Stability Analysis of Difference Methods for Parabolic Initial Value Problems

Ercília Sousa[1]

A decomposition of the numerical solution can be defined by the normal mode representation, that generalizes further the spatial eigenmode decomposition of the von Neumann analysis by taking into account the boundary conditions which are not periodic. In this paper we present some new theoretical results on normal mode analysis for a linear and parabolic initial value problem. Furthermore we suggest an algorithm for the calculation of stability regions based on the normal mode theory.

1. INTRODUCTION

The von Neumann analysis is based on a decomposition of the numerical solution in Fourier modes, implying periodic boundary conditions. While the matrix analysis is based on a decomposition in eigenmodes of the space discretisation operator, including the boundary conditions, nevertheless the normal mode analysis involves a decomposition of a more general form than the one based on the eigenmodes of the space discretisation and allows a detailed investigation of the influence of boundary conditions on stability.

The normal mode analysis constitutes perhaps the most powerful method for local analysis of the influence of boundary conditions. This method was initially presented by Godunov and Ryabenkii [5] and developed by Kreiss [9] and Osher [14]. The original Godunov–Ryabenkii theory provided a necessary condition for stability. A necessary and

---

[1] Departamento de Matemática, Universidade de Coimbra, Apartado 3008, 3000 Coimbra, Portugal. E-mail: ecs@mat.uc.pt

sufficient condition for stability was later developed by Gustafsson, Kreiss and Sundstrom [7], henceforth called GKS theory. This theory covers linear, first order hyperbolic systems in one space dimension. Since 1971, when GKS theory was first presented, related work has been done by Varah [27] for parabolic problems, by Strikwerda [21] for semi-discretised equations, by Michelson [11] for multidimensional problems and by Trefethen [25,26] where a relation between the GKS theory and group velocity is established.

Additional work on applying the normal mode analysis can be found in [1,6,12,13,15,17,19] showing us that this method often leads to very complex calculations. To overcome the difficulty of the theoretical approach, recently Thuné proposed a numerical algorithm to calculate GKS stability for linear hyperbolic equations [23] and linear hyperbolic systems [24].

We are interested in parabolic problems with convective and diffusive coefficients. The GKS theory that leads to necessary and sufficient conditions for stability was proven for hyperbolic problems. For parabolic problems we can apply the Godunov–Ryabenkii method which theoretically lead us to necessary conditions for stability although in a vast number of cases they appear to be also sufficient conditions.

The paper is organized as follows. In Sec. 2, we lay down a concise overview of the Godunov–Ryabenkii theory. We introduce some new developments of the theory in Sec. 3. These results centre on the determinant condition, the main condition we need to manipulate when performing the stability analysis. The determinant condition involves a discontinuous function. The discontinuity of this function is related to the multiplicity of the normal modes that are solutions of an equation called the characteristic equation. Here, we obtain a continuous function to replace the discontinuous function and thus increase our understanding of the phenomenom of normal modes with a multiplicity higher than one. In Sec. 4, we use the same continuous function to develop an algorithm to determine the stability of a numerical scheme with numerical boundary conditions.

2. GODUNOV–RYABENKII STABILITY ANALYSIS

The following is an adaptation of the von Neumann method for problems subject to non periodic boundary conditions and numerical boundary conditions. One essential aspect of normal mode analysis for the investigation of the influence of boundary conditions on the stability of a scheme is that the initial value problem needs to be stable for the Cauchy problem, which is best analysed with the von Neumann method. To make this paper reasonably self-contained, we give a general overview of the

Godunov–Ryabenkii theory in this section. For a complete description of the theory see for instance [8, 16, 22].

We present the stability theory for a quarter-plane $x, t \geqslant 0$. If there are two physical boundaries, then the theory shows that each boundary can be analysed separately. Thus, it is sufficient to study quarter-plane problems.

The model problem we consider is a linear and parabolic initial boundary problem. We have a convection-diffusion problem defined on a half-real line, for V and D positive constants:

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} = D \frac{\partial^2 u}{\partial x^2}, \quad 0 \leqslant x < \infty, \ t \geqslant 0, \tag{1}$$

$$u(x, 0) = f(x), \tag{2}$$

$$u(0, t) = 0, \tag{3}$$

$$u(\cdot, t) \in \mathrm{L}_2(0, \infty) \quad \text{for every fixed } t. \tag{4}$$

Suppose we have approximations U_j^n to the values $u(x_j, t_n)$ at the mesh points $x_j, \ j = 0, 1, 2, \ldots$ and assume we approximate the problem (1)–(4) by the difference scheme

$$U_j^{n+1} = Q U_j^n, \quad j = r, r+1, \ldots, \tag{5}$$

$$Q = \sum_{j=-r}^{p} a_j E^j, \quad E U_j^n = U_{j+1}^n, \tag{6}$$

where a_{-r} and a_p are non-zero. The a_j's also depend of parameters μ and ν defined by

$$\mu = \frac{D \Delta t}{\Delta x^2} \quad \text{and} \quad \nu = \frac{V \Delta t}{\Delta x}, \tag{7}$$

where Δt is the time step and Δx the space step. An important assumption is made: The finite difference scheme (5) is von Neumann stable and dissipative. We recall that a scheme is dissipative if the amplification factor, z, is of the form

$$|z(\theta)| \leqslant 1 - \delta \theta^{2r}, \quad \text{when } |\theta| < \pi \tag{8}$$

for $\delta \in \mathbb{R}^+$ and positive integer r.

The requirement of the scheme to be dissipative, arises quite naturally for difference schemes for parabolic equations. We require the scheme to be dissipative to strengthen the Godunov–Ryabenkii condition as it is explained in the end of this section in the context of the normal mode analysis.

Considering the finite difference scheme (5) we observe that as Q uses r points upstream, the basic approximation can not be used at $x_0, x_1, x_2, \ldots, x_{r-1}$, so there we need to apply boundary conditions. In our particular case the boundary given by the physical problem is associated only with the point x_0. At the other points the boundary conditions, called numerical boundary conditions, affect the difference scheme. Let us assume that the boundary conditions can be written as

$$U_\beta^{n+1} = \sum_{j=0}^{q} l_{\beta j} U_j^n, \quad \beta = 0, 1, \ldots, r-1. \tag{9}$$

The general technique is based on Laplace transforms of nodal values, which vary continuously with time. Therefore we define

$$U_j(t) = U_j^n \quad \text{for } t_n \leqslant t \leqslant t_{n+1}.$$

Applying the Laplace transform

$$\tilde{U}_j(s) = \int_0^\infty e^{-st} U_j(t) dt$$

to (5)–(9) gives (see [8] for details),

$$z\tilde{U}_j = Q\tilde{U}_j, \quad j = r, r+1, \ldots,$$

$$z\tilde{U}_\beta = \sum_{j=0}^{q} l_{\beta j} \tilde{U}_j, \quad \beta = 0, 1, \ldots, r-1,$$

$$\tilde{U} \in l_2(0, \infty),$$

where $z = e^{s\Delta t}$ and the coefficients $l_{\beta j}$ depend on ν and μ, and we assume that μ is constant when Δx varies.

Therefore, the eigenvalue problem associated with our approximation is:

$$z\phi_j = Q\phi_j, \quad j = r, r+1, \ldots, \tag{10}$$

$$z\phi_\beta = \sum_{j=0}^{q} l_{\beta j} \phi_j, \quad \beta = 0, 1, \ldots, r-1, \tag{11}$$

$$\phi \in l_2(0, \infty). \tag{12}$$

Lemma 2.1. (Godunov–Ryabenkii Condition). The approximation is unstable if the eigenvalue problem (10)–(12) has an eigenvalue $z \in \mathbb{C}$ with $|z| > 1$.

This follows from observing that if z is an eigenvalue of (10)–(12) with eigenfunction ϕ_j, it is also true that $U_j^n = z^n \phi_j$ is a solution of (5)–(9). At a fixed time t, we have

$$U_j^{t/\Delta t} = z^{t/\Delta t} \phi_j$$

and for decreasing Δt the solution grows without bound. □

At this point, we discuss how to solve the eigenvalue problem (10)–(12). Lemma 2.1 tells us that when z is such that $|z| > 1$, then we have an instability. What follows are results that help us to find an instability by looking for an eigenvalue z such that $|z| > 1$.

Consider the characteristic equation of the interior scheme. This equation is generated by substituting U_j^n by $z^n \kappa^j$ in the interior numerical scheme (5)–(6), obtaining

$$z - \sum_{j=-r}^{p} a_j \kappa^j = 0. \tag{13}$$

It can be proved that the general solution, ϕ_j, of the eigenvalue problem (10)–(12) can be written in terms of certain solutions, κ_a, of the characteristic equation (13) (see [8] for details). The general solution is given by

$$\phi_j = \sum_{|\kappa_a| < 1} P_a(j) \kappa_a^j, \quad \kappa_a = \kappa_a(z), \quad |z| > 1, \tag{14}$$

which depends on r free parameters $\sigma = (\sigma_1, \ldots, \sigma_r)^T$, $P_a(j)$ is a polynomial in j and its order is at most $m_a - 1$ where m_a is the multiplicity of κ_a. We will frequently refer to the solutions of (13) as the κ modes.

Remark 2.2. Note that if the roots κ_a are simple, from Eq. (14), the solution has the form

$$\phi_j = \sum_{|\kappa_a| < 1} \sigma_a \kappa_a^j, \tag{15}$$

for some constants σ_a.

In order to know how many modes κ, that are solutions of (13), are involved in the general solution (14) we present the following lemma.

Lemma 2.3. For $z \in \mathbb{C}$ such that $|z| > 1$, there is no solution of Eq. (13) with $|\kappa| = 1$ and there are exactly r solutions, counted according to multiplicity, with $|\kappa| < 1$.

Assume that there is a root $\kappa = e^{i\xi}, \xi \in \mathbb{R}$. Then Eq. (13) implies that $z = \sum_{j=-r}^{p} a_j e^{ij\xi}$. Because we have assumed that the approximation is von Neumann stable, we necessarily have $|z| = |\sum_{j=-r}^{p} a_j e^{ij\xi}| \leqslant 1$. This is a contradiction to the hypothesis $|z| > 1$; that is, there are no solutions κ with $|\kappa| = 1$. The solutions κ are continuous functions of z and cannot cross the unit circle. Therefore, the number of solutions with $|\kappa| < 1$ is constant for $|z| > 1$, and we can determine their number from the limit $|z|$ tends to $+\infty$. In this limit, the solutions with $|\kappa| < 1$ converge to zero and are, because $a_{-r} \neq 0$, to first approximation, determined by

$$z - a_{-r} \kappa^{-r} = 0.$$

This equation has exactly r solutions $\kappa = O(z^{-1/r})$. This proves the lemma.
□

Substituting (14) into the boundary conditions (11) yields a system of equations

$$C(z) \, \sigma = 0,$$

$\sigma = (\sigma_1, \ldots, \sigma_r)^T$, and we can rephrase Lemma 2.1 in the following form.

Lemma 2.4. The approximation is unstable for some $z \in \mathbb{C}$ with $|z| > 1$, if

$$\det C(z) = 0. \tag{16}$$

Direct application of Lemma 2.1.   □

Summarizing, the theory states that the interior scheme needs to be von Neumann stable and when considered in the half-plane $x \geqslant 0$, a mode κ^j with $|\kappa| > 1$ will lead to an unbounded solution in space, that is, κ^j will increase without bound when j goes to infinity. Therefore $|\kappa|$ should be less than one, and the Godunov–Ryabenkii (necessary) stability condition states that all the modes with $|\kappa| \leqslant 1$, generated by the boundary conditions, should correspond to $|z| < 1$. The form of the solution is very similar to the assumed Fourier modes, except that the amplitude of the spatial oscillation decays exponentially with j away from the boundary.

We can now explain in the context of normal mode analysis, the meaning of the interior difference scheme being dissipative. The dissipativity condition that we assume for the interior difference scheme (5), tells us that $|z| < 1$ for $\kappa = e^{\mathrm{i}\xi}, \xi \in \mathbb{R}, \xi \neq 0$. This condition strengthens the Godunov–Ryabenkii condition because the latter is a necessary condition that does not take into account the instability mechanism associated with $|z| = |\kappa| = 1$.

3. NEW RESULTS ON NORMAL MODE ANALYSIS

The function, $\det C(z)$, that appears on the left side of the determinant condition (16) is a discontinuous function in z, since the analytical form of the function changes according to whether the κ modes present multiple roots or not. In this section, we define another function that is continuous and has the same z roots as the previously mentioned function and is also used in Sec. 4 to develop an algorithm.

We consider a finite difference approximation of (1)–(4). The crucial parameters for the stability of the approximations are μ and ν defined by (7). Equations (13) and (16) form an implicit condition on μ and ν: the approximation of (1)–(4) is practically unstable for those values of μ and ν for which (13) and (16) have solutions with $|z| > 1$, $|\kappa_a| < 1$, $a = 1, \ldots, r$.

Let g be the function

$$g(z) = \det \ C(z). \tag{17}$$

The essential task when performing the stability analysis is to find the eigenvalues $z \in \mathbb{C}$ with $|z| > 1$ that are solutions of the determinant condition

$$g(z) = 0. \tag{18}$$

The form of the general solution ϕ_j of problem (10)–(12) in (14) changes according to the multiplicity of the roots $\kappa_a, a = 1, \ldots, r$, of Eq. (13), for each $|z| > 1$. Therefore the function g is not a continuous function. To see it more clearly, we write the matrix $C(z)$ explicitly in both cases, when the κ's are all simple and when there are some κ's that are confluent, that is, some of the κ's are multiple roots. We denote $C_r(z) := C(z)$, for z such that the κ's are simple roots, and $C_{\mathrm{conf}}(z) := C(z)$, for z such that the κ's are confluent.

Although, in what follows the κ's depend on z and it would be more accurate to write $\kappa(z)$, we will omit the variable z to have a clearer definition of the matrices.

The matrix $C_r(z)$ has the form:

$$
\begin{pmatrix}
z - \sum_{j=1}^{q} l_{0j}\kappa_1^j & \cdots & z - \sum_{j=1}^{q} l_{0j}\kappa_a^j & \cdots & z - \sum_{j=1}^{q} l_{0j}\kappa_r^j \\
\vdots & & \vdots & & \vdots \\
z\kappa_1^\beta - \sum_{j=1}^{q} l_{\beta j}\kappa_1^j & \cdots & z\kappa_a^\beta - \sum_{j=1}^{q} l_{\beta j}\kappa_a^j & \cdots & z\kappa_r^\beta - \sum_{j=1}^{q} l_{\beta j}\kappa_r^j \\
\vdots & & \vdots & & \vdots \\
z\kappa_1^{r-1} - \sum_{j=1}^{q} l_{r-1 j}\kappa_1^j & \cdots & z\kappa_a^{r-1} - \sum_{j=1}^{q} l_{r-1 j}\kappa_a^j & \cdots & z\kappa_r^{r-1} - \sum_{j=1}^{q} l_{r-1 j}\kappa_r^j
\end{pmatrix}
$$

Each column of the matrix $C_r(z)$ is associated with one of the κ's. We denote the columns of the matrix C_r by $\ell(\kappa_a)$, $a = 1, \ldots, r$, and the columns of the confluent matrix $C_{\text{conf}}(z)$ by $\ell_{\text{conf}}(\kappa_a)$, $a = 1, \ldots, r$. Suppose that one of the κ's, which we take to be the first one, κ_1, has multiplicity m, and the others are all simple. Then $\ell_{\text{conf}}(\kappa_a) = \ell(\kappa_a)$ for $a > m$, while for $1 < a \leqslant m$ we have

$$
\ell_{\text{conf}}(\kappa_a) =
\begin{pmatrix}
-\sum_{j=1}^{q} l_{0j} j^{a-1} \kappa_a^j \\
\\
z\kappa_a - \sum_{j=1}^{q} l_{1j} j^{a-1} \kappa_a^j \\
\vdots \\
z\beta^{a-1}\kappa_a^\beta - \sum_{j=1}^{q} l_{\beta j} j^{a-1} \kappa_a^j \\
\vdots \\
z(r-1)^{a-1}\kappa_a^{r-1} - \sum_{j=1}^{q} l_{r-1 j} j^{a-1} \kappa_a^j
\end{pmatrix}.
$$

Next we give the definition of a Vandermonde matrix and a confluent Vandermonde matrix since we use it below.

A Vandermonde matrix of order r is a matrix of the form

$$
V_r = V_r(\kappa_1, \kappa_2, \ldots, \kappa_r) =
\begin{pmatrix}
1 & 1 & \cdots & 1 \\
\kappa_1 & \kappa_2 & \cdots & \kappa_r \\
\vdots & \vdots & & \vdots \\
\kappa_1^{r-1} & \kappa_2^{r-1} & \cdots & \kappa_r^{r-1}
\end{pmatrix}.
\tag{19}
$$

By a confluence of the lth column into the ath column we mean the following limit operation: Replace in the lth column κ_l by $\kappa_a + \epsilon$ and subtract from it the ath column; divide this new column by ϵ and then let $\epsilon \to 0$. The resulting matrix is denoted by V_{conf}:

$$V_{\text{conf}} = \begin{pmatrix} 1 & \cdots & 1 & 0 & 1 & \cdots & 1 \\ \kappa_1 & \cdots & \kappa_{l-1} & 1 & \kappa_{l+1} & \cdots & \kappa_r \\ \kappa_1^2 & \cdots & \kappa_{l-1}^2 & 2\kappa_a & \kappa_{l+1}^2 & \cdots & \kappa_r^2 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \kappa_1^{r-1} & \cdots & \kappa_{l-1}^{r-1} & (r-1)\kappa_a^{r-2} & \kappa_{l+1}^{r-1} & \cdots & \kappa_r^{r-1} \end{pmatrix}.$$

In other words, V_{conf} is the same matrix as V_r except for the lth column, which is the derivative of the ath column. A matrix obtained from (19) by one or more confluences of columns is called a confluent Vandermonde matrix, see [2–4] for more details about Vandermonde matrices.

We denote by U_{bc} the vector of the approximate solution in the first r points of the mesh where we need to have numerical boundary conditions. We have that

$$U_{bc} = V(z)\ \sigma, \tag{20}$$

where $\sigma = (\sigma_1, \ldots, \sigma_r)^T$ are the coefficients in the solution (14) and $V(z)$ is the Vandermonde matrix or confluent Vandermonde matrix depending on the multiplicity of the κ's for each z.

Consider the vector $b = (b_1, \ldots, b_r)^T$ defined as

$$b = C(z)\ \sigma. \tag{21}$$

From (21) and (20) we obtain

$$b = C(z)V^{-1}(z)U_{bc} \quad \text{for each } z, \tag{22}$$

where $V(z)$ is invertible for $\kappa_i \neq \kappa_j$.

For different eigenvalues z the multiplicity of the κ's can change and consequently the matrices $C(z)$ and $V(z)$ change accordingly. In order to conclude that the matrix $C(z)V^{-1}(z)$ is continuous, we first prove some useful lemma.

Lemma 3.1. Assume that κ_1 has multiplicity of order m and that $\kappa_1, \ldots, \kappa_m$ are the confluent roots at the eigenvalue $z = z_{\text{conf}}$. Then there is a family of matrices $E(z)$ such that:

(a) $\lim\limits_{z \to z_{\text{conf}}} V_r(z)E(z) = V_{\text{conf}}(z_{\text{conf}})$,

(b) $\lim\limits_{z \to z_{\text{conf}}} C_r(z)E(z) = C_{\text{conf}}(z_{\text{conf}})$.

We have $V_r(z) = V_r(\kappa_1(z), \ldots, \kappa_l(z), \ldots, \kappa_r(z))$, for $z \neq z_{\text{conf}}$. The roots κ_1, $\ldots$, κ_m are confluent at $z = z_{\text{conf}}$. Then for $\alpha = 2, \ldots, m$

$$\kappa_\alpha(z) \to \kappa_1(z_{\text{conf}}), \quad \text{when } z \to z_{\text{conf}}.$$

We can write, for $z \neq z_{\text{conf}}$,

$$\kappa_2(z) = \kappa_1(z_{\text{conf}}) + \epsilon_1(z),$$
$$\kappa_3(z) = \kappa_1(z_{\text{conf}}) + \epsilon_2(z),$$
$$\vdots$$
$$\kappa_{m(z)} = \kappa_1(z_{\text{conf}}) + \epsilon_{m-1}(z).$$

For $\epsilon(z)$ defined as

$$\epsilon(z) = (\epsilon_1(z), \ldots, \epsilon_{m-1}(z)),$$

we want to prove that there exists a family of matrices $E(\epsilon(z))$ such that

$$\lim_{\epsilon(z) \to (0, \ldots, 0)} V_r(\epsilon(z)) E(\epsilon(z)) = V_{\text{conf}}(z_{\text{conf}}) \tag{23}$$

and

$$\lim_{\epsilon(z) \to (0, \ldots, 0)} C_r(\epsilon(z)) E(\epsilon(z)) = C_{\text{conf}}(z_{\text{conf}}), \tag{24}$$

where

$$V_r(\epsilon(z)) = V_r(\kappa_1(z_{\text{conf}}), \kappa_1(z_{\text{conf}}) + \epsilon_1(z), \ldots, \kappa_1(z_{\text{conf}}) + \epsilon_{m-1}(z), \ldots, \kappa_r(z_{\text{conf}})).$$

The matrices $C_r(\epsilon(z))$ are defined similarly.

The differences between $V_{\text{conf}}(z_{\text{conf}})$, $C_{\text{conf}}(z_{\text{conf}})$ and $V_r(z_{\text{conf}})$, $C_r(z_{\text{conf}})$, respectively, are in column 2 to column m. The $(i+1)$th column of the confluent matrices is the derivative of the ith column, in order to κ_1, for $1 \leqslant i \leqslant m-1$.

We have the following system with $m-1$ equations and $m-1$ asymptotic approximations of the derivatives in order to κ_1:

$$\ell(\kappa_2(z)) \approx \ell(\kappa_1(z_{\text{conf}})) + \sum_{i=1}^{m-1} \frac{\epsilon_1^i(z)}{i!} \ell^{(i)}(\kappa_1(z_{\text{conf}})),$$

$$\ell(\kappa_3(z)) \approx \ell(\kappa_1(z_{\text{conf}})) + \sum_{i=1}^{m-1} \frac{\epsilon_2^i(z)}{i!} \ell^{(i)}(\kappa_1(z_{\text{conf}})),$$

$$\vdots \qquad\qquad\qquad\qquad (25)$$

$$\ell(\kappa_m(z)) \approx \ell(\kappa_1(z_{\text{conf}})) + \sum_{i=1}^{m-1} \frac{\epsilon_{m-1}^i(z)}{i!} \ell^{(i)}(\kappa_1(z_{\text{conf}})),$$

where $\ell(\kappa_i(z))$, $i = 1, \ldots, m$ represent the i-column of $C_r(z)$ or $V_r(z)$. From the fact that the system (25) has a solution we can conclude that $\ell^{(i)}(\kappa_1(z_{\text{conf}}))$, $i = 1, \ldots, m-1$ can be written as a linear combination of $\ell(\kappa_i(z))$ $i = 1, \ldots, m$, where the linear coefficients depend on $\epsilon_1(z), \ldots, \epsilon_{m-1}(z)$. Therefore we conclude that there is a family of operators $E(\epsilon(z))$ that satisfies the conditions (23) and (24). $\qquad\square$

Example 3.2. This is an example on how to define explicitly the matrices $E(\epsilon(z))$, whose existence is proved in Lemma 3.1. We omit the variable z for the sake of clarity.

Suppose we have a 3 by 3 matrix $C = [\ell(\kappa_1)\ \ell(\kappa_2)\ \ell(\kappa_3)]$, where $\ell(\kappa_i)$, $i = 1, 2, 3$ are 3×1 columns. We assume that κ_1, κ_2 and κ_3 are confluent roots such that

$$\kappa_2 = \kappa_1 + \epsilon_1,$$
$$\kappa_3 = \kappa_1 + \epsilon_2$$

and we define $\epsilon = (\epsilon_1, \epsilon_2)$. The system (25), for this particular case, has two equations and by solving the system we get,

$$\ell'(\kappa_1) \approx -\frac{\epsilon_2 + \epsilon_1}{\epsilon_1 \epsilon_2} \ell(\kappa_1) - \frac{\epsilon_2}{\epsilon_1(\epsilon_1 - \epsilon_2)} \ell(\kappa_2) + \frac{\epsilon_1}{\epsilon_2(\epsilon_1 - \epsilon_2)} \ell(\kappa_3),$$

$$\ell''(\kappa_1) \approx \frac{2}{\epsilon_1 \epsilon_2} \ell(\kappa_1) + \frac{2}{\epsilon_1(\epsilon_1 - \epsilon_2)} \ell(\kappa_2) - \frac{2}{\epsilon_2(\epsilon_1 - \epsilon_2)} \ell(\kappa_3).$$

Therefore the matrix $E(\epsilon)$ such that

$$\lim_{\epsilon \to (0,0)} C(\epsilon) E(\epsilon) = [\ell(\kappa_1)\ \ell'(\kappa_1)\ \ell''(\kappa_1)]$$

is given by

$$E(\epsilon) = \begin{pmatrix} 1 & -\dfrac{\epsilon_1 + \epsilon_2}{\epsilon_1 \epsilon_2} & \dfrac{2}{\epsilon_1 \epsilon_2} \\[2ex] 0 & -\dfrac{\epsilon_2}{\epsilon_1(\epsilon_1 - \epsilon_2)} & \dfrac{2}{\epsilon_1(\epsilon_1 - \epsilon_2)} \\[2ex] 0 & \dfrac{\epsilon_1}{\epsilon_2(\epsilon_1 - \epsilon_2)} & -\dfrac{2}{\epsilon_2(\epsilon_1 - \epsilon_2)} \end{pmatrix}.$$

Theorem 3.3. The matrix $C(z)V^{-1}(z)$ is continuous.

The continuity problem of the matrix $C(z)V^{-1}(z)$ arises when for certain eigenvalues z the matrices $C(z)$ and $V(z)$ change because of the multiplicity of the κ's. Hence the matrix is continuous if

$$\lim_{z \to z_{\text{conf}}} C_r(z)V_r^{-1}(z) = C_{\text{conf}}(z_{\text{conf}})V_{\text{conf}}^{-1}(z_{\text{conf}}), \tag{26}$$

where $z = z_{\text{conf}}$ is the eigenvalue where the roots κ are confluent.

The equality (26) follows directly from the fact that exists a family of matrices $E(z)$ such that

$$\lim_{z \to z_{\text{conf}}} C_r(z)E(z) = C_{\text{conf}}(z_{\text{conf}}), \qquad \lim_{z \to z_{\text{conf}}} V_r(z)E(z) = V_{\text{conf}}(z_{\text{conf}}). \tag{27}$$

$\square$

Corollary 3.4. The function

$$d(z) = \det C(z)/\det V(z) \tag{28}$$

is continuous.

This result follows from the continuity of $C(z)V^{-1}(z)$ and the definition of the determinant. $\square$

Remark 3.5. For $z \in \mathbb{C}$, $g(z) = 0$ if and only if $d(z) = 0$.

The function $d(z)$ has the same roots as $g(z)$, it is a continuous function and also independent of ordering of κ's. Therefore on trying to solve $g(z) = 0$ for $|z| > 1$, we can have advantages in solving $d(z) = 0$ as it happens in Sec. 4.

4. A NUMERICAL APPLICATION

In this section we present an algorithm for the stability of linear parabolic equations defined by Godunov–Ryabenkii theory. The central point of our algorithm is the way we find the eigenvalues $z \in \mathbb{C}$, with $|z| > 1$, that are solutions of the determinant condition (18) and responsible for the phenomenon of instability.

4.1. General Idea

The most recent approach to the creation of an algorithm for stability investigation according to normal mode theory was given by Thuné [23, 24]. To investigate stability he used GKS theory for hyperbolic equations [23] and hyperbolic systems [24], taking advantage of the special structure of the system of algebraic equations whose solutions govern stability.

One of the difficulties associated with the determinant condition (18) is that there are discontinuities, due to the changes of the multiplicities of the κ's. Thuné [24] writes about the problem of multiplicities in Sec. 5.1. There, he notes (Lemma 6) that if $\kappa(z_0)$ has multiplicity greater than one, then $g_1(z_0) = 0$, where $g_1(z)$ is the determinant for the case when all eigenvalues have multiplicity one. Thus, he always used the form g_1 of g. If, during the iterative process, a solution z_0 was found such that $g_1(z_0) = 0$, then he subsequently checked whether the corresponding κ's had multiplicity one. If they had not, he went on to formulate $g_*(z)$, the correct $g(z)$ with respect to these multiplicities. If $g_*(z_0) = 0$, then z_0 was truly a solution, otherwise z_0 was considered a false alarm.

In our case, one of the important tools that is applied to determine the roots of $g(z)$ is the choice of the function $d_r(z)$ defined by

$$d_r(z) = \det\ C_r(z)/\det\ V_r(z). \tag{29}$$

Suppose that the function $d_r(z)$ is a meromorphic function, meaning that is analytic except for a finite number of poles, in a domain D,

$$D = \{z \in \mathbb{C} : |z| < 1 + r\}, \quad r \text{ real and positive.} \tag{30}$$

We assume those poles do not lie on the circle $|z| = 1$. Considering $d_r(z)$ analytic in the annulus A, where $A \subset D$ and is defined as

$$A = \{z \in \mathbb{C} : 1 - \alpha_1 < |z| < 1 + \alpha_2\}, \quad \alpha_1, \alpha_2 \text{ real and positive}, \tag{31}$$

we can approximate $d_r(z)$ by a Laurent series on A. Then for $z \in A$

$$d_r(z) = \sum_{k=-\infty}^{+\infty} b_k z^k \quad \text{with} \quad b_k = \frac{1}{2\pi i} \oint_C \frac{d_r(w)}{w^{k+1}} dw \qquad (32)$$

for all integer k, where C is the unit circle.

Our concern is to have a good approximation of the function $d_r(z)$ around the unit circle since the onset of instability is when a root z crosses the unit circle as μ and ν change. We denote $S_{n_1}^{n_2}(z)$ the truncated Laurent series that approximates $d_r(z)$,

$$S_{n_1}^{n_2}(z) = \sum_{k=-n_1}^{n_2} b_k z^k. \qquad (33)$$

The functions $d_r(z)$ and $S_{n_1}^{n_2}(z)$ also depend on the parameters μ and ν defined by (7), although we have omitted this in the notation, in the interest of clarity. In the same way the z roots of $d_r(z)$ and $S_{n_1}^{n_2}(z)$ depend on μ and ν, namely $z(\mu, \nu)$.

The basic idea of our method is: for each (μ, ν) we first check the practical von Neumann stability for the difference scheme. Inside the von Neumann stability region we know the exact number of κ's, the roots of the characteristic equation, that are less than one in modulus for each z with $|z| > 1$, namely r. Once we find these roots κ, we order them by magnitude so we can select the smallest first r roots and compute the function $d_r(z)$. The next step consists not in finding the z roots of the function $d_r(z)$ but in counting the number of z roots, of the approximated function $S_{n_1}^{n_2}(z)$, that lie inside the unit circle. With this counting process we expect to detect when one of the z's crosses the unit circle. Our numerical algorithm is implemented using MATLAB.

Algorithm:

```
for ν = ν₁ step ν_step until ν_s
        for μ = μ₁ step μ_step until μ_s
            for θ = 0 step θ_step until 2π
                    a. Check von Neumann condition, by doing Uⁿⱼ = zⁿeⁱʲθ
                        and impose |z| ⩽ 1.
            endfor
            if von Neumann unstable then
                b. instability.
            else
                for θ = 0 step θ_step until 2π
                        c. z = eⁱθ.
```

             d. Compute roots $\kappa_a(z)$ of the characteristic
               equation.
             e. Order roots by magnitude to find the κ's
               inside the unit circle.
             f. Compute $d_r(z) = \det C_r(z)/\det V_r(z)$.
          endfor
             g. Compute Laurent series $S_{n_1}^{n_2}(z)$ that approximates
               $d_r(z)$, by using the Fast Fourier Transform algorithm
               incorporated in MATLAB.
             h. Compute roots of $z^{n_1} S_{n_1}^{n_2}(z)$.
             i. Count number of roots of $z^{n_1} S_{n_1}^{n_2}(z)$ that are less than
               one in modulus.
        endif
      endfor
    endfor

Suppose we fix ν. Then, the roots of $d_r(z)$ are continuous functions of μ, $z(\mu)$. The instability is found when for some μ_0 there is an eigenvalue z_0 such that $|z_0(\mu_0)| > 1$. When we approximate $d_r(z)$ by $S_{n_1}^{n_2}(z)$ we determine the instability point by counting the number of roots of $S_{n_1}^{n_2}(z)$ that are less than one in modulus, as μ changes. If the number of z's that are less than one in modulus decreases at a certain μ_0, then we have found the instability point, since one of the roots that was less than one has become larger than one.

In Sec. 4.2 we apply the numerical algorithm described above to some particular problems. By doing this we hope to give a better understanding of how the algorithm works.

4.2. Test Problems

We consider a finite difference scheme associated with different numerical boundary conditions: one of the examples was taken from [19], where the stability of the problem is discussed analytically by using the Godunov–Ryabenkii theory and the other examples were taken from [20], where the stability analysis is studied by using matrix analysis. These problems were also discussed in [18]. In this section we obtain the stability results by using the algorithm described in Sec. 4.1, and they are consistent with those results obtained in [19, 20].

We consider a finite difference scheme called Quickest that was introduced by Leonard [10] in 1979 as an alternative to upwind and central differences. The Quickest scheme for the convection–diffusion equation (1) is given by,

$$U_j^{n+1} = [1 - 2c_1\Delta_0 + c_2\delta^2 + c_3\delta^2\Delta_-]U_j^n, \tag{34}$$

where $c_1 = v/2$, $c_2 = v^2/2 + \mu$ and $c_3 = (v/6)(1 - v^2 - 6\mu)$. The operators in (34) are the usual central, backward and second difference operators,

$$\Delta_0 U_j := \tfrac{1}{2}(U_{j+1} - U_{j-1}), \quad \Delta_- U_j := U_j - U_{j-1}, \quad \delta^2 U_j := U_{j+1} - 2U_j + U_{j-1}.$$

Quickest uses a third-order upwinding on the convective derivatives yielding a four-point scheme. Therefore, it uses two points upstream and this approximation can not be used at the points x_0 and x_1. At the point x_0 we apply the Dirichlet boundary condition associated with the physical boundary condition (3),

$$U_0^{n+1} = 0 \tag{35}$$

and at the point x_1 we need to consider a numerical boundary condition. We first consider the following numerical boundary condition:

$$U_1^{n+1} = [1 - 2c_1\Delta_0 + c_2\delta^2 + c_3\delta^2\Delta_+]U_1^n, \tag{36}$$

where $\Delta_+ U_j := U_{j+1} - U_j$.

The eigenvalue problem associated with our approximation is given by (34)–(36):

$$z\phi_j = [1 - 2c_1\Delta_0 + c_2\delta^2 + c_3\delta^2\Delta_-]\phi_j, \quad j \geqslant 2, \tag{37}$$

$$\phi_0 = 0, \tag{38}$$

$$z\phi_1 = [1 - 2c_1\Delta_0 + c_2\delta^2 + c_3\delta^2\Delta_+]\phi_1. \tag{39}$$

The Godunov–Ryabenkii condition tell us that if the system (37)–(39) has an eigenvalue $z \in \mathbb{C}$ with $|z| > 1$, then the approximation (34)–(36) is unstable. As explained before these eigenvalues $z \in \mathbb{C}$ with $|z| > 1$ are the same as the roots of the function $d(z)$ defined by (28).

The characteristic equation for the interior scheme (34) is a cubic equation of the form,

$$\kappa^3(-c_1 + c_2 + c_3) + \kappa^2(-z + 1 - 2c_2 - 3c_3) + \kappa(c_1 + c_2 + 3c_3) - c_3 = 0. \tag{40}$$

From Lemma 2.3 this equation has not solutions $\kappa = e^{i\xi}$, ξ real for $|z| > 1$ and there are exactly two solutions κ_i, $i = 1, 2$ with $|\kappa_i| < 1$ for $|z| > 1$.

In this particular example, the matrices $C_r(z)$ and $V_r(z)$ defined in the beginning of Sec. 3 are given by

$$C_r(z) = \begin{pmatrix} 1 & 1 \\ g_1(\kappa_1, z, \mu, v) & g_1(\kappa_2, z, \mu, v) \end{pmatrix}, \qquad V_r(z) = \begin{pmatrix} 1 & 1 \\ \kappa_1 & \kappa_2 \end{pmatrix},$$

where κ_1 and κ_2 are the solutions of the cubic equation (40) and the function $g_1(\kappa, z, \mu, \nu)$, associated with the numerical boundary condition (36), is defined as

$$g_1(\kappa, z, \mu, \nu) = \kappa^3 c_3 + \kappa^2(-c_1 + c_2 - 3c_3) + \kappa(1 - 2c_2 + 3c_3 - z).$$

Using the matrices defined above, the function $d_r(z)$, given by (29), can be written in the form

$$d_r(z) = \frac{g_1(\kappa_2, z, \mu, \nu) - g_1(\kappa_1, z, \mu, \nu)}{\kappa_2 - \kappa_1}. \tag{41}$$

The function defined in (41) must be meromorphic in order to be approximated by a Laurent series. We have that κ_1 and κ_2 are functions of z. For $z \in \mathbb{C}$ such that $\kappa_1(z) \neq \kappa_2(z)$, we can show that

$$d_r(z) = (\kappa_1^2 + \kappa_1\kappa_2 + \kappa_2^2)c_3 + (\kappa_1 + \kappa_2)(-c_1 + c_2 - 3c_3) + 1 - 2c_2 + 3c_3 - z. \tag{42}$$

Consequently, the analyticity of the function $d_r(z)$ is associated with the functions $\kappa_i(z)$, $i = 1, 2$ that are solutions of the cubic equation (40).

Suppose we write the cubic equation (40) as

$$\kappa^3 + a_2\kappa^2 + a_1\kappa + a_0 = 0, \tag{43}$$

where

$$a_2 = \frac{-z + 1 - 2c_2 - 3c_3}{-c_1 + c_2 + c_3}, \quad a_1 = \frac{c_1 + c_2 + 3c_3}{-c_1 + c_2 + c_3}, \quad a_0 = \frac{-c_3}{-c_1 + c_2 + c_3}. \tag{44}$$

Then for $\kappa_i = x_i - \lambda$, with $\lambda = (1/3)a_2$ and defining

$$p = \frac{3a_1 - a_2^2}{3} \quad \text{and} \quad q = \frac{9a_1 - 27a_0 - 2a_2^3}{27} \tag{45}$$

the cubic equation (43) can be written in the form

$$x^3 + px + q = 0. \tag{46}$$

The solutions, x_i's, of (46) satisfy the Newton's identities and after some algebra we have that

$$x_1^2 + x_1x_2 + x_2^2 = -p \tag{47}$$

and therefore the roots, κ_i's, of the cubic equation (43) verify

$$\kappa_1^2 + \kappa_1\kappa_2 + \kappa_2^2 = -p + 3\lambda^2 - \lambda\frac{q}{p}. \tag{48}$$

According to (48) the function $d_r(z)$ given by (42) has poles at $z \in \mathbb{C}$ such that $p(z) = 0$, this is, at $z \in \mathbb{C}$ such that $a_2(z) = \pm \sqrt{3a_1}$. Note that only the coefficient a_2 of (43) depends on z. Furthermore, from [19] we know that $\kappa_1 + \kappa_2$ has a pole. Then, the function $d_r(z)$ is analytic except for a finite number of poles, this is, it is a meromorphic function.

We approximate the function $d_r(z)$ by the truncated Laurent series $S_{n_1}^{n_2}(z)$ defined on (33). Recall that the functions $d_r(z)$ and $S_{n_1}^{n_2}(z)$ depend also on the parameters μ and ν. The coefficients b_k on (33) are computed using the fast Fourier transform algorithm in MATLAB and then we calculate the roots of the polynomial function $z^{n_1} S_{n_1}^{n_2}(z)$. For each μ and ν we count the roots, of this polynomial function, that are inside the unit circle and we detect that for a certain value of μ and ν one of the z roots that was inside the unit circle travels to the outside. The fact that the polynomial function $S_{n_1}^{n_2}(z)$ approximates $d_r(z)$ on the unit circle assures us that when a root of $z^{n_1} S_{n_1}^{n_2}(z)$ crosses the unit circle, this root approximates one of the roots of $d_r(z)$.

In Fig. 1, we plot the z roots of the polynomial function $z^{15} S_{15}^{16}(z)$ for $\nu = 0.4$ and μ changing, and we can observe that at some point one of the z roots crosses the unit circle at $z = -1$. We show only the roots z with $|z| < 2$, although we have 31 roots in total. Figure 2 plots the output of the algorithm and we observe it is in agreement with the theoretical approach presented in [19]. The crossing at $z = -1$ in Fig. 1 was also determined analytically in the same paper.

Next we apply the numerical algorithm to two additional numerical boundary conditions discussed in [20] and called, the Lax–Wendroff numerical boundary condition and the numerical boundary condition using a fictitious point, given by

$$U_1^{n+1} = [1 - 2c_1 \Delta_0 + c_2 \delta^2] U_1^n \qquad (49)$$

and

$$U_1^{n+1} = [1 - 2c_1 \Delta_0 + c_2 \delta^2] U_1^n + c_3 (U_2 - 3U_1 + \beta U_1), \qquad (50)$$

respectively, where $\beta = (-c_1 + c_2)/(c_1 + c_2)$. The interior scheme is still defined by (34) with the Dirichlet boundary condition (35). However, since we have a different numerical boundary condition the matrix $C_r(z)$ changes and instead of the function $g_1(\kappa, z, \mu, \nu)$ we have the functions $g_2(\kappa, z, \mu, \nu)$ and $g_3(\kappa, z, \mu, \nu)$ associated, respectively, with the Lax–Wendroff numerical boundary condition and the numerical boundary condition using a fictitious point. Thus the functions g_2 and g_3 are as
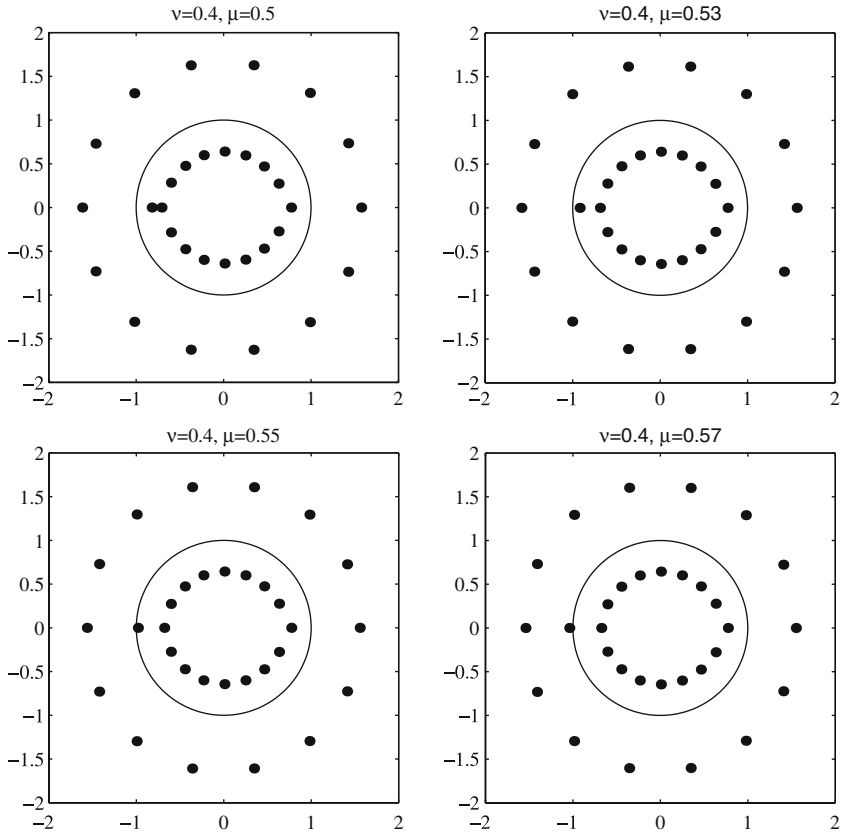
Fig. 1. Roots of the approximated function $S_{15}^{16}(z)$ and the unit circle $|z|=1$.

follows:

$$g_2(\kappa, z, \mu, \nu) = \kappa^2(c_1 - c_2) + \kappa(z - 1 + 2c_2),$$
$$g_3(\kappa, z, \mu, \nu) = \kappa^2(c_1 - c_2 - c_3) + \kappa(z - 1 + 2c_2 + (3 - \beta)c_3).$$

We obtain the stability regions plotted in Figs. 3a and b that are according to the results in the paper by Sousa and Sobey [20]. In Fig. 3b we do not have any Godunov–Ryabenkii eigenmodes since the method is stable in all the von Neumann stability region.
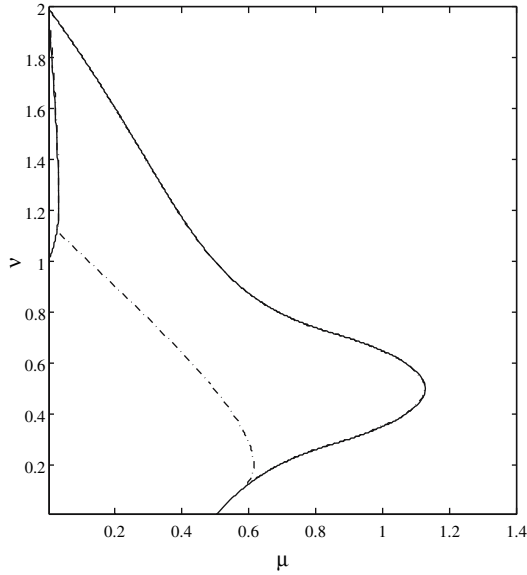
Fig. 2. Stability region for the downwind third difference numerical boundary condition: von Neumann condition (–) and Godunov–Ryabenkii condition (- · -) using the numerical algorithm with the approximated function $S_{15}^{16}(z)$.
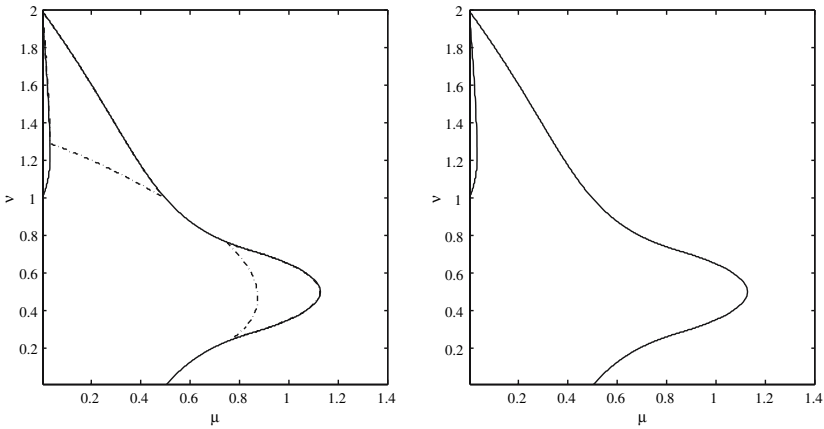

Fig. 3. von Neumann condition (–) using the numerical algorithm with the approximated function $S_{15}^{16}(z)$: (a) Stability region for the Lax–Wendroff numerical boundary condition; (b) Stability region for the numerical boundary condition using a fictitious point.

5. CONCLUDING REMARKS

The essential task when performing normal mode stability analysis is to find the solutions of the determinant condition. In this paper, we have presented a continuous function associated with the determinant condition and the main idea amounts to a new way of looking at the phenomenon of multiple roots. We have then taken advantage of this function to build an algorithm for the implementation of stability analysis according to Godunov–Ryabenkii theory.

ACKNOWLEDGMENTS

REFERENCES

1. Carpenter, M. H., Gottlieb, D. and Abarbanel, S. (1993). The stability of numerical boundary treatments for compact high-order finite-difference schemes. *J. Comput. Phys.* **108**, 272–295.
2. Gautschi, W. (1962). On inverses of Vandermonde and confluent Vandermonde matrices. *Numerische Mathematik* **4**, 117–123.
3. Gautschi, W. (1963). On inverses of Vandermonde and confluent Vandermonde matrices II. *Numerische Mathematik* **5**, 425–430.
4. Gautschi, W. (1978). On inverses of Vandermonde and confluent Vandermonde matrices III. *Numerische Mathematik* **29**, 445–450.
5. Godunov, S. K. and Ryabenkii, V. S. (1963). Spectral criteria for the stability of boundary problems for non-self-adjoint difference equations. *Uspekhi Mat. Nauk.* **18**(3) (In Russian).
6. Goldberg, M., and Tadmor, E., (1978). Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. *Math. Comput.* **32**, 1097–1107.
7. Gustafsson, B., Kreiss, H. -O. and Sundstrom, A. (1972). Stability theory of difference approximations for mixed initial boundary value problems, II. *Math. Comput.* **26**, 649–686.
8. Gustafsson, B., Kreiss, H. -O. and Oliger, J. (1995). Time-dependent Problems and Difference Methods, Wiley-Interscience, New York.
9. Kreiss, H. -O. (1968). Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comput.* **22**, 703–714.
10. Leonard, B. P. (1979). A stable and accurate convective modelling procedure based on quadratic upstream interpolation. *Comput. Methods Applied Mechanics and Engineering* **19**, 59–98.
11. Michelson, D. (1983). Stability theory of difference approximations for multidimensional initial-boundary value problems. *Math. Comput.* **40**, 1–45.
12. Oliger, J. (1974). Fourth order difference methods for the initial boundary-value problem for hyperbolic equations. *Math. Comput.* **28**, 15–25.

13. Oliger, J. (1976). Hybrid difference methods for the initial boundary-value problem for hyperbolic equations. *Math. Comput.* **30**, 724–738.
14. Osher, S. (1969). Stability of difference approximations of dissipative type for mixed initial-boundary value problems. *Math. Comput.* **23**, 335–340.
15. Otto, K., and Thuné, M. (1989). Stability of a Runge-Kutta method for the Euler equations on a substructured domain. *SIAM J. Scien Stat. Comput.* **10**, 154–174.
16. Richtmyer, R. D., and Morton, K. W. (1967). *Difference Methods for Initial-value Problems*, 2nd ed., Wiley-Interscience, New York.
17. Sloan, D. M. (1983). Boundary conditions for a fourth order hyperbolic difference scheme. *Math. Comput.* **41**, 1–11.
18. Sousa, E. (2001). Finite Differences for the Convection-Diffusion Equation: On Stability and Boundary Conditions. Ph.D. Thesis, Oxford University.
19. Sousa, E. (2001). A Godunov–Ryabenkii instability for a Quickest scheme. *Lecture Notes Comput. Sci.* **1988**, 732–740.
20. Sousa, E., and Sobey, I. J. (2002). On the influence of boundary conditions. *Appl. Numerical Math.* **41**, 325–344.
21. Strikwerda, J. (1980). Initial boundary value problems for the method of lines. *J. Comput. Phys.* **34**, 94–107.
22. Strikwerda, J. (1989). *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks, California.
23. Thuné, M. (1986). Automatic GKS stability analysis. *SIAM J. Sci. Stat. Comput.* **7**, 959–977.
24. Thuné, M. (1990). A numerical algorithm for stability analysis of difference methods for hyperbolic systems. *SIAM J. Sci. Stat. Comput.* **11**, 63–81.
25. Trefethen, L. N. (1983). Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundstrom. *J. Comput. Phys.* **49**, 199–217.
26. Trefethen, L. N. (1984). Instability of difference models for hyperbolic initial boundary value problems. *Comm. Pure Appl. Math.* **37**, 329–367.
27. Varah, J. M. (1971). Stability of difference approximations to the mixed initial boundary value problems for parabolic systems. *SIAM J. Numerical Anal.* **8**, 598–615.