
      write (8,15)
      write (8,55) eru2,erux2,eruxx2
      close (unit=8)
10    format (' OUTPUT FOR CODE A3')
15    format (' ******************************************************')
20    format (' it =',i8,' k =',i8,' dt =',g14.7)
25    format (' a =',g14.7,' CFL =',g14.7)
```

```
30    format (' t =',g14.7)
35    format (' x =',g14.7,'******************************************')
40    format (' ue =',g14.7,xx,' u =',g14.7,xx,' eru =',g14.7)
45    format (' uxe =',g14.7,x,' ux =',g14.7,x,' erux =',g14.7)
50    format (' uxxe =',g14.7,' uxx =',g14.7,' eruxx =',g14.7)
55    format (' eru2 =',g14.7,' erux2 =',g14.7,' eruxx2 =',g14.7)
      stop
      end
```

```
c                     Appendix D. Code "a2.for"
c
      implicit real*8(a-h,o-z)
      parameter (nxd=2000)
      dimension u(nxd), un(nxd), ux(nxd), uxn(nxd)
c
c     Code ``a2.for''. This code implememts the dual ``a'' scheme
c     with a sine-wave initial data and periodic boundary conditions.
c     The sine-wave is of unit wavelength. It can be shown that
c     both exact and numerical solutions must be spatially periodic
c     with unit wavelength too.
c
c     ux and uxn represent the numerical analogues of normalized
c     spatial derivatives during time marching. However, the output
c     contains the non-normalized values.
c
c     ****************************************************************
c     input:
c     it = no. of time-marching steps.
c     k = no. of spatial intervals per unit length.
c     cn = Courant number.
c     a = the advection speed.
c     iop = output selector. Only the global L2 error norms eru2 and
c         and erux2 (which correspond to u, and ux, respectively) along
c         with the problem defining parameters will be included in
c         the output if iop = 0. Otherwise, all local exact solution
c         and numerical solution values (ue, uxe, u, and ux) along with
c         the error values (eru and erux) will also be included.
c     ****************************************************************
c
c     The computational domain is 0 .le. x .le. 1. In the output, (i)
c     ue and uxe are local exact solution values, (ii) u and ux are
c     local numerical solution values, (iii) eru = u - ue and
c     erux = ux - uxe, and (iv) eru2 and erux2 (which are given at the
c     end) are the global L2 error norms for u and ux, respectively.
c
      it = 20000
      k = 200
      cn = 0.5d0
      a = 1.d0
      iop = 0
c
      pi = 3.1415926535897932d0
      dx = 1.d0/dfloat(k)
      dt = cn*dx/a
      hdx = dx/2.d0
      t = dt*dfloat(it)
      k1 = k + 1
      k2 = k + 2
      tp = pi*2.d0
      tpat = tp*a*t
      pdx = pi*dx
      cns = (1.d0 - cn**2)/2.d0
      cnp = (1.d0 + cn)/2.d0
      cnm = (1.d0 - cn)/2.d0
c
      open (unit=8,file='a2.opt')
      write (8,10)
      write (8,15)
      write (8,20) it,k,dt
      write (8,25) a,cn
      write (8,30) t
      do 100 j = 1,k2
      tpx = tp*dfloat(j-1)*dx
      u(j) = dsin(tpx)
      ux(j) = pdx*dcos(tpx)
100   continue
      do 400 i = 1,it
      do 200 j = 2,k1
      un(j) = cnm*u(j+1) + cnp*u(j-1) + cns*(ux(j-1) - ux(j+1))
      uxn(j) = (u(j+1) - u(j-1))/2.d0 - cnp*ux(j+1) - cnm*ux(j-1)
200   continue
      do 300 j = 2,k1
      u(j) = un(j)
      ux(j) = uxn(j)
300   continue
```

```
      u(k2) = u(2)
      ux(k2) = ux(2)
      u(1) = u(k1)
      ux(1) = ux(k1)
400   continue
      eru2 = 0.d0
      erux2 = 0.d0
      do 500 j = 1,k1
      x = dfloat(j-1)*dx
      tpx = tp*x
      ue = dsin(tpx-tpat)
      uxe = tp*dcos(tpx-tpat)
      ux(j) = ux(j)/hdx
      eru = u(j) - ue
      erux = ux(j) - uxe
      eru2 = eru2 + eru**2
      erux2 = erux2 + erux**2
      if (iop.eq.0) goto 500
      write (8,15)
      write (8,35) x
      write (8,40) ue,u(j),eru
      write (8,45) uxe,ux(j),erux
500   continue
      eru2 = dsqrt(eru2/dfloat(k))
      erux2 = dsqrt(erux2/dfloat(k))
      write (8,15)
      write (8,50) eru2,erux2
      close (unit=8)
10    format (' OUTPUT FOR CODE A2')
15    format (' ******************************************************')
20    format (' it =',i8,' k =',i8,' dt =',g14.7)
25    format (' a =',g14.7,' CFL =',g14.7)
30    format (' t =',g14.7)
35    format (' x =',g14.7)
40    format (' ue =',g14.7,xx,' u =',g14.7,xx,' eru =',g14.7)
45    format (' uxe =',g14.7,x,' ux =',g14.7,x,' erux =',g14.7)
50    format (' eru2 =',g14.7,' erux2 =',g14.7)
      stop
      end
```

```
c                     Appendix E. Code "fa3.for"
c
      implicit real*8(a-h,o-z)
c
c     Code ``fa3.for''
c
c     Let a, dx, and dt denote the advection speed, the spatial mesh
c     interval and the time step size, respectively. Let (i) the
c     Courant number cn = a*dt/dx; and (ii) theta denote the phase
c     angle of a Fourier mode. Then, according to Proposition 1(a),
c     for any pair of cn and theta with |cn| .le. 1/2 and -pi .lt.
c     theta .le. pi, the three amplification factors of the a3 scheme
c     are all of unit magnitude.
c
c     Assuming that |cn| .le. 0.5, this code can be used to evaluate
c     the dispersive errors of the principal amplification factor
c     (per dt) of the a3 scheme and compare them with the
c     corresponding errors of the dual a scheme (see AIAA 2006-4779).
c
c     Let lambda = the wavelength of a Fourier mode. Then (i) theta
c     = 2*pi*dx/lambda; (ii) the phase angle of the analytical
c     amplification factor is -cn*theta. Let phi denote the phase
c     angle of the principal amplification factor of the dual a scheme
c     or the a3 scheme. Then phi = f(cn,theta) with
c     f(-cn,theta) = f(cn,-theta) = -f(cn,theta) (see Eq.(3.14) in
c     AIAA 2007-4321).
c     Thus, without any loss of generality, one may assume that
c     0 .le. cn .le. 0.5  and 0 .le. theta .le. pi in numerical
c     computations.
c
c     nz = number of intervals over the domain 0 .le. theta .le. pi.
c     ep = the error bound below which the Newton's iteration is
c     terminated.
c
c     In the output, (i) cn is the courant number (ii) z is the value
c     of theta, (iii) era is the value of [1-phi/(-cn*theta)] for
c     the dual a scheme, (iv) era3 is that for the a3 scheme, and
c     (v) k is number of Newton's interations required for convergence.
c
c     The domain of cn is 0 .le. cn .le. 0.5.
c
      cn = 0.001d0
      nz = 100
      ep = 1.d-14
      pi = 3.1415926535897932d0
      dz = pi/dfloat(nz)
      z = 0.d0
c
      open (unit=8,file='fa3.opt')
      write (8,10)
      write (8,20) cn,nz,ep
      do 300 i = 1,nz
      z = z + dz
      a = 2.d0*cn**2*(1.d0 - dcos(z))
      b = cn*dsin(z)
      p0 = -cn*z
      era = 1.d0 - datan(-b/dsqrt(1.d0 - b**2))/p0
      p = p0
      k = 0
100   pn = p - ((dsin(p))**2 - a*(1.d0 + dcos(p)) - b*dsin(p))/
     *      (dsin(2.d0*p) + a*dsin(p) - b*dcos(p))
      k = k + 1
      if (dabs(pn/p - 1.d0).lt.ep) goto 200
      p = pn
      goto 100
200   era3 = 1.d0 - pn/p0
      write (8,30) z,era,era3,k
300   continue
      close (unit=8)
10    format (' ***********',' Output for Code fa3.for',' ***********')
20    format (' cn =',g14.7,' nz =',i6,' ep =',g14.7)
30    format (' z =',g14.7,' era =',g14.7,' era3 =',g14.7,' k=',i6)
      stop
      end
```

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. Vogt 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, X.Y. Wang, and C.Y. Chow, The method of Space-Time Conservation Element and Solution Element–Application to One-Dimensional and Two-Dimensional Time-Marching Flow Problems, AIAA Paper 95-1754-CP, appears in *A Collection of Technical Papers, Part 2, pp. 1258–1291, 12th AIAA CFD Conference, June 19-22, 1995, San Diego, California*, Also published as NASA TM 106915 (1995).
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.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.
7. T. Molls and F. Molls, Space-Time Conservation Method Applied to Saint Venant Equations, *J. of Hydraulic Engr.*, **124(5)**, 501 (1998).
8. C. Zoppou and S. Roberts, Space-Time Conservation Method Applied to Saint Venant Equations: A Discussion, *J. of Hydraulic Engr.*, **125(8)**, 891 (1999).
9. 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).
10. 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).
11. 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).
12. J. Qin, S.T. Yu, Z.C. Zhang, and M.C. Lai, Direct Calculations of Cavitating Flows by the Space-Time CE/SE Method, *J. Fuels & Lubricants, SAE Transc.*, **108(4)**, 1720 (2000).
13. 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).
14. 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).
15. 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).
16. 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).
17. 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).
18. S. Cioc and T.G. Keith, Application of the CE/SE Method to One-Dimensional Flow in Fluid Film Bearings, *STLE Tribology Transactions*, **45**, 167 (2002).
19. S. Cioc and T.G. Keith, Application of the CE/SE Method to Two-Dimensional Flow in Fluid Film Bearings, *International J. of Numer. Methods for Heat & Fluid Flow*, **13(2)**, 216 (2003).

20. S. Cioc, F. Dimofte, T.G. Keith, and D.P. Fleming, Computation of Pressurized Gas Bearings Using the CE/SE Method, *STLE Tribology Transactions*, **46(1)**, 128 (2003).
21. S. Cioc, F. Dimofte, and T.G. Keith, Application of the CE/SE Method to Wave Journal Bearings, *STLE Tribology Transactions*, **46(2)**, 179 (2003).
22. 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).
23. 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).
24. 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).
25. 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 (CE/SE) Method, *Computers and Chemical Engineering*, **28**, 1309 (2004)
26. 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).
27. 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).
28. 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).
29. 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).
30. 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).
31. S. Jerez, J.V. Romero, and M.D. Rosello, A Semi-Implicit Space-Time CE-SE Method to Improve Mass Conservation through Tapered Ducts in Internal Combustion Engines, *Math. and Computer Modeling*, **40**, 941 (2004).
32. 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).
33. 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).
34. T.I. Tseng and R.J. Yang, Simulation of the Mach Reflection in Supersonic Flows by the CE/SE Method, *Shock Waves*, **14(4)**, 307 (2005).
35. B. Wang, H. He, and S.-T.J. Yu, Direct Calculation of Wave Implosion for Detonation Initiation, *AIAA J.*, **43(10)**, 2157 (2005).
36. T.I. Tseng and R.J. Yang, Numerical Simulation of Vorticity Production in Shock Diffraction, *AIAA J.*, **44(5)**, 1040 (2006).
37. 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).
38. S. Qamar and G. Warnecke, A Space-Time Conservation Method for Hyperbolic Systems with Stiff and Non Stiff Source Terms, *Commun. Comput. Phys.*, **1(3)**, 449 (2006).
39. X.Y. Wang, C.Y. Chow, and S.C.Chang, *Numerical Simulation of Flows Caused by Shock-Body Interaction*, AIAA Paper 96-2004 (1996).
40. C.Y. Loh, L.S. Hultgren and S.C. Chang, *Vortex Dynamics Simulation in Aeroacoustics by the Space-Time Conservation Element and Solution Element Method*, AIAA Paper 99-0359 (1999)
41. X.Y. Wang, S.C. Chang and P.C.E. Jorgenson, *Accuracy Study of the Space-Time CE/SE Method for Computational Aeroacoustics Problems Involving Shock Waves*, AIAA Paper 2000-0474 (2000).

42. C.Y. Loh, L.S. Hultgren, S.C. Chang and P.C.E. Jorgenson, *Noise Computation of a Supersonic Shock-Containing Axisymmetric Jet by the CE/SE Method*, AIAA Paper 2000-0475 (2000).
43. C.Y. Loh, X.Y. Wang, S.C. Chang, and P.C.E. Jorgenson, Computation of Feedback Aeroacoustic System by the CE/SE Method, in *Proceedings of the First International Conference on Computational Fluid Dynamics, Kyoto, Japan, 10-14 July, 2000*, edited by N. Satofuka, (Springer-Verlag Berlin Heidelberg 2001), p. 555.
44. C.Y. Loh, L.S. Hultgren and P.C.E. Jorgenson, *Near Field Screech Noise Computation for An Underexpanded Supersonic Jet by the CE/SE Method*, AIAA Paper 2001-2252 (2001).
45. X.Y. Wang, S.C. Chang, and P.C.E. Jorgenson, Numerical Simulation of Aeroacoustic Field in a 2D Cascade Involving a Downstream Moving Grid Using the Space-Time CE/SE method, in *Proceedings of the First International Conference on Computational Fluid Dynamics, Kyoto, Japan, 10-14 July, 2000*, edited by N. Satofuka, (Springer-Verlag Berlin Heidelberg 2001), p. 543.
46. X.Y. Wang, S.C. Chang, A. Himansu, and P.C.E. Jorgenson, *Gust Acoustic Response of A Single Airfoil Using the Space-Time CE/SE Method*, AIAA Paper 2002-0801 (2002).
47. 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).
48. S.T. Yu and S.C. Chang, Applications of the Space-Time Conservation Element / Solution Element Method to Unsteady Chemically Reactive Flows," AIAA Paper 97-2099, in *A Collection of Technical Papers, 13th AIAA CFD Conference*, June 29-July 2, 1997, Snowmass, CO.
49. 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.
50. 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.
51. N.S. Liu and K.H. Chen, *Flux: An Alternative Flow Solver for the National Combustion Code*, AIAA Paper 99-1079.
52. G. Cook, *High Accuracy Capture of Curved Shock Front Using the Method of Conservation Element and Solution Element*, AIAA Paper 99-1008.
53. S.C. Chang, Y. Wu, X.Y. Wang, and V. Yang, Local Mesh Refinement in the Space-Time CE/SE Method, in *Proceedings of the First International Conference on Computational Fluid Dynamics, Kyoto, Japan, 10-14 July, 2000*, edited by N. Satofuka, (Springer-Verlag Berlin Heidelberg 2001), p. 61.
54. S.C. Chang, Z.C. Zhang, S.T. John Yu, and P.C.E. Jorgenson, A Unified Wall Boundary Treatment for Viscous and Inviscid Flows in the CE/SE Method, in *Proceedings of the First International Conference on Computational Fluid Dynamics, Kyoto, Japan, 10-14 July, 2000*, edited by N. Satofuka, (Springer-Verlag Berlin Heidelberg 2001), p. 671.
55. Z.C. Zhang, S.T. John Yu, S.C. Chang, and P.C.E. Jorgenson, Calculations of Low-Mach-Number Viscous Flows without Preconditioning by the Space-Time CE/SE method, in *Proceedings of the First International Conference on Computational Fluid Dynamics, Kyoto, Japan, 10-14 July, 2000*, edited by N. Satofuka, (Springer-Verlag Berlin Heidelberg 2001), p. 127.
56. A. Himansu, P.C.E. Jorgenson, X.Y. Wang, and S.C. Chang, Parallel CE/SE Computational via Domain Decomposition, in *Proceedings of the First International Conference on Computational Fluid Dynamics, Kyoto, Japan, 10-14 July, 2000*, edited by N. Satofuka, (Springer-Verlag Berlin Heidelberg 2001), p. 423.
57. Y. Wu, V. Yang, and S.C. Chang, Space-Time Method for Chemically Reacting Flows with Detailed Kinetics, in *Proceedings of the First International Conference on Computational Fluid Dynamics, Kyoto, Japan, 10-14 July, 2000*, edited by N. Satofuka, (Springer-Verlag Berlin Heidelberg 2001), p. 207.
58. I.S. Chang, *Unsteady Rocket Nozzle Flows*, AIAA Paper 2002-3884.
59. S.C. Chang, *Courant Number Insensitive CE/SE Schemes*, AIAA Paper 2002-3890 (2002).
60. I.S. Chang, *Unsteady Underexpanded Jet Flows*, AIAA Paper 2003-3885.

61. S.C. Chang and X.Y. Wang, *Multidimensional Courant Number Insensitive CE/SE Euler Solvers for Applications Involving Highly Nonuniform Meshes*, AIAA Paper 2003-5280.
62. 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.
63. B.S. Venkatachari, G.C. Cheng, and S.C. Chang, *Courant Number Insensitive Transient Viscous Flow Solver Based on CE/SE Framework*, AIAA Paper 2005-00931.
64. S.C. Chang, *Explicit von Neumann Stability Conditions for the c-τ Scheme—A Basic Scheme in the Development of the CE-SE Courant Number Insensitive Schemes*, NASA TM 2005-213627, April 2005.
65. J.C. Yen and D.A. Wagner, *Computational Aeroacoustics Using a Simplified Courant Number Insensitive CE/SE Method*, AIAA Paper 2005-2820.
66. I.S. Chang, C.L. Chang, and S.C. Chang, *Unsteady Navier-Stokes Rocket Nozzle Flows*, AIAA Paper 2005-4353.
67. S.C. Chang, *Courant Number and Mach Number Insensitive CE/SE Euler Solvers*, AIAA Paper 2005-4355
68. 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 CE/SE 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, 2005*, 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.
69. I.S. Chang, C.L. Chang, and S.C. Chang, *3D Unsteady Navier-Stokes Rocket Nozzle Flows*, AIAA Paper 2006-4775.
70. C.L. Chang, *Time-accurate, Unstructured-Mesh Navier-Stokes Computations with the Space-time CESE Method*, AIAA Paper 2006-4780.
71. S.C. Chang, *On Space-Time Inversion Invariance and Its Relation to Non-Dissipativeness of a CESE Core Scheme*, AIAA Paper 2006-4779.
72. S.C.Chang, *The $a(4)$ Scheme—A High Order Neutrally Stable CESE Solver*, AIAA Paper 2007-5820.
73. Other CESE references are posted on: http://www.grc.nasa.gov/www/microbus.
74. G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, 1986.
75. B. Noble and J.W. Daniel, *Applied linear Algebra*, Prentice-Hall Inc. (1977).
76. R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

TABLE 1.—NUMERICAL RESULTS OF THE *a*(3) AND DUAL *a* SCHEMES

$v = 0.1$				$t = 9.876$	
		$K = 25, n = 2{,}469$	$K = 50, n = 4{,}938$	$K = 100, n = 9{,}876$	$K = 200, n = 19{,}752$
E	$a(3)$	0.131×10^{-3}	0.143×10^{-4}	0.883×10^{-6}	0.549×10^{-7}
	a	0.452	0.115	0.287×10^{-1}	0.716×10^{-2}
E_x	$a(3)$	0.445×10^{-1}	0.977×10^{-3}	0.611×10^{-4}	0.382×10^{-5}
	a	2.90	0.732	0.182	0.454×10^{-1}
E_{xx}	$a(3)$	0.225	0.169	0.406×10^{-1}	0.100×10^{-1}

TABLE 2.—NUMERICAL RESULTS OF THE *a*(3) AND DUAL *a* SCHEMES

$v = 0.1$				$t = 10.00$	
		$K = 25, n = 2{,}500$	$K = 50, n = 5{,}000$	$K = 100, n = 10{,}000$	$K = 200, n = 20{,}000$
E	$a(3)$	0.228×10^{-3}	0.110×10^{-4}	0.628×10^{-6}	0.384×10^{-7}
	a	0.469	0.118	0.292×10^{-1}	0.727×10^{-2}
E_x	$a(3)$	0.154×10^{-1}	0.992×10^{-3}	0.623×10^{-4}	0.390×10^{-5}
	a	2.89	0.728	0.182	0.455×10^{-1}
E_{xx}	$a(3)$	0.473	0.316×10^{-1}	0.199×10^{-2}	0.124×10^{-3}

TABLE 3.—NUMERICAL RESULTS OF THE *a*(3) AND DUAL *a* SCHEMES

$v = 0.5$				$t = 49.38$	
		$K = 25, n = 2{,}469$	$K = 50, n = 4{,}938$	$K = 100, n = 9{,}876$	$K = 200, n = 19{,}752$
E	$a(3)$	0.168×10^{-3}	0.471×10^{-5}	0.294×10^{-6}	0.183×10^{-7}
	a	1.34	0.429	0.109	0.271×10^{-1}
E_x	$a(3)$	0.583×10^{-1}	0.856×10^{-12}	0.261×10^{-11}	0.678×10^{-11}
	a	8.73	2.73	0.686	0.171
E_{xx}	$a(3)$	1.01	0.942×10^{-1}	0.235×10^{-1}	0.587×10^{-2}

TABLE 4.—NUMERICAL RESULTS OF THE *a*(3) AND DUAL *a* SCHEMES

$v = 0.5$				$t = 50.00$	
		$K = 25, n = 2{,}500$	$K = 50, n = 5{,}000$	$K = 100, n = 10{,}000$	$K = 200, n = 20{,}000$
E	$a(3)$	0.362×10^{-13}	0.140×10^{-12}	0.229×10^{-12}	0.262×10^{-12}
	a	1.35	0.440	0.111	0.275×10^{-1}
E_x	$a(3)$	0.162×10^{-12}	0.845×10^{-12}	0.261×10^{-11}	0.682×10^{-11}
	a	8.73	2.73	0.689	0.172
E_{xx}	$a(3)$	0.172×10^{-9}	0.282×10^{-8}	0.185×10^{-7}	0.840×10^{-7}

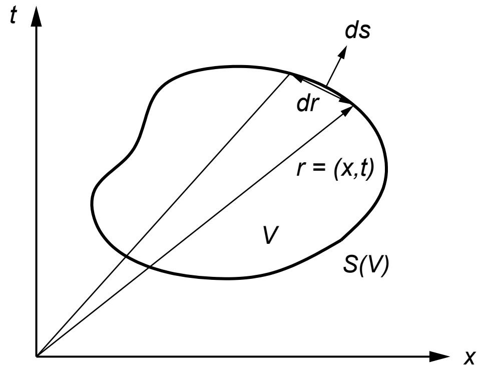
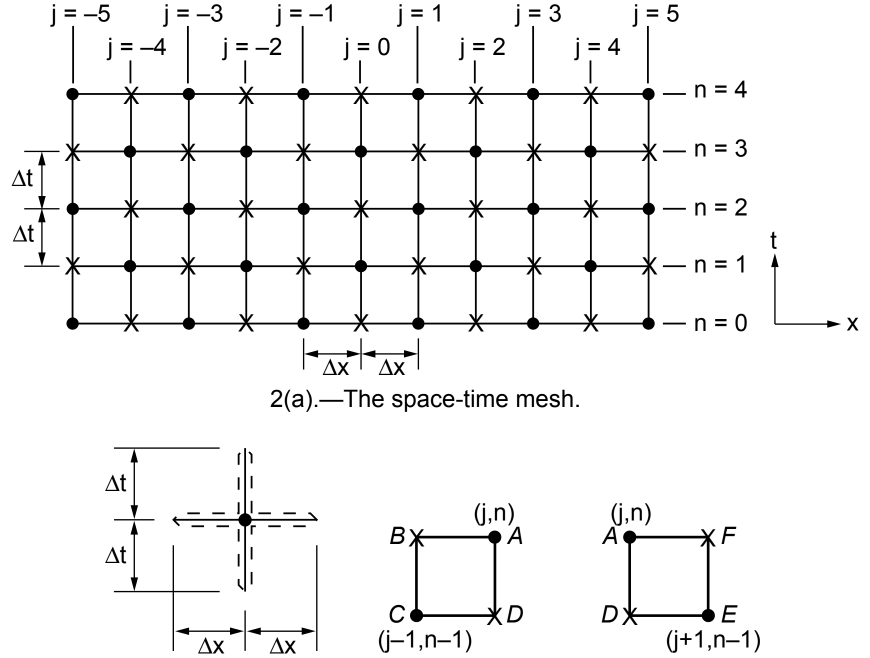


Figure 1.—A surface element on the boundary *S*(*V*) of an arbitrary space-time volume *V*.



2(a).—The space-time mesh.

2(b).—SE(j,n).

2(c).—CE_(j,n).

2(d).—CE$_+$(j,n).

Figure 2.—The SEs and CEs.

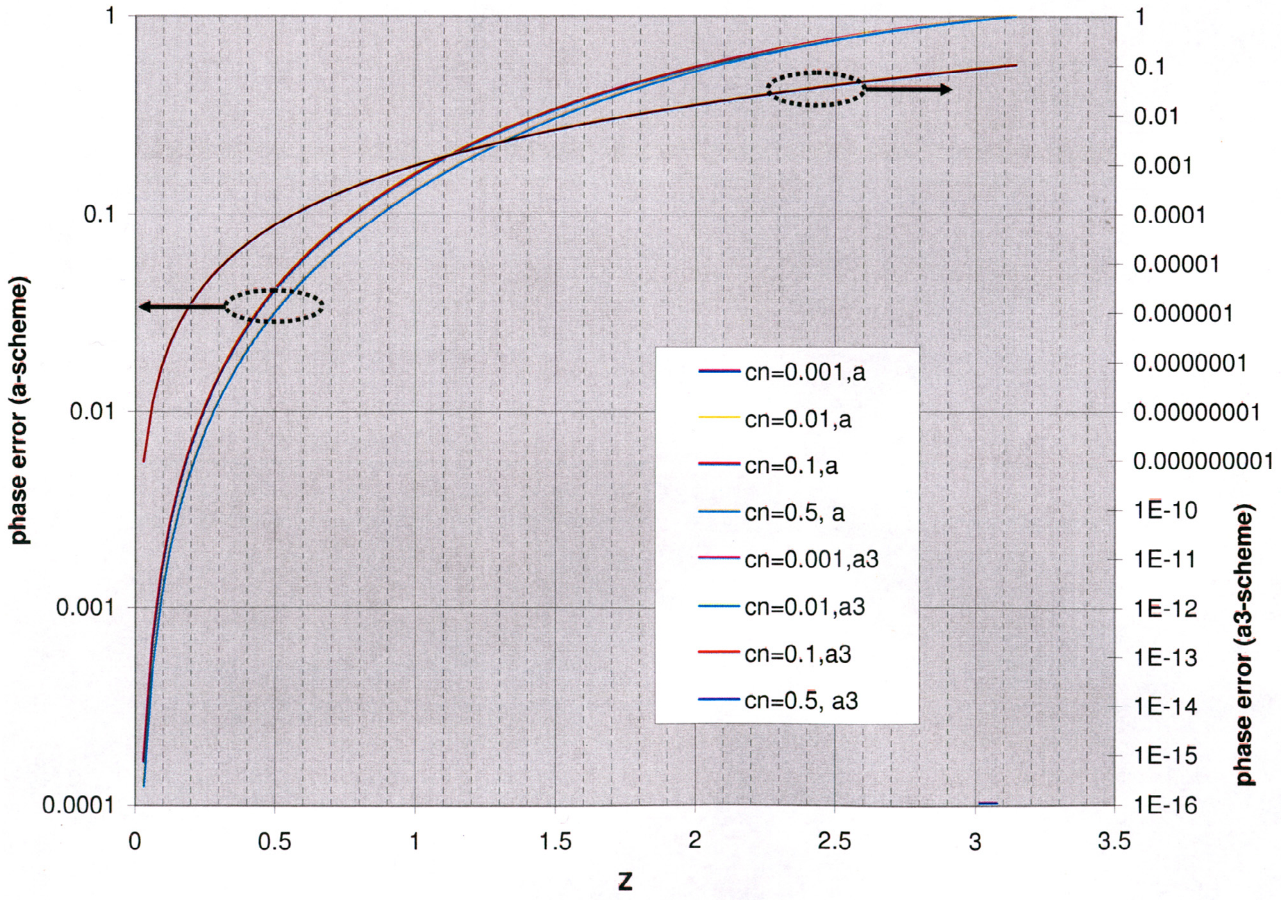


Figure 3.—Phase errors of the *a*(3) scheme and the dual *a* scheme.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 01-04-2008	**2. REPORT TYPE** Technical Memorandum	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE The *a*(3) Scheme--A Fourth-Order Space-Time Flux-Conserving and Neutrally Stable CESE Solver	**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-16150-2
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546-0001	**10. SPONSORING/MONITORS ACRONYM(S)** NASA **11. SPONSORING/MONITORING REPORT NUMBER** NASA/TM-2008-215138

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified-Unlimited
Subject Categories: 34, 64, 67, 70, and 71
Available electronically at http://gltrs.grc.nasa.gov
This publication is available from the NASA Center for AeroSpace Information, 301-621-0390

13. SUPPLEMENTARY NOTES

This report is a revised and expanded version of AIAA-2007-4321 presented at the 18th Computational Fluid Dynamics Conference sponsored by the American Institute of Aeronautics and Astronautics, Miami, Florida, June 25-28, 2007.

14. ABSTRACT

The CESE development is driven by a belief that a solver should (i) enforce conservation laws in both space and time, and (ii) be built from a non-dissipative (i.e., neutrally stable) core scheme so that the numerical dissipation can be controlled effectively. To initiate a systematic CESE development of high order schemes, in this paper we provide a thorough discussion on the structure, consistency, stability, phase error, and accuracy of a new 4th-order space-time flux-conserving and neutrally stable CESE solver of an 1D scalar advection equation. The space-time stencil of this two-level explicit scheme is formed by one point at the upper time level and three points at the lower time level. Because it is associated with three independent mesh variables (the numerical analogues of the dependent variable and its 1st-order and 2nd-order spatial derivatives, respectively) and three equations per mesh point, the new scheme is referred to as the *a*(3) scheme. Through the von Neumann analysis, it is shown that the *a*(3) scheme is stable if and only if the Courant number is less than 0.5. Moreover, it is established numerically that the *a*(3) scheme is 4th-order accurate.

15. SUBJECT TERMS

PT invariance; Neutral stable scheme; High order scheme; Space-time CESE method

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	74	**19b. TELEPHONE NUMBER** *(include area code)* 301-621-0390