SIAM J. SCI. COMPUT.
Vol. 16, No. 1, pp. 88–118, January 1995

007

# CONVERGENCE ANALYSIS OF THE DEFECT-CORRECTION ITERATION FOR HYPERBOLIC PROBLEMS[*]

J.-A. DÉSIDÉRI[†] AND P. W. HEMKER[‡]

**Abstract.** This paper explains some of the convergence behaviour of iterative implicit and defect-correction schemes for the solution of the discrete steady Euler equations. Such equations are also commonly solved by (pseudo) time integration, the steady solution being achieved as the limit (for $t \to \infty$) of the solution of a time-dependent problem. Implicit schemes are then often chosen for their favourable stability properties, permitting large timesteps for efficiency. An important class of implicit schemes involving first- and second-order accurate upwind discretisations is considered. In the limit of an infinite timestep, these implicit schemes approach defect-correction algorithms. Thus our analysis is informative for both types of construction.

Simple scalar linear model problems are introduced for the one-dimensional and two-dimensional cases. These model problems are analyzed in detail by both Fourier and matrix analyses. The convergence behaviour appears to be strongly dependent on a parameter $\beta$ that determines the amount of upwinding in the discretisation of the second-order scheme.

In general, in the convergence behaviour of the iteration, after an impulsive initial phase a slower pseudo-convective (or Fourier) phase can be distinguished, and then a faster asymptotic phase. The extreme parameter values $\beta = 0$ (no upwinding) and $\beta = 1$ (full second-order upwinding) both appear as special cases for which the convergence behaviour degenerates. They are not recommended for practical use. For the intermediate values of $\beta$ the pseudo-convection phase is less significant. Fromm's scheme ($\beta = 1/2$) or van Leer's third-order scheme ($\beta = 1/3$) show quite satisfactory convergence behaviour.

In the last section experiments for the steady Euler equations are discussed. Comments are given on how well phenomena, understood for the scalar linear model problem, are recognised for this system of more complex nonlinear equations.

**Key words.** defect correction, hyperbolic problems, iterative implicit methods, upwind schemes, Euler flows

**AMS subject classifications.** 65L, 65M, 65N

## 1. Introduction.

The Euler equations govern the motion of inviscid compressible flow. When the flow is supersonic, the steady Euler equations are hyperbolic, and they can be integrated by a "space-marching" procedure that sweeps the domain starting from one boundary where it is appropriate to specify the data. These methods are very efficient because they are noniterative. However, for many situations of practical interest, the flow is not supersonic throughout the domain. For example, when a blunt body is in supersonic motion, a shock wave forms in front of the body and it is well known in aerodynamics that in such a flow there always exists a subsonic pocket between the shock wave and the body. There the steady Euler equations are elliptic in nature, and a direct space-marching procedure is not applicable. (In addition, the boundaries of the elliptic region are usually not known a priori.) The solution method is then necessarily iterative. The efficiency of the iterative method is then of crucial importance, since the discretisation may yield a large number of unknowns.

One way to construct an iteration that yields the solution to the steady Euler equations at convergence is to integrate forward in time the time-dependent Euler equations which are hyperbolic, regardless of the flow regime. The initial solution is then arbitrary and the timestep is viewed as a relaxation parameter. This paper focuses on the iterative properties of a certain class of implicit (pseudo-)time-integration methods commonly employed in compressible flow computations to solve, at convergence, the steady Euler equations.

[*] Received by the editors December 19, 1990; accepted for publication (in revised form) November 18, 1993.

[†] Institut National de Recherche en Informatique et en Automatique, B.P. 93, 06902 Sophia Antipolis Cedex, France

[‡] Centrum voor Wiskunde en Informatica, Centre for Mathematics and Computer Science, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

In order to perform a theoretical analysis of the rate of convergence, a very simple model is introduced. For this, recall that in two dimensions the Euler equations can be written in the following quasi-linear form:

$$w_t + Aw_x + Bw_y = 0, \tag{1}$$

in which $A = A(w)$ and $B = B(w)$ are the usual $4 \times 4$ Jacobian matrices, which can be diagonalised explicitly. To allow a *linear analysis* of numerical schemes, one may construct a hyperbolic model equation by setting $A$ and $B$ to constant matrices, subject to the condition that any linear combination of $A$ and $B$ should be diagonalisable.

In one dimension, and after diagonalisation, the above system reduces to a set of *convection equations*, and an appropriate model is given by the following quarter-plane problem:

$$\begin{cases} u_t + cu_x = 0 & (c > 0;\ t > 0,\ x > 0), \\ u(x, 0) = 0 & (x > 0), \\ u(0, t) = 1 & (t > 0). \end{cases} \tag{2}$$

This is a purely convective problem, in which information travels without dissipation along characteristics. The process of convergence (to steady state, on a fixed spatial interval $[0, X]$) is therefore distinct from that of a dissipative phenomenon. However, in discrete models, dissipation may exist in the form of *artificial dissipation*.

For the model problem, various differencing schemes may be employed to represent the spatial derivative $u_x$: central differencing, $\delta_x^c$; first-order backward differencing, $\delta_{x,1}^u$; second-order backward differencing, $\delta_{x,2}^u$.[1] The matrix analogues of these operators can be written down in a precise way that accounts for the left-boundary condition, and assumes that the central difference operator is replaced by backward differencing at the right boundary. These matrix models differ only by a small number of elements from those one would construct to study a similar linear, hyperbolic, purely initial value model problem obtained from the above by replacing the boundary condition with the assumption of a spatially periodic solution. It is emphasised that these models are very distinct in nature. In the periodic case, one usually avoids the matrix notation, in which linear operators are represented by circulant matrices that are all simultaneously diagonalised by the discrete Fourier transform. Hence, the analysis is directly carried in terms of eigenvalues (discrete Fourier analysis). The periodic model, which does not contain the effect of boundary conditions, is adequate to yield L2-stability conditions, phase-error, and wavespeed evaluations. On the other hand, the nonperiodic model, (2), is more appropriate for (asymptotic) iterative convergence rate estimations.

For both discrete equations the following general statements can be made.

(1) All the *eigenvalues* $\lambda_m$ of any acceptable differencing scheme *fall on the same half-plane*

$$\forall m, \quad \Re(\lambda_m) \geq 0. \tag{3}$$

(2) The *central-difference* operator can be *diagonalised*. In the periodic case, the eigenvalues are purely imaginary, which indicates the absence of artificial viscosity. In the nonperiodic case, with $N$ equations, all but one eigenvalue can be shown to satisfy $\Re(\lambda_m) = O(\log(N)/N)$.

(3) In the nonperiodic case, pure *upwind schemes* are represented by defective (i.e., nondiagonalisable) matrices with multiple eigenvalues, whose real parts are of order 1.

[1] Backward differences are considered in the definition of upwind schemes since $c > 0$ has been assumed.

We observe that in the case of the Euler equations, upwind schemes satisfying statement (1) are constructed via flux or flux-difference splitting, which isolates the contributions from the positive and the negative eigenvalues prior to applying a backward- or forward-type differencing scheme.

**1.1. Implicit schemes.** We now examine more closely the time-discretisation method, sometimes referred to as the *solver*, when only the converged, steady-state solution is of interest. Then implicit schemes are attractive because they are not limited by the CFL stability condition, and therefore allow rapid convergence to steady state when large timesteps are employed. Here we concentrate on the *linearised backward Euler scheme* for the solution of the semidiscrete system $\mathbf{u}_t + D_h(\mathbf{u}) = 0$:

$$M_h\left(\mathbf{u}^{n+1} - \mathbf{u}^n\right) = -\Delta t D_h(\mathbf{u}^n)\,, \tag{4}$$

where $\Delta t$ is the timestep and the operator $M_h$, which is defined by

$$M_h = I + \Delta t\left(\frac{\partial D_h(\mathbf{u}^n)}{\partial \mathbf{u}^n}\right), \tag{5}$$

involves the Jacobian of the discrete set of equations to be solved, $D_h(\mathbf{u}^n) = 0$. To evaluate the stability of this method we consider again the linear hyperbolic model, for which $D_h$ and $M_h$ can be thought of as matrices constant during the iteration and satisfying

$$M_h = I + \Delta t D_h\,, \tag{6}$$

so that an amplification matrix $G_{\Delta t}$ can be defined by $\mathbf{u}^{n+1} = G_{\Delta t}\mathbf{u}^n + \mathbf{b}$. Here $\mathbf{b}$ is a constant vector containing prescribed boundary terms, and $G_{\Delta t}$ turns out to be

$$G_{\Delta t} = I - \Delta t M_h^{-1} D_h = I - \Delta t\,(I + \Delta t D_h)^{-1} D_h\,. \tag{7}$$

Thus, if the eigenvalues of $D_h$ are denoted as before by $\lambda_m$ ($m = 1, 2, \ldots, N$), those of $G_{\Delta t}$ are given by

$$g_m(\Delta t) = 1 - \frac{\Delta t \lambda_m}{1 + \Delta t \lambda_m} = \frac{1}{z_m + 1}\,, \tag{8}$$

where $z_m = \lambda_m \Delta t$. Since for all $m$, $\Re(z_m) \geq 0$, it follows that for all $\Delta t$

$$|g_m(\Delta t)| \leq 1, \tag{9}$$

thus proving that the method is *unconditionally stable for the associated linear hyperbolic problem.* Furthermore,

$$\lim_{\Delta t \to \infty} g_m(\Delta t) = 0\,. \tag{10}$$

Of course these results, valid for a linear model, may not entirely extend to the nonlinear case. However, the Euler implicit method is stable for values of $\Delta t$ that are *not limited by the CFL condition.* It becomes more dissipative with larger timesteps while the steady-state solution, which is independent of $\Delta t$, is only determined by the differencing operator $D_h$ appearing explicitly on the right-hand side. (This in contrast to the Lax–Wendroff type schemes, for example, for which the steady state depends on the timestep used.) If one lets $\Delta t \to \infty$, we recognise Newton's method.[2]

[2] For a linear problem, the amplification matrix $G_\infty$ is then equal to the null matrix and the process converges in one iteration step. This confirms that the ability to use stable, very large timesteps is a highly desirable feature for the solver. This property does not hold for factored implicit schemes.

We now examine the algorithmic standpoint. The application of the algorithm defined in eq. (4) is usually performed in three steps: (i) the *explicit* or *physical phase*: evaluation of the right-hand side vector $R = -\Delta t D_h \mathbf{u}^n$; (ii) the *implicit* or *mathematical phase*: solution of the system $M_h \Delta \mathbf{u}^n = R$, in which the unknown is the vector $\Delta \mathbf{u}^n$; (iii) the *update*: $\mathbf{u}^{n+1} = \mathbf{u}^n + \Delta \mathbf{u}^n$.

The implicit phase preconditions the system in a way that enhances the stability of the method but has no effect on steady-state accuracy. The physical phase alone defines the converged solution. Therefore we require that the operator $D_h$ be at least second-order accurate in regions where the solution is smooth. This is achieved either by a *central differencing scheme* [1], [12] or a *second-order upwind scheme*. The latter alternative has gained some popularity in recent papers [10], [11], [13], [14] because it yields schemes that have better monotonicity properties and thus produce more physically relevant solutions near discontinuities. Another alternative is to combine a central discretisation with an upwind discretisation [9], [15]. In any case, the evaluation of a second-order accurate space discretised operator is generally attained with moderate difficulty.

Constructing a second-order accurate preconditioner is a much more complex task, however. Partly for this reason it is interesting to consider the case of a first-order approximation of the Jacobian in the preconditioner. This alternative is also attractive because for a constant-coefficient hyperbolic model problem, the system of linear equations to be solved at each timestep is diagonally dominant. Hence its solution can be carried by relaxation (e.g., Gauss–Seidel or Jacobi iteration). However, with a first-order implicit preconditioner, some inconsistency is introduced in the formulation, and the efficiency of the method at large timesteps can no longer be that of Newton's method. The rate of convergence to steady state is the main subject of the present article, in particular for a class of implicit schemes in which a parameter $\beta$ $(0 < \beta < 1)$ controls the degree of upwinding introduced in the second-order differencing scheme of the implicit part.

**1.2. The model implicit upwind schemes.** Here we introduce some notation to analyze the convergence of the iterative implicit scheme applied to the model problem. A *second-order* difference operator of *adjustable upwinding* is employed in the explicit phase

$$D_h = \frac{c}{\Delta x} \delta^{\beta}_{x,2}. \tag{11}$$

Here $\delta^{\beta}_{x,2}$, $(0 \le \beta \le 1)$, combines the second-order backward differencing scheme with the central differencing scheme

$$\delta^{\beta}_{x,2} = \beta \delta^{u}_{x,2} + (1 - \beta) \delta^{c}_{x} . \tag{12}$$

As mentioned before, this discretisation gained considerable popularity since it was introduced in aerodynamics [15]. A *first-order* upwind scheme is applied in the implicit phase

$$M_h = I + \frac{c \Delta t}{\Delta x} \delta^{u}_{x,1} . \tag{13}$$

With these expressions for $M_h$ and $D_h$, the amplification matrix, $G_{\Delta t}$, is completely determined (eq. (7)). When $\Delta t \to \infty$, the amplification matrix $G_{\Delta t}$ approaches

$$G_\infty = I - \left(\delta^{u}_{x,1}\right)^{-1} \delta^{\beta}_{x,2}. \tag{14}$$

Thus, although the timestep is infinite and the problem linear, the amplification matrix is nonzero. Thus, the iteration is not equivalent to Newton's method, and the asymptotic convergence can at best be linear.

**1.3. The defect-correction method.** The iteration (4), derived in the previous section, can be seen as an application of the defect-correction method [2]. In a *defect-correction iteration* the solution $\mathbf{u}^*$ of a linear or nonlinear equation,

$$\Phi_2(\mathbf{u}) = \mathbf{f}, \tag{15}$$

is found by iteration with a simpler, approximate equation for the same problem. For example, let $\Phi_1(\mathbf{u})$ and $\Phi_2(\mathbf{u})$ be a first-order and a second-order discrete approximation, respectively, to the same equation. Then the iterative process starts by first solving

$$\Phi_1(\mathbf{u}^1) = \mathbf{f} \tag{16}$$

for the unknown $\mathbf{u}^1$, and then solving, for $n = 1, 2, \ldots$, the equation

$$\Phi_1(\mathbf{u}^{n+1}) = \Phi_1(\mathbf{u}^n) - \Phi_2(\mathbf{u}^n) + \mathbf{f}. \tag{17}$$

In this way only "simple" equations of the type $\Phi_1(\mathbf{u}^{n+1}) = \mathbf{r}^n$ are solved, and it it is immediate that a fixed point of the iteration yields a solution of (15). Such a construction is common also when a steady problem is accurately approximated by a spectral method denoted by $\Phi_2$ (associated with a full matrix), while an approximation of simpler type, e.g., finite-difference type, denoted by $\Phi_1$ (associated with a band matrix), is introduced to construct a simple iteration.

If the operators $\Phi_1$ and $\Phi_2$ are differentiable, with nonsingular Jacobian matrices $D\Phi_1$ and $D\Phi_2$, then a small error $\mathbf{e}^n = \mathbf{u}^n - \mathbf{u}^*$ approximately satisfies the relation

$$D\Phi_1\, \mathbf{e}^{n+1} = D\Phi_1\, \mathbf{e}^n - D\Phi_2\, \mathbf{e}^n\,. \tag{18}$$

Hence, the linear error amplification operator is given by

$$G_\infty = I - (D\Phi_1)^{-1} D\Phi_2\,. \tag{19}$$

In this way, accurate discrete operators are evaluated only to form right-hand sides in (17). Inversions only involve the simpler first-order discrete operator. In our case, for the steady state (i.e., $\Delta t \to \infty$) we have

$$\Phi_1 = \frac{c}{\Delta x}\delta^u_{x,1}, \quad \Phi_2 = \frac{c}{\Delta x}\delta^\beta_{x,2}, \tag{20}$$

and the identification is obvious. In a similar way, for finite $\Delta t$, the fully implicit method is identified with a defect-correction iteration for which $\Phi_1 = M_h$, $\Phi_2 = \Delta t D_h$.

Defect-correction iteration has interesting implications from the point of view of stability and accuracy. In the linear case, because only the operator $\Phi_1$ is inverted, for a stable approximate solution $\mathbf{u}^n$ $(n = 1, 2, \ldots, )$ only stability of the operator $\Phi_1$ is required (see e.g. [2] or [6, §14.2.2]). This is true for any fixed $n$, but the stability bound degenerates for $n \to \infty$. On the other hand, it is simply verified that if $\Phi_1$ is stable, and if $\Phi_k$ is a $p_k$th-order discretisation of a continuous operator $\Phi$, (k = 1, 2), $p_1 < p_2$, then $\mathbf{u}^n$ is an $O(h^{\min(np_1, p_2)})$ approximation to the true solution of the continuous problem. This implies that *smooth components* in the discretisation error of $\mathbf{u}^1$ converge rapidly. In our case, there is a factor $O(h)$ per iteration step. It is interesting to know that the *high-frequency components*, which are slowly converging to the solution of $\Phi_2(\mathbf{u}) = \mathbf{f}$, are essentially the same error components that give a poor approximation to the continuous problem. This is illustrated by the results of Koren ([9, Fig. 5.7]).

Although these arguments suggest that a small number of iterations is sufficient to obtain accurate results, and that, in general, it is unwise to iterate (17) until convergence is observed, in this paper we study the general convergence behaviour of the above iteration procedure.

## 2. One-dimensional analysis.

### 2.1. Fourier type analysis.

**2.1.1. The interior domain.** In this section we first give the analysis of the defect-correction iteration, neglecting the effect of the boundaries. In many cases this gives a good impression of the convergence behaviour in the initial phase of the iteration. We consider the operators $\delta_{x,1}$ and $\delta_{x,2}^{\beta}$ working on a uniform discretisation of the line $(-\infty, +\infty)$ . Then $\delta_{x,1}$ and $\delta_{x,2}^{\beta}$ can be represented by infinite Toeplitz matrices of the form (in stencil notation)

$$\delta_{x,1} = \text{Trid}\,[-1, 1, 0]\,, \tag{21}$$

and

$$\begin{aligned} \delta_{x,2}^{\beta} &= (1-\beta)\,\text{Trid}\,[-\tfrac{1}{2}, 0, \tfrac{1}{2}] + \beta\,\text{Pentad}\,[\tfrac{1}{2}, -2, \tfrac{3}{2}, 0, 0] \\ &= \tfrac{1}{2}\,\text{Pentad}\,[\beta, -3\beta-1, 3\beta, 1-\beta, 0]. \end{aligned} \tag{22}$$

Any discrete $l^2$-function $u_h$, defined on $(\ldots, -2h, -h, 0, h, 2h, \ldots)$ can be decomposed in its Fourier modes $u_\omega$, with $u_\omega(hj) = e^{i\omega hj}$, by

$$u_h(jh) = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{i\omega hj} F(u_h)(\omega) d\omega, \tag{23}$$

where $F(u_h) \in L^2(-\frac{\pi}{h}, \frac{\pi}{h})$, such that $F(u_h)(\omega) = \frac{h}{\sqrt{2\pi}} \sum_{-\infty}^{\infty} e^{-i\omega hj} u_h(jh)$ is the Fourier transform of $u_h$. It is easily shown that $F(\delta_{x,1}u_h)(\omega) = F(\delta_{x,1})(\omega) F(u_h)(\omega)$, where

$$F(\delta_{x,1})(\omega) = -e^{-i\omega h} + 1 = 2ie^{-i\omega h/2}\sin(\omega h/2)\,. \tag{24}$$

Similarly

$$\begin{aligned} F(\delta_{x,2}^{\beta})(\omega) = 2i\, e^{-i\omega h/2}[&\sin(\omega h/2) + \\ &i\, \cos(\omega h/2)\sin^2(\omega h/2) + (2\beta - 1)\sin^3(\omega h/2)]. \end{aligned} \tag{25}$$

The amplification operator of the defect correction is given by $G_\infty = I - (\delta_{x,1})^{-1}\delta_{x,2}^{\beta}$ and, hence,

$$\begin{aligned} F(G_\infty)(\omega) &= 1 - (F(\delta_{x,1})(\omega))^{-1}\, F(\delta_{x,2}^{\beta})(\omega) \\ &= i\, \sin(\omega h/2)\cos(\omega h/2) + \kappa\sin^2(\omega h/2), \end{aligned} \tag{26}$$

where $\kappa = 1 - 2\beta$. Thus we find

$$\text{sup}_{\omega\in(-\pi/h,\pi/h)}|F(G_\infty)(\omega)| = \text{sup}_{t\in(0,1)}\sqrt{\kappa^2 t^2 + t(1-t)}, \tag{27}$$

and as the upperbound

$$\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)| = \begin{cases} \frac{1}{2}\frac{1}{\sqrt{1-\kappa^2}} & \text{for } \kappa^2 \le 1/2, \\ |\kappa| & \text{for } 1/2 \le \kappa^2 \le 1. \end{cases} \tag{28}$$

Special cases are

$$\begin{aligned} &\text{sup}_{\omega\in(-\pi/h,\pi/h)}|F(G_\infty)(\omega)| = 1 && \text{for } \beta = 0 \quad \text{or} \quad \beta = 1, \\ &\text{sup}_{\omega\in(-\pi/h,\pi/h)}|F(G_\infty)(\omega)| = 1/2 && \text{for } \beta = 1/2, \\ &\text{sup}_{\omega\in(-\pi/h,\pi/h)}|F(G_\infty)(\omega)| = \tfrac{3}{8}\sqrt{2} \approx 0.530 && \text{for } \beta = 1/3. \end{aligned} \tag{29}$$

We can use $\sup_{\omega\in(-\pi/h,\pi/h)}|F(G_\infty)(\omega)|$ as an estimate of the convergence factor of the defect-correction iteration, in the case that there is no significant influence of any of the two boundaries. In Fig. 1 we give a picture of this amplification operator (28), as a function of $\beta$.

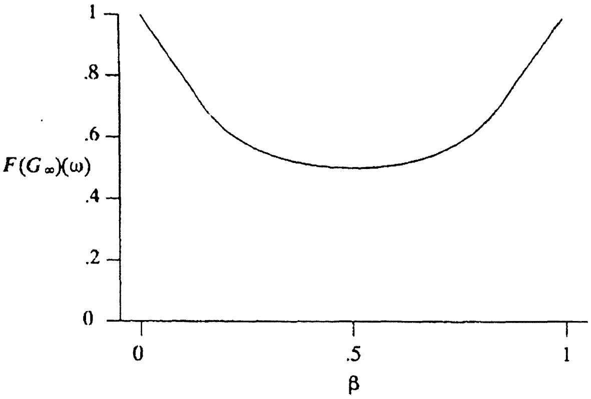


FIG. 1. *Fourier amplification factor (one-dimensional).*

From this analysis we see that convergence can be expected for $\beta\in(0, 1)$. However, for $|1 - 2\beta|^2 > 1/2$ convergence can be slow, and the high frequencies, for which $\sin^2(\omega h/2) \approx 1$, are the slowly damped components responsible for this behaviour.

**2.1.2. The influence of the boundary.** In general, computations are made in a bounded domain and the influence of the inflow boundary cannot be neglected. Therefore the above Fourier analysis can only be of limited value. Nevertheless, as we shall see in §2.3, in many cases the results give a reasonable impression of the iterative behaviour in the initial phase of the iteration. As can be expected, this initial phase of the iteration takes longer if the number of points in the domain gets larger.

To obtain an impression of the influence of the inflow Dirichlet boundary, we consider grid functions on a uniform partition $\{x_i = ih,\ i = 0, 1, 2, \dots,\}$ of the half-line $[0, \infty)$ and we restrict ourselves to error components that vanish for large $x_i$. The operators $\delta_{x,1}$ and $\delta^{\beta}_{x,2}$ are again described by (21), (22), except for the first two equations in the system, which are determined by the boundary discretisation.

The amplification operator $G_\infty$ of the defect-correction iteration is given by (14) and we are interested in the behaviour of its eigenvalues. The eigenfunctions $u_\lambda$ of $G_\infty$ and the corresponding eigenvalues $\lambda$ satisfy the relation

$$\delta^{\beta}_{x,2}\, u_\lambda = (1-\lambda)\,\delta_{x,1}\, u_\lambda\,, \tag{30}$$

and from (21), (22) it follows that $u_\lambda$ has the form

$$u_\lambda(jh) = A_0 + A_1\mu_1^j + A_2\mu_2^j\,, \tag{31}$$

where $\mu_1$ and $\mu_2$ are roots of the equation

$$(1-\beta)\mu^2 + (2\beta + 2\lambda - 1)\mu - \beta = 0\,. \tag{32}$$

The constants $A_i$, $i = 0, 1, 2$ are determined by the boundary condition and the two equations used near the boundary. Since we are only interested in real errors that lie in the neighbourhood of the boundary, the relevant eigenfunctions are restricted to those with $|\mu| \le 1$. Further, because the eigenfunctions should be real, we see from (31) that either $\mu_1, \mu_2 \in \mathbb{R}$ or $|\mu_1| = |\mu_2|, \mu_1 \neq \mu_2$ if $\mu_1, \mu_2 \in \mathbb{C}$.

In the case of real $\mu$, only those $\lambda \in \mathbb{R}$ for which $|\mu_1| \leq 1$ and $|\mu_2| \leq 1$ are acceptable. Eigenfunctions vanishing at infinity (i.e., as $x_i \to \infty$) are such that $A_0 = 0$, and the eigenfunctions satisfy the Dirichlet condition $u_\lambda(x_0) = 0$ if, in addition, $A_1 + A_2 = 0$ holds. This combines to

$$u_\lambda(x_k) = A_1\mu_1^k - A_1\mu_2^k, \quad k = 0, 1, 2, \ldots, \tag{33}$$

which implies

$$u_\lambda(x_2)/u_\lambda(x_1) = \mu_1 + \mu_2 = -(2\beta + 2\lambda - 1)/(1 - \beta) \in \mathbb{R}\,. \tag{34}$$

However, this leads to a contradiction because the discretisation near the boundary requires

$$(1 - \beta)\, u_\lambda(x_2) = (2(1 - \lambda) - 2\beta)\, u_\lambda(x_1)\,. \tag{35}$$

This implies that real $\lambda$ are only possible for $\lambda = 1/2 - \beta$.

In the case of complex $\mu$ we have $|\mu_1| = |\mu_2|, \mu_1 \neq \mu_2$. This implies $\mu_2 = \mu_1 e^{2i\theta}$, $\theta \neq 0 \bmod(\pi)$. Now, for $0 < \beta < 1$ we know $-\beta/(1-\beta) = \mu_1\mu_2 = \mu_1^2\, e^{2i\theta}$, and we obtain $\mu_{1,2} = i\sqrt{\beta/(1-\beta)}\, e^{\pm i\theta}$. Further, the relation $\mu_1 + \mu_2 = (1 - 2\beta - 2\lambda)/(1 - \beta)$ yields the following expression for the eigenvalue:

$$\lambda = 1/2 - \beta \pm i\sqrt{\beta(1-\beta)}\, \cos(\theta), \quad \theta \neq 0 \bmod(\pi). \tag{36}$$

This expression describes the location of the eigenvalues in the complex plane. It follows that $|\lambda| \leq \sqrt{(1/2 - \beta)^2 + \beta(1-\beta)} = 1/2$, which is a bound for the spectral radius, independent of $\beta$. As we shall see in the next section, a similar result is obtained for a finite interval, provided the matrix models account for the proper left-boundary specification and assume backward differencing at the right boundary.

### 2.2. Matrix analysis.

**2.2.1. Standard schemes.** For $0 < \beta < 1$, the eigenvalues of the amplification operator for the discrete system with $N + 1$ nodal points ($x_i = ih$; $i = 0, 1, \ldots, N$) approximating the model problem (2) are similar to those obtained in (36), provided backward differencing is used at the right boundary:

$$\left\{\begin{array}{l} \lambda_0 = 0, \\ \lambda_m = \frac{1}{2} - \beta + i\sqrt{\beta(1-\beta)}\cos\frac{m\pi}{N}, \quad (m = 1, 2, \ldots, N-1). \end{array}\right. \tag{37}$$

A diagram of these eigenvalues, in which $\beta$ is a parameter, is given in Fig. 2. The eigenvectors can be expressed explicitly, each one being a simple function of the corresponding eigenvalue; the matrix $G_\infty$ is diagonalisable[3] and the convergence is (immediately) dissipative, at a rate slightly more rapid than that of the sequence $2^{-n}$, since the spectral radius is given by

$$\begin{aligned} \rho(\beta) &= |\lambda_1(\beta)| \\ &= \tfrac{1}{2}\sqrt{1 - 4\beta(1-\beta)\sin^2\tfrac{\pi}{N}} < \tfrac{1}{2}, \quad \text{and hence} \approx \tfrac{1}{2}. \end{aligned} \tag{38}$$

This result is remarkable since it implies that the iterative convergence rate is independent of meshsize. The best separation of the eigenvalues and the best condition for the system of eigenvectors are realised by the *Fromm* scheme ($\beta = 1/2$).

[3] For $\beta = 1/2$ we are discarding the case where $N$ is even, for which the eigenvalue $\lambda = 0$ is double and the matrix defective; however, this has no severe effect on the convergence rate since only one eigenvector is missing, as it will be explained in the next section.

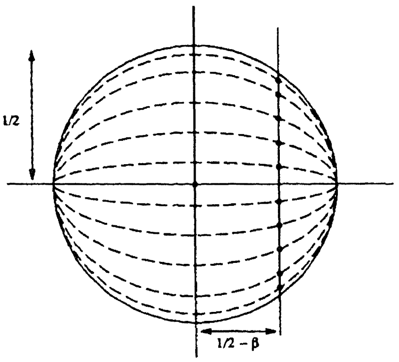


FIG. 2. *Locus of the eigenvalues of the amplification operator in the complex plane relative to the circle of radius* 1/2; *for the one-dimensional nonperiodic model problem and infinite timestep with the second-order β-upwind scheme and, for preconditioning, the first-order upwind scheme.*

**2.2.2. Pathological schemes.** We start this section by examining the application of the simple explicit first-order upwind scheme to (2):

$$u_j^{n+1} = u_j^n - c\Delta t \frac{u_j^n - u_{j-1}^n}{\Delta x}.$$

With a Courant number, $\nu = c\Delta t/\Delta x$, equal to 1, the method reduces to the method of characteristics (which is exact): $u_j^{n+1} = u_{j-1}^n$. If we let $\mathbf{u}^n = \left(u_1^n, u_2^n, \ldots, u_N^n\right)^T$, where $N$ denotes the number of mesh intervals, we have

$$\mathbf{u}^{n+1} = G\mathbf{u}^n + \mathbf{b}, \tag{39}$$

where

$$G = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{40}$$

The amplification matrix $G$ is defective, and although the spectral radius $\rho = 0$, the steady-state solution is not found in one iteration only, but in $N - 1$ iterations. The matrix $G$ is a Jordan block of order $N \times N$; $N - 1$ eigenvectors are missing, and the convergence process begins with a phase of *transfer* of the components of the error-vector, $\mathbf{u}^n - \mathbf{u}^\infty$ ($\mathbf{u}^\infty$ denotes the steady-state solution), *from one generalised eigenvector to the next.* This phase extends over $N - 1$ iterations. Only then, after the error content is "flipped" into the only true eigenvector, does the dissipative phase begin.[4]

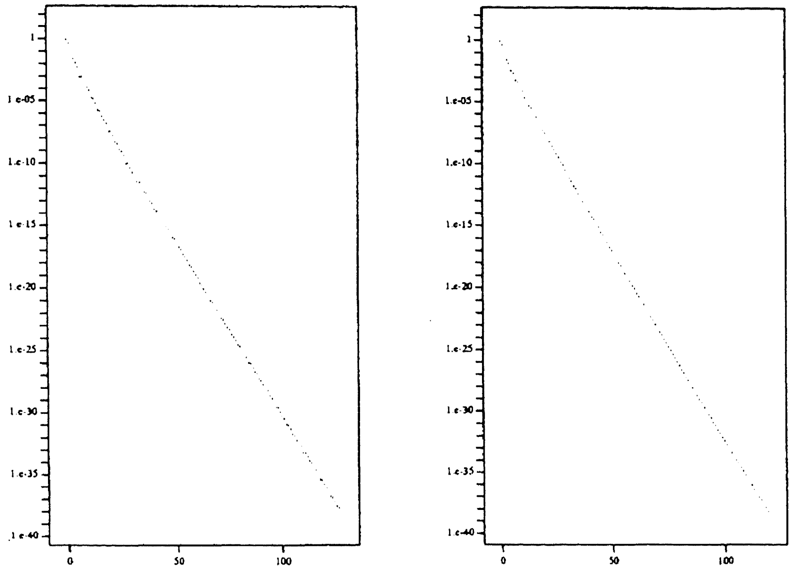
The pattern of convergence that exhibits a phase of significant extent during which the norm of the residual (expressed in the basis of the generalised eigenvectors) is not reduced can be observed any time the iteration is defective and the number of missing eigenvectors is large. In an analogy to the convergence of the simple explicit method, we refer to this phase as one of *pseudo-* or *artificial convection.*

[4] Here this phase reduces to immediate annihilation (in one iteration) since $\rho = 0$.

In particular, this is the case for the implicit methods under study when the timestep is infinite and the upwinding parameter $\beta$ is set to either limit 0 or 1. To see this, return to Fig. 2. In either limit, the spectral radius is equal to $1/2$. However, $N-1$ eigenvalues are identical and the corresponding eigenvectors coalesce because they all can be expressed in closed-form as $\mathbf{v}^m = \chi(\lambda_m)$ for the same known vector-valued function $\chi(\lambda)$. Hence the matrix is defective, and the number of missing eigenvectors, $N-2$, is large. Consequently *the pseudo-convection phase extends over a number of iterations equivalent to, for N large,* $N/(1-\rho) = 2N$ (for a detailed analysis see [5]). Here, the propagation speed has no physical meaning, the wave travels one mesh interval every two iterations, and the Courant number is infinite. The phenomenon is therefore a *numerical pathology*.

### 2.3. Numerical experiments.

**2.3.1. Iteration with a regular scheme.** To illustrate the above results, we show some experiments made for the simple linear model problem (16), (17), (20). In Fig. 3 we show results for iterations applied with $\beta = 1/3$ or $\beta = 1/2$ on a mesh with 100 intervals. We notice that the asymptotic rate is $\rho = \lim_{n\to\infty} \rho_n = 0.5$, where $\rho_n = \|e_n\|_\infty / \|e_{n-1}\|_\infty$ and $e_n$ is the error after $n$ iterations. For these examples all components of the initial error $e_0$ were chosen randomly from the interval (0, 1) with a uniform distribution. The convergence rate corresponds to what is expected from the analysis.



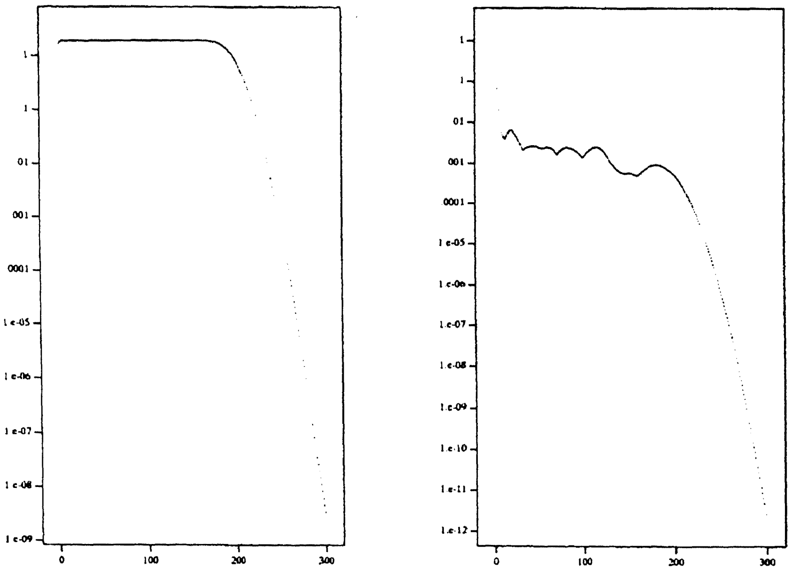
Case a: $N = 100$, $\beta = 1/3$, random initial error.

Case b: $N = 100$, $\beta = 1/2$, random initial error.

FIG. 3. *Convergence history of standard methods.*

**2.3.2. Iteration with the central scheme.** In Fig. 4 we show results for an iteration with $\beta = 0$. The discretisation is on $N = 100$ nodes, and an oscillating or random initial error is used. The *oscillating initial error* $e_0$ is defined by the element values $e_{0,i} = (-1)^i$, $i = 1, 2, \ldots,$; in the *random initial error* the error at all nodes is randomly chosen, uniformly distributed in the interval (0, 1).

The oscillating error, for which $\sin(\omega h/2) \approx 1$, is the most persistent error component: the convergence factor is approximately 1.0 for the first $2N$ iteration steps. Only after the



Case a: $N = 100$, $\beta = 0$, oscillating initial error.

Case b: $N = 100$, $\beta = 0$, random initial error.

FIG. 4. *Convergence histories of iteration with central scheme.*

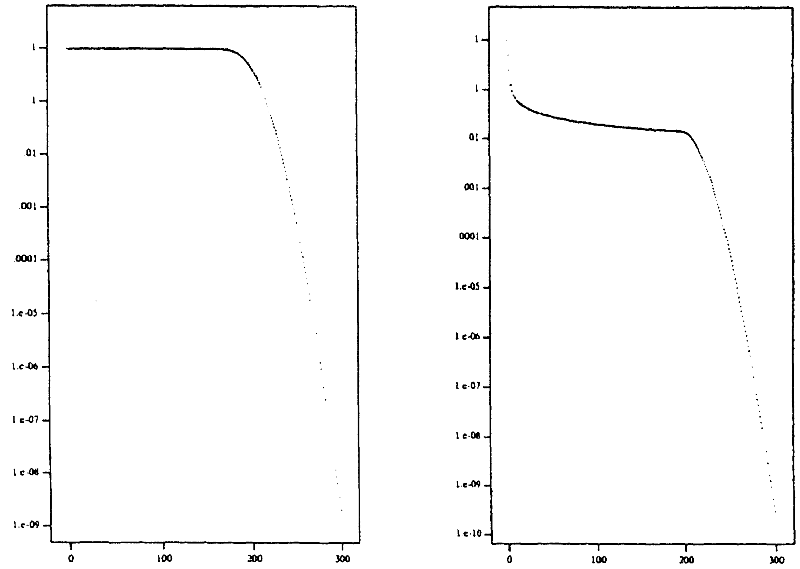
$2N$th iteration does the spectral radius of the amplification operator start to determine the convergence rate. The asymptotic convergence rate is again 0.5. This is in agreement with the theoretical findings.

**2.3.3. Iteration with the fully upwind scheme.** For the fully upwind scheme ($\beta = 1$), we observe in Fig. 5 a behaviour very much similar to that for the central scheme ($\beta = 0$).

As an example of the evolution of the error during the iteration process, in Fig. 6 we show the behaviour of an initially oscillating error for $\beta = 0$, $\beta = 1/2$, and $\beta = 1$ on $N = 10$ nodes. The Dirichlet boundary condition was taken at the left-hand side. We see that for $\beta = 0$ or 1 it takes $2N$ iterations before the error has moved out of the domain. For $\beta = 0$ the error moves to the left, for $\beta = 1$ to the right. Further, for $\beta = 1$ we see that the error changes sign at each iteration, which can be related to the corresponding eigenvalue being $-1$.

**2.3.4. Iteration with a near-pathological scheme.** In Fig. 7 we show the convergence of the defect-correction iteration for different values of $\beta$ that are close to either $\beta = 0$ or $\beta = 1$. For these near-pathological cases, we clearly distinguish the pseudo-convection phase, in which the convergence rate $\rho_n = |1 - 2\beta|$ is predicted by Fourier analysis. After $2N$ iterations, the convergence behaviour is dominated by the asymptotic convergence rate 1/2.

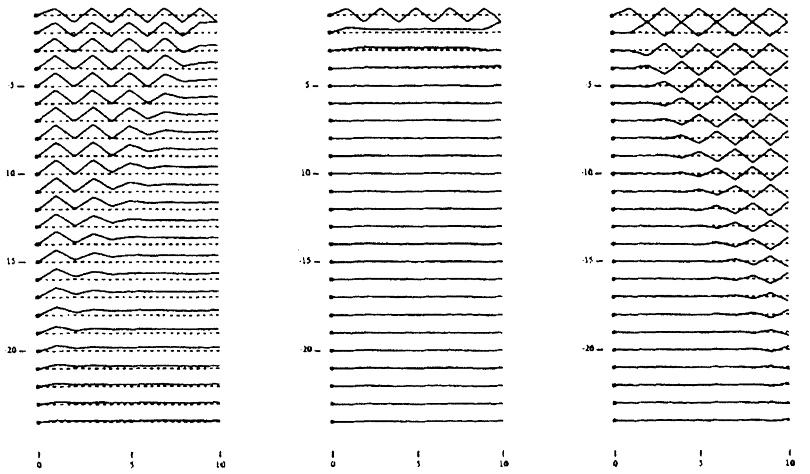
**Summary.** For the one-dimensional problem we distinguish different phases in the convergence of the iterated defect correction. Generally, we first observe an impulsive start, where all components corresponding to small eigenvalues are damped. For the regular schemes ($\beta$ different from 0 or 1) soon an asymptotic rate of 1/2 is obtained. For the (near) pathological cases ($\beta$ close to 0 or 1), after the impulsive start, we distinguish first a Fourier (or pseudo-convection) phase for about $2N$ iterations, in which the convergence is described by the Fourier analysis. After $2N$ iterations the asymptotic rate 1/2 is found. In the real degenerate cases ($\beta = 0$ or $\beta = 1$) we recognise a Fourier (pseudo-convection) phase, where the error does



Case a: $N = 100$, $\beta = 1$, oscillating initial error.

Case b: $N = 100$, $\beta = 1$, random initial error.

FIG. 5. *Convergence histories of iteration with fully upwind scheme.*



Case a: $\beta = 0$

Case b: $\beta = 1/2$

Case c: $\beta = 1$

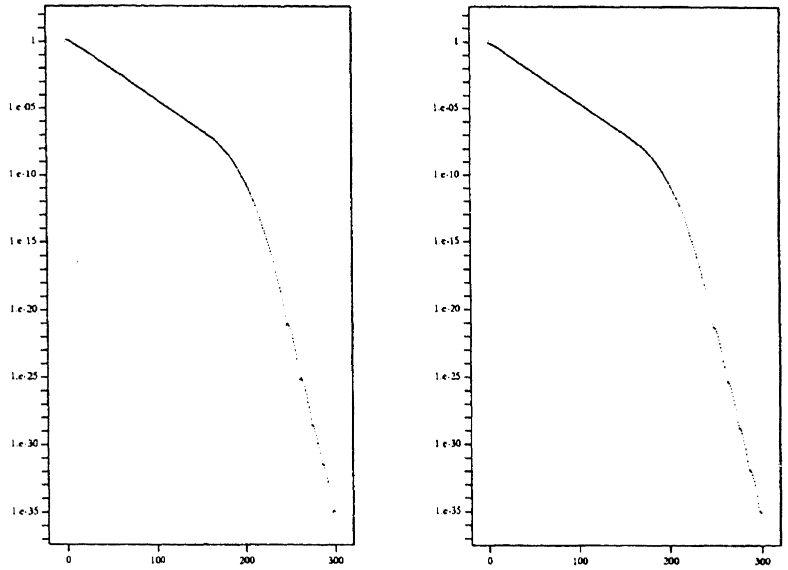
FIG. 6. *Evolution of the error on a mesh with* $N = 10$ *points. The solutions are shown after i iteration cycles* ($i = 1, \ldots, 24$, *from top to bottom*).

not decrease for $2N$ iterations, and the logarithmic asymptotic rate is due to the large Jordan block in the eigenvalue decomposition.

## 3. Two-dimensional analysis.

**3.1. Fourier analysis.** Analogous to the treatment of the one-dimensional problem, here we give the Fourier analysis for the discretisation of a steady problem of the form

$$u_t + au_x + bu_y = f\,, \tag{41}$$



Case a: $N = 100$, $\beta = 0.05$,    Case d: $N = 100$, $\beta = 0.95$,

FIG. 7. *Convergence histories of near-pathological methods. Oscillating initial error.*

in which $a > 0, b > 0$. Because we are only interested in steady solutions, the time derivative solely serves to define the "direction" of the flow, described by $u$. The stencils for the discrete central and upwind operators are given by

$$\delta_1^u = \left[\begin{array}{ccc} & 0 & \\ -a & a+b & 0 \\ & -b & \end{array}\right], \tag{42}$$

$$\delta_2^c = \frac{1}{2}\left[\begin{array}{ccc} & b & \\ -a & 0 & a \\ & -b & \end{array}\right], \tag{43}$$

$$\delta_2^u = \frac{1}{2}\left[\begin{array}{ccccc} & & 0 & & \\ & & 0 & & \\ a & -4a & 3(a+b) & 0 & 0 \\ & & -4b & & \\ & & b & & \end{array}\right]. \tag{44}$$

Similar to the previous section, we define

$$\delta_2^\beta = \frac{1}{2}\left[\begin{array}{ccccc} & & 0 & & \\ & & (1-\beta)b & & \\ \beta a & -(1+3\beta)a & 3\beta(a+b) & (1-\beta)a & 0 \\ & & -(1+3\beta)b & & \\ & & \beta b & & \end{array}\right]. \tag{45}$$

Without loss of generality we may take $a + b = 1$.

The Fourier transforms of these difference operators are introduced, analogous to the one-dimensional case. If we define the Fourier modes by $u_\omega(hj) = e^{i(\omega_1 h_1 j_1 + \omega_2 h_2 j_2)}$, where the subscripts 1 and 2 refer to the $x$- and $y$-directions, respectively, we find

$$F(\delta_1) = 2ia\ e^{-i\omega_1 h_1/2}\sin(\omega_1 h_1/2) + 2ib\ e^{-i\omega_2 h_2/2}\sin(\omega_2 h_2/2) \tag{46}$$

and

$$\begin{aligned} F(\delta_2^\beta) = \quad & 2iae^{-i\omega_1 h_1/2} S_1(C_1^2 + iS_1C_1 + 2\beta S_1^2) \\ & +2ibe^{-i\omega_2 h_2/2} S_2(C_2^2 + iS_2C_2 + 2\beta S_2^2)\ , \end{aligned} \tag{47}$$

where, for brevity, we have used $S_1 = \sin(\omega_1 h_1/2)$, $S_2 = \sin(\omega_2 h_2/2)$, $C_1 = \cos(\omega_1 h_1/2)$, and $C_2 = \cos(\omega_2 h_2/2)$, and for symmetry, $a_1 = a$, $a_2 = b$, $h_1 = \Delta x$, $h_2 = \Delta y$, so that

$$\frac{F(\delta_1) - F(\delta_2^\beta)}{F(\delta_1)} = \frac{a_1 e^{-i\omega_1 h_1/2} S_1^2[\kappa S_1 - iC_1] + a_2 e^{-i\omega_2 h_2/2} S_2^2[\kappa S_2 - iC_2]}{a_1 e^{-i\omega_1 h_1/2} S_1 + a_2 e^{-i\omega_2 h_2/2} S_2}, \tag{48}$$

where $\kappa = 1 - 2\beta$. As the amplification factor we find

$$g(\omega) = \left\| \frac{F(\delta_1) - F(\delta_2^\beta)}{F(\delta_1)} \right\| = \sqrt{\frac{(a_1 S_1^2(1-2\beta S_1^2) + a_2 S_2^2(1-2\beta S_2^2))^2 + 4\beta^2(a_1 S_1^3 C_1 + a_2 S_2^3 C_2)^2}{(a_1 S_1^2 + a_2 S_2^2)^2 + (a_1 S_1 C_1 + a_2 S_2 C_2)^2}}. \tag{49}$$

This expression can be used to determine the convergence rate for the separate modes.

In the neighbourhood of the origin, for small $\omega_1$ and $\omega_2$, we can set $S_1 \approx (\omega_1 h_1/2)$, $S_2 \approx (\omega_2 h_2/2)$, $C_1 \approx 1$, $C_2 \approx 1$ and obtain

$$g(\omega) \approx \sqrt{\frac{(a_1\omega_1^2 h_1^2 + a_2\omega_2^2 h_2^2)^2}{(a_1\omega_1^2 h_1^2 + a_2\omega_2^2 h_2^2)^2 + 4(a_1\omega_1 h_1^2 + a_2\omega_2 h_2^2)^2}}\ . \tag{50}$$

From this expression we see that the amplification factor becomes one in the neighbourhood of the origin where $a_1\omega_1 h_1^2 + a_2\omega_2 h_2^2 = 0$. Therefore, introducing new coordinates $z = (a_1\omega_1 h_1 + a_2\omega_2 h_2)/2$ and $w = \sqrt{a_1 a_2}\ (\omega_1 h_1 - \omega_2 h_2)/2$, we obtain in the neighbourhood of the origin

$$g(\omega) \approx \sqrt{\frac{(z^2 + w^2)^2}{(z^2 + w^2)^2 + z^2}}\ . \tag{51}$$

This implies that the level curves for $g(\omega)$ are a family of circles through the origin that are all tangent to the line $z = 0$ (see Fig. 8). This means that the origin is a singular point for the function $g(\omega)$, and

$$\lim_{(\omega \to 0, a_1\omega_1 + a_2\omega_2 = 0)} g(\omega) = 1, \tag{52}$$

and

$$\lim_{(\omega \to 0, a_1\omega_1 + a_2\omega_2 = c \neq 0)} g(\omega) = 0. \tag{53}$$

This shows that for the hyperbolic problem there are always low-frequency modes $u_\omega$ for which $g(\omega) = 1$. These modes, associated with low frequencies $\omega$ for which $a_1\omega_1 + a_2\omega_2 \approx 0$, are related with functions that are constant in the characteristic direction of the hyperbolic equation. Such modes form the null-space of the differential operator, and the corresponding solution components are determined by the boundary condition. The zero eigenvalue for these eigenmodes is inherited (to some order of accuracy) by all consistent discrete operators (hence also by $\delta_1$ and $\delta_2^\beta$). Consequently, to quantify the convergence behaviour in the two-dimensional case, we cannot use $\sup_{\omega,\omega\neq 0} g(\omega)$. This is a fundamental difference between the one- and the two-dimensional cases. For further remarks associated with this problem see [3]. Nevertheless, the function $g(\omega)$, of which level curves are shown in Fig. 8, gives a good qualitative impression about the convergence behaviour. In Fig. 8 we show $g(\omega)$ for some special cases. We take a convection direction of 45°, $h_1 a_1 = h_2 a_2 = 1$, and $\beta = 0, \frac{1}{2}$, or 1. It is observed that for $\beta = 0$ only low-frequency modes that are perpendicular to the characteristic modes are damped. For $\beta = 1$ we observe many high-frequency modes that will not converge, and for $\beta = 1/2$ we see that all modes converge except the low-frequency characteristic modes.

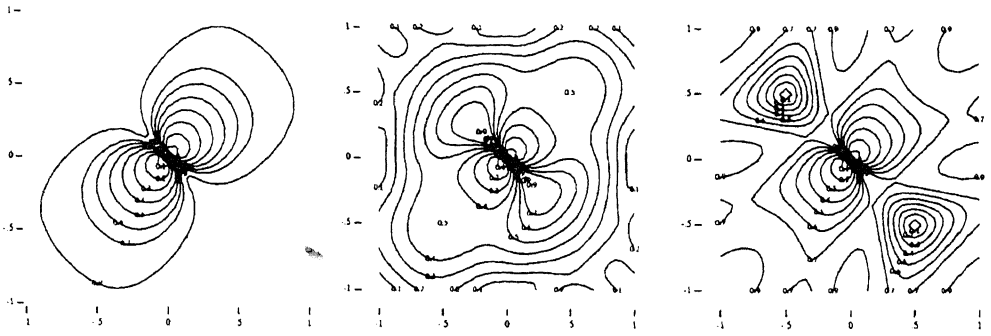

Case a: $\beta = 0$ Case b: $\beta = 1/2$ Case c: $\beta = 1$

FIG. 8. *Level curves of Fourier amplification factor* ($a = b$).

### 3.2. Matrix analysis.

**3.2.1. General remarks.** To a certain extent, the one-dimensional matrix analysis can be extended to two dimensions in the case of the following model problem:

$$(54)\qquad \begin{cases} u_t + a\,u_x + b\,u_y = 0, & a > 0,\ b > 0;\ x, y, t \geq 0; \\ u(x, y, 0) = u^0(x, y), & \\ u(0, y, t) = u(x, 0, t) = 0 & \forall x, y, t \geq 0\,. \end{cases}$$

Assuming a mesh of $N_x \times N_y$ gridpoints, and employing the Kronecker product notation, a finite-difference analogue of the operator

$$(55)\qquad \mathcal{D} = a\,\frac{\partial}{\partial x} + b\,\frac{\partial}{\partial y}$$

can be represented by an $N_x N_y \times N_x N_y$ matrix having the following Kronecker sum structure:

$$(56)\qquad \mathcal{D}_h = \mathcal{D}_x \oplus \mathcal{D}_y,$$

in which $\mathcal{D}_x$ and $\mathcal{D}_y$ are $N_x \times N_x$ and $N_y \times N_y$ matrix representations of finite-difference analogues of the operators $a\,\frac{\partial}{\partial x}$ and $b\,\frac{\partial}{\partial y}$, respectively (assuming specified data along the axes). In particular, if first-order backward differences are employed we can form

$$(57)\qquad \mathcal{D}_{h,1} = (\nu_x\,\delta^u_{x,1}) \oplus (\nu_y\,\delta^u_{y,1})$$

($\nu_x = a/\Delta x$ , $\nu_y = b/\Delta y$), whereas second-order partially upwind differences yield

$$\mathcal{D}_{h,2} = (\nu_x\, \delta^{\beta}_{x,2}) \oplus (\nu_y\, \delta^{\beta}_{y,2}) \tag{58}$$

if the same value of the upwinding parameter $\beta$ is used in the two directions. From these definitions the expression of the amplification matrix $G_\infty$ follows:

$$G_\infty = I - \left(\mathcal{D}_{h,1}\right)^{-1} \mathcal{D}_{h,2}. \tag{59}$$

Again we are led to examine the eigenproblem associated with the matrix $G_\infty$. One seeks $\lambda \in \mathbb{C}$ such that there exists a nonzero vector $u \in \mathbb{C}^N$ ($N = N_x N_y$) satisfying

$$\mathcal{D}_{h,2}\, u = (1 - \lambda)\, \mathcal{D}_{h,1}\, u. \tag{60}$$

Replacing these operators by their respective expressions as Kronecker sums and rearranging yields

$$\left(A_x(\lambda) \oplus A_y(\lambda)\right) u = 0, \tag{61}$$

where

$$\begin{aligned} A_x(\lambda) &= \nu_x \left(\delta^{\beta}_{x,2} + (\lambda - 1)\, \delta^{u}_{x,1}\right), \\ A_y(\lambda) &= \nu_y \left(\delta^{\beta}_{y,2} + (\lambda - 1)\, \delta^{u}_{y,1}\right). \end{aligned} \tag{62}$$

We were not able to solve this (generalised) eigenproblem analytically; however, several conclusions can readily be drawn.

*Remark* 1. Let $N_x/N_y = p/q$ ($p < N_x$ and $q < N_y$) and $\beta$ be arbitrary. This is the case in particular if $N_x = N_y$ and $p = q = 1, 2, \ldots, N_x - 1$. According to (37),

$$\lambda_{x,p} = \lambda_{y,q} = \lambda \tag{63}$$

if $\lambda_{x,p}$ and $\lambda_{y,q}$ are, respectively, the $p$th and $q$th eigenvalues of one-dimensional problems defined over meshes of $N_x$ and $N_y$ gridpoints. Then let $u_x$ and $u_y$ be $N_x \times 1$ and $N_y \times 1$ associated eigenvectors, so that

$$\begin{aligned} A_x(\lambda)\, u_x &= 0, \\ A_y(\lambda)\, u_y &= 0; \end{aligned} \tag{64}$$

then (61) holds for $u = u_x \otimes u_y$. This proves that any eigenvalue common to the $x$- and $y$-associated one-dimensional problems is also an eigenvalue of the two-dimensional problem.

In the mesh-refinement limit $N_x, N_y \to \infty$, the spectra of these one-dimensional problems identify to the same continuum (since $\beta$ is assumed to be the same in both directions), and all the eigenvalues of the one-dimensional problems can be considered common to the $x$ and $y$ directions; thus, they also are eigenvalues of the two-dimensional problem. Hence, when $N_x, N_y \to \infty$, the spectral radius of the iteration matrix $G_\infty$ is greater or equal to the one-dimensional value for the same $\beta$ (given by (38)):

$$\forall \beta, \quad \rho_{2D}(\beta) \geq \rho(\beta). \tag{65}$$

One can expect that for some values of $\beta$, the critical eigenvector is the discrete form in two dimensions of the highest-frequency mode, and that it is the tensor product of the highest-frequency modes of the associated one-dimensional problems. If so, the eigenvalue and the spectral radius assume the same values as in one dimension. In fact, we will observe by numerical experiment that when $1/2 \leq \beta \leq 1$ the equality sign holds in (65). Before this, we observe the following fact.

*Remark* 2. With varying $\lambda$, let the eigenvalues of the matrices $A_x(\lambda)$ and $A_y(\lambda)$ be denoted by $\big(\alpha_x(\lambda)\big)_j$, $j = 1, 2, \ldots, N_x$, and $\big(\alpha_y(\lambda)\big)_k$, $k = 1, 2, \ldots, N_y$, respectively. The eigenvalues of the Kronecker sum $A_x(\lambda) \oplus A_y(\lambda)$ are the numbers

$$(66) \qquad \big(\alpha_{x\oplus y}(\lambda)\big)_{j,k} = \big(\alpha_x(\lambda)\big)_j + \big(\alpha_y(\lambda)\big)_k$$

for all possible couples $(j, k)$. Therefore, the eigenvalues of the two-dimensional problem are the solutions of the following set of equations:

$$(67) \qquad \big(\alpha_x(\lambda)\big)_j + \big(\alpha_y(\lambda)\big)_k = 0, \qquad j = 1, 2, \ldots, N_x\,,\ k = 1, 2, \ldots, N_y.$$

This result allows us to treat the case $\beta = 1$ analytically.

*Remark* 3. When $\beta = 1$, the matrices $A_x(\lambda)$ and $A_y(\lambda)$ are lower triangular and defective. The corresponding eigenvalues are directly found in the main diagonal

$$(68) \qquad \begin{aligned} &\big(\alpha_x(\lambda)\big)_1 = \lambda\,,\ \big(\alpha_x(\lambda)\big)_2 = \big(\alpha_x(\lambda)\big)_3 = \cdots = \big(\alpha_x(\lambda)\big)_{N_x} = \lambda + \tfrac{1}{2}\,,\\ &\big(\alpha_y(\lambda)\big)_1 = \lambda\,,\ \big(\alpha_y(\lambda)\big)_2 = \big(\alpha_y(\lambda)\big)_3 = \cdots = \big(\alpha_y(\lambda)\big)_{N_y} = \lambda + \tfrac{1}{2}\,. \end{aligned}$$

This gives the following equations for $\lambda$:

$$(69) \qquad \begin{cases} \lambda + \lambda = 0 & \text{once,}\\ \lambda + (\lambda + \tfrac{1}{2}) = 0 & (N_x - 1) + (N_y - 1) \quad \text{times,}\\ (\lambda + \tfrac{1}{2}) + (\lambda + \tfrac{1}{2}) = 0 & (N_x - 1)(N_y - 1) \quad \text{times.} \end{cases}$$

Thus one finds only three distinct eigenvalues: $\lambda_1 = 0$ (simple), $\lambda_2 = -1/4$ (multiplicity $N_x + N_y - 2$), and $\lambda_3 = -1/2$ (multiplicity $(N_x - 1)(N_y - 1)$). Consequently, the spectral radius is

$$(70) \qquad \rho_{2D}(1) = \rho(1) = \frac{1}{2},$$

as in one dimension.

We now return to the general case ($\beta \neq 1$). Since we were not able to obtain an analytic expression for the eigenvalues, we instead computed them numerically for different combinations of the parameters $(N_x, N_y)$, $(\nu_x, \nu_y)$, and $\beta$.

The results are reported in Fig. 9, where the locus of the eigenvalues of the amplification matrix $G_\infty$ is represented in the complex plane relatively to the circle of radius 1/2, assuming a 9 × 9 mesh, for increasing values of the upwinding parameter $\beta$ and convection directions corresponding to $\nu_x = \nu_y$ and $\nu_x = 100\nu_y$.

Recall that for the one-dimensional model problem, all the eigenvalues other than 0 lie on the chord of the circle parallel to the imaginary axis at the abscissa corresponding to $\Re(\lambda) = 1/2 - \beta$. For $N_x = N_y$, these complex numbers also are eigenvalues of the two-dimensional model problem, but some other eigenvalues appear, forming a cloud. The chord reduces to a point when $\beta = 0$ or 1.

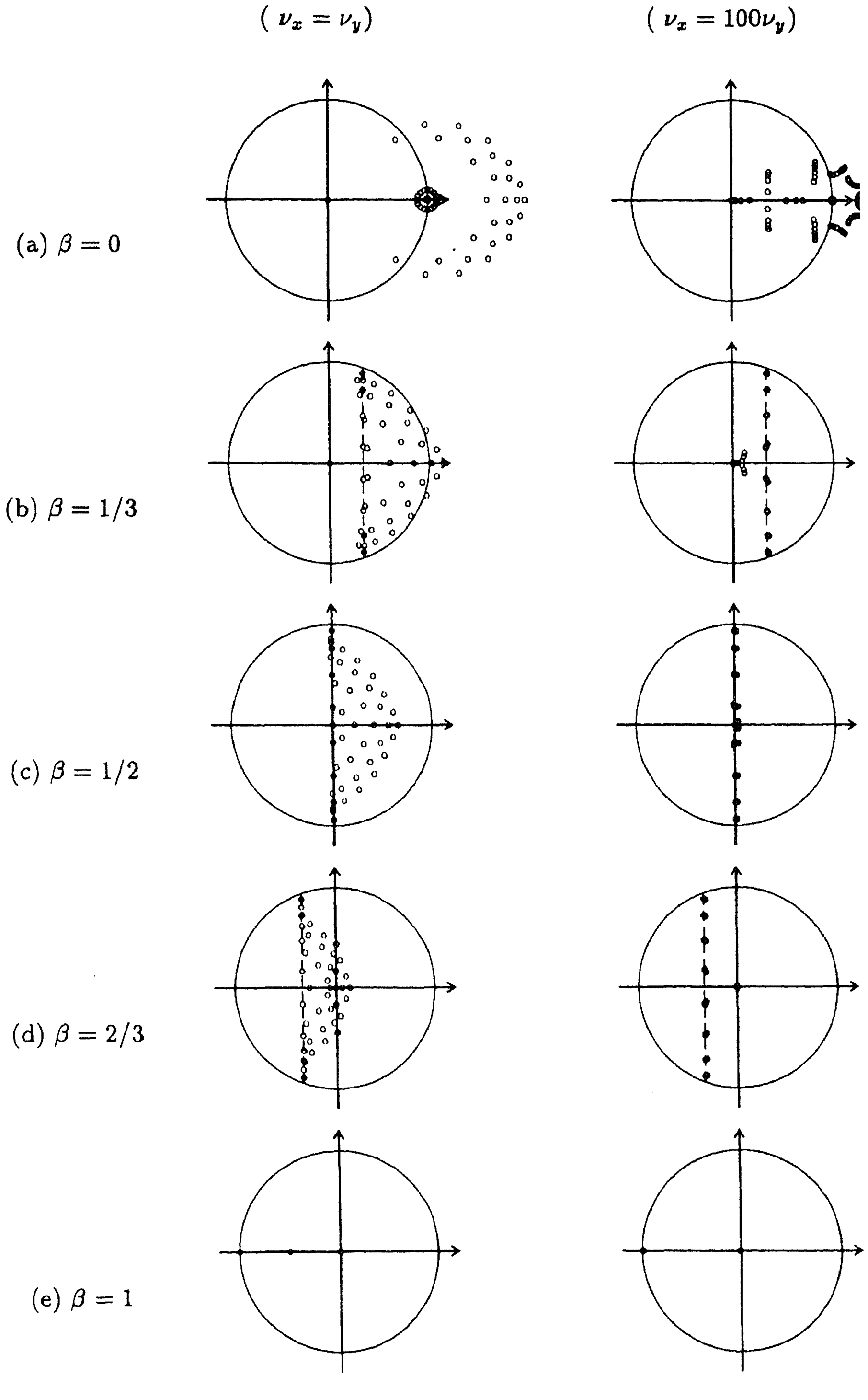


FIG. 9. *Locus of the eigenvalues of the amplification matrix $G_\infty$ of the two-dimensional model problem relative to the circle of radius* 1/2.

In the case most different from the one-dimensional case, a convection direction of 45$^o$ with the grid ($\nu_x = \nu_y$), the cloud of new eigenvalues appears around the one-dimensional eigenvalue spectrum mostly to the right, and it shifts with it to the left as $\beta$ increases. For small values of $\beta$, the eigenvalue of largest modulus is real positive and exterior to the disk of radius 1/2. For $\beta$ larger than some value between 1/3 and 1/2, the cloud lies entirely in the disk; for some $\bar{\beta} < 1/2$, the eigenvalue of largest modulus is, for all $\beta \geq \bar{\beta}$, that of the largest imaginary part, and since it belongs to the one-dimensional spectrum, the spectral radius assumes the same value (less than and close to 1/2) as in one dimension. For $\beta = 1$, as previously established, only three distinct eigenvalues are found: 0, $-1/4$, and $-1/2$. Finally we observe that in contrast to the one-dimensional case, the eigenvalue spectra of two schemes defined by values of $\beta$ symmetrical with respect to 1/2 are not symmetrical with respect to the origin, and the corresponding spectral radii are different.

For the cases shown on the right side of Fig. 9, ($\nu_x = 100\nu_y$), convection in the $x$ direction dominates convection in the $y$ direction. As a result, with $\epsilon = \nu_y/\nu_x$,

$$(71)\qquad \begin{aligned} G_\infty &= I_x \otimes I_y - \left(\delta^u_{1,x} \otimes I_y + \epsilon I_x \otimes \delta^u_{1,y}\right)^{-1} \left(\delta^\beta_{2,x} \otimes I_y + \epsilon I_x \otimes \delta^\beta_{2,y}\right) \\ &= (G_\infty)_x \otimes I_y + O(\epsilon)\,. \end{aligned}$$

Hence, the two-dimensional algebraic problem is close to the repetition of $N_y$ identical one-dimensional algebraic subproblems of $N_x$ unknowns. Consequently, as is seen in Fig. 9, the eigenvalue spectrum is found much closer to the one-dimensional spectrum, particularly for values of $\beta$ different from 0. For $\beta = 0$, the matrix $(G_\infty)_x \otimes I_y$ is defective; hence, as $\epsilon \to 0$, the eigenvalue problem may be viewed as a non-standard perturbation problem, and this may be the reason why the cloud is more diffuse. However, for $\beta = 1$, the same matrix is also defective, but this obviously does not have the same effect.

**3.2.2. Spectral radius.** We now examine more closely the behaviour of the spectral radius $\rho$, which is given in Tables 1 to 4 for various combinations of the parameters. In Tables 1 and 2, all possible situations with respect to the parities of the numbers $N_x$ and $N_y$ are examined; evidently, these parities have no significant effect on the spectral radius. Tables 1 and 2 give the results of experiments with equal, comparable, and very different parameters $\nu_x$, $\nu_y$, and for meshes of comparable sizes.

The first major observation is that for $N_x = N_y$, there exists a value $\bar{\beta} < 1/2$ and $\approx 1/2$ such that

$$(72)\qquad \rho_{2D}(\beta) \begin{cases} > \rho(\beta) & \text{if } \beta < \bar{\beta} < 1/2, \\ = \rho(\beta) & \text{if } \beta \geq \bar{\beta}, \end{cases}$$

i.e., the spectral radius equals the theoretical spectral radius for the one-dimensional case.

TABLE 1
*Spectral radius* ($\nu_x = \nu_y$).

| $N_x \times N_y$ | $9 \times 9$ | $10 \times 9$ | $10 \times 10$ |
|---|---|---|---|
| $\beta = 0$ | 0.98693 | 0.98866 | 0.99040 |
| 0.1 | 0.87353 | 0.87549 | 0.87746 |
| 1/3 | 0.56854 | 0.57045 | 0.57235 |
| 1/2 | 0.46985* | 0.47270 | 0.47553* |
| 2/3 | 0.47329* | 0.47581 | 0.47831* |
| 0.9 | 0.48936* | 0.49034 | 0.49133* |
| 1 | 0.5* | 0.5* | 0.5* |
| 0.49 | 0.46986* | 0.47271 | 0.47554* |
| 0.51 | 0.46986* | 0.47271 | 0.47554* |

*One-dimensional theoretical value.

In the case of Table 2, $\nu_x = 100\nu_y$, convection is preponderant in the $x$ direction. As a result, the algebraic system behaves more like that in one dimension. For $0 \leq \beta < \bar{\beta}$, since $\epsilon$ is small but finite the spectral radius, although different from the one-dimensional theoretical value, is found closer to it than the analogous value in the first column of Table 1 or 2.

TABLE 2
*Spectral radius.*

| | $\nu_x = 2\nu_y$ | | | | $\nu_x = 100\nu_y$ |
|---|---|---|---|---|---|
| $N_x \times N_y$ | $9 \times 9$ | $9 \times 10$ | $10 \times 9$ | $10 \times 10$ | $9 \times 9$ |
| $\beta = 0$ | 0.93387 | 0.93596 | 0.93750 | 0.939380 | 0.64278 |
| 0.1 | 0.83216 | 0.83417 | 0.83346 | 0.83548 | 0.49869 |
| 1/3 | 0.56350 | 0.56456 | 0.56635 | 0.567413 | 0.47653 |
| 1/2 | 0.46985* | 0.47175 | 0.47364 | 0.47553* | 0.46985* |
| 2/3 | 0.47329* | 0.47497 | 0.47665 | 0.47831* | 0.47329* |
| 0.9 | 0.48936* | 0.49001 | 0.49067 | 0.49133* | 0.48936* |
| 1 | 0.5* | 0.5* | 0.5* | 0.5* | 0.5* |
| 0.49 | 0.46986* | 0.47176 | 0.47365 | 0.47554* | 0.46986* |
| 0.51 | 0.46986* | 0.47176 | 0.47365 | 0.47554* | 0.46986* |

*One-dimensional theoretical value.

In the one-dimensional case, we have found that as $N_x \to \infty$, the spectral radius $\rho$ tends to 1/2 for all $\beta$. In contrast to this, in the two-dimensional case and for the same limit, the third-order upwind-biased method ($\beta = 1/3$) is less efficient than Fromm's scheme ($\beta = 1/2$). For example, in Table 3, as $N_x$ and $N_y$ increase, the spectral radius for $\beta = 1/2$ remains equal to the one-dimensional theoretical value which is bounded by 1/2, while this bound is violated by the third-order method.

TABLE 3
*Spectral radius* ($\nu_x = \nu_y$).

| $N_x \times N_y$ | $\beta = 1/3$ | $\beta = 1/2$ |
|---|---|---|
| $5 \times 5$ | 0.52253 | 0.40451* |
| $10 \times 10$ | 0.57235 | 0.47553* |
| $20 \times 20$ | 0.58423 | 0.49384* |
| $30 \times 30$ | 0.58633 | 0.49726* |

*One-dimensional theoretical value.

**3.2.3. Condition number.** In the last set of experiments of this section, an attempt was made to compute, along with the spectral radius $\rho$, the condition number $\kappa$ of the eigenvector matrix $U$:

$$\kappa = \kappa_U = \| U \|_2 \, \| U^{-1} \|_2. \tag{73}$$

This parameter becomes infinite when approaching a case where the matrix $G_\infty$ is defective. Hence we expect that such cases can be detected by the increase of this parameter. The results in Table 4 are given for a case of convection across the grid ($\nu_x = \nu_y$) and a case of convection almost along the grid ($\nu_x = 100\nu_y$).

The symbol "$\infty$" indicates that the smallest eigenvalue found is either 0 or a very small number due to loss of precision. In the two cases ($\nu_x = \nu_y$ and $\nu_x = 100\nu_y$), the number $N_x = N_y = 9$ is odd[5] and the results tend to indicate that only $\beta = 0$ and $\beta = 1$ result in a

[5] For the one-dimensional problem, it was observed that for $N_x$ even and $\beta = 1/2$, the matrix $G_\infty$ was defective ($\lambda = 0$ double) with no serious consequence since only one eigenvector was missing. In two dimensions, with $N_x = N_y$ even and $\nu_x/\nu_y$ sufficiently large, the same situation should be approached, according to (71), for $\beta = 1/2$ (this was verified experimentally for $N_x = N_y = 10$) with no more consequence on the iterative process. Only the $2 \times 2$ sub-block $\left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ appears $N_y$ times in the Jordan reduced form of the matrix $G_\infty$.

TABLE 4

*Spectral radius and condition number* ($N_x \times N_y = 9 \times 9$).

| | $\nu_x = \nu_y$ | | $\nu_x = 100\nu_y$ | |
|---|---|---|---|---|
| $\beta$ | $\rho$ | $\kappa$ | $\rho$ | $\kappa$ |
| 0 | 0.98693 | $\infty$* | 0.64278 | $\infty$* |
| 0.1 | 0.87353 | $0.10\times 10^9$-$0.12\times 10^9$ | 0.49869 | $0.59\times 10^7$ |
| 1/3 | 0.56854 | $0.37\times 10^5$ | 0.47653 | $0.41\times 10^5$ |
| 1/2 | 0.46985* | $0.3\times 10^4$-$0.9\times 10^4$ | 0.46985* | $0.13\times 10^5$ |
| 2/3 | 0.47329* | $0.61\times 10^5$ | 0.47329* | $0.41\times 10^5$ |
| 0.9 | 0.48936* | $0.1\times 10^8$-$0.6\times 10^8$ | 0.48936* | $0.49\times 10^8$ |
| 1 | 0.5* | $\infty$* | 0.5* | $\infty$* |
| 0.45 | 0.47016* | $0.612\times 10^4$ | 0.47016* | $0.795\times 10^4$ |
| 0.46 | 0.47005* | $0.598\times 10^4$ | 0.47005* | $0.779\times 10^4$ |
| 0.47 | 0.46996* | $0.601\times 10^4$ | 0.46996* | $0.799\times 10^4$ |

*One-dimensional theoretical value

defective amplification matrix. A minimum condition number is achieved near $\beta = 1/2$, at $\beta \approx 0.46$.

**3.3. Numerical experiments for the two-dimensional wave equation.** Numerical experiments were made for the two-dimensional linear wave equation (41), for a range of angles for the convection direction $\arctan(b/a)$, and for a range of differently shaped rectangular meshes. No major differences were seen for the different skew angles and for the different (not degenerate) shapes of the domain.

In this section we show the convergence behaviour for the linear wave equation for which the convection direction is skew to the grid, $a = b = 1$ (convection angle $45^o$). We use a square grid with $N \times N$ ($N = 10, 20, 40, 80$) gridpoints. Similar initial errors were used as for the one-dimensional case. The *oscillating initial error* is defined by $e_{ij} = (-1)^{i+j}$, $i, j = 1, 2, \ldots$. In the *random initial error* the error at all nodes is randomly chosen and uniformly distributed in the interval $(0, 1)$.

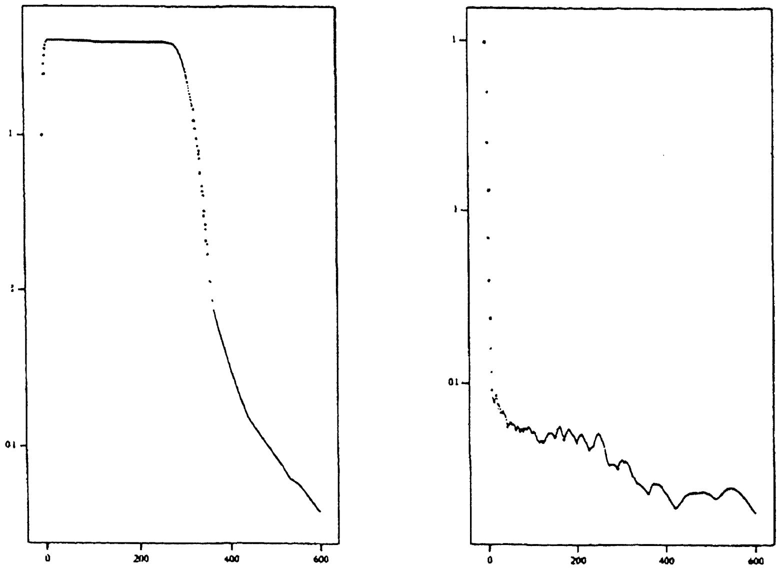
First, in Fig. 10 we show results for an iteration applied with the central scheme ($\beta = 0$). We see that the iteration does not converge. For a random initial error we see that some error components are rapidly damped in the first few steps, but after a couple of iterations the convergence hampers. As was seen by the Fourier theory, there are error components (both with low and high frequencies and in a large range of different directions) that cannot be damped. Also the matrix analysis shows that some eigenvalues may tend to 1.0 in this case.

For the nonpathological cases in Figs. 11 and 12, for $\beta = 1/2$ and $\beta = 1/3$, respectively, we observe an asymptotic rate of convergence corresponding with the values as given in §3.2.2, Table 3. For $\beta = 1/2$ (Fig. 11) the true rate is hard to observe on finer meshes because an additional effect is seen. Therefore it appears that for all nets there is a secondary phenomenon. This looks like the effect of a large Jordan box, corresponding to an eigenvalue that is close to (but definitely smaller than) the largest eigenvalue. For the coarser grids we see that this effect has disappeared after $2N$ iterations.

For the fully upwind scheme ($\beta = 1$), in Fig. 13 we see that there is again a pseudo-convection phase for $O(N)$ iterations, but the situation is much more complex than in the one-dimensional case.

## 4. Euler flow experiments.

**4.1. Introduction.** Several illustrative experiments have been carried out in a much more complex context than that of the analysis. In this section, the (steady) Euler equations are solved in two dimensions by the defect-correction method of Hemker, Spekreijse, and Koren[7], [8]. It is a finite volume method on a structured quadrilateral grid. It uses Osher's approximate Riemann solver for the numerical flux function, both in the first-order and the second-order method. The second-order approximation is computed by the MUSCL approach (without



Case a: oscillating initial error

Case b: random initial error

FIG. 10. *Convergence histories of iteration with central scheme.* $\beta = 0$, *two-dimensional computation over* $80 \times 80$ *mesh.*

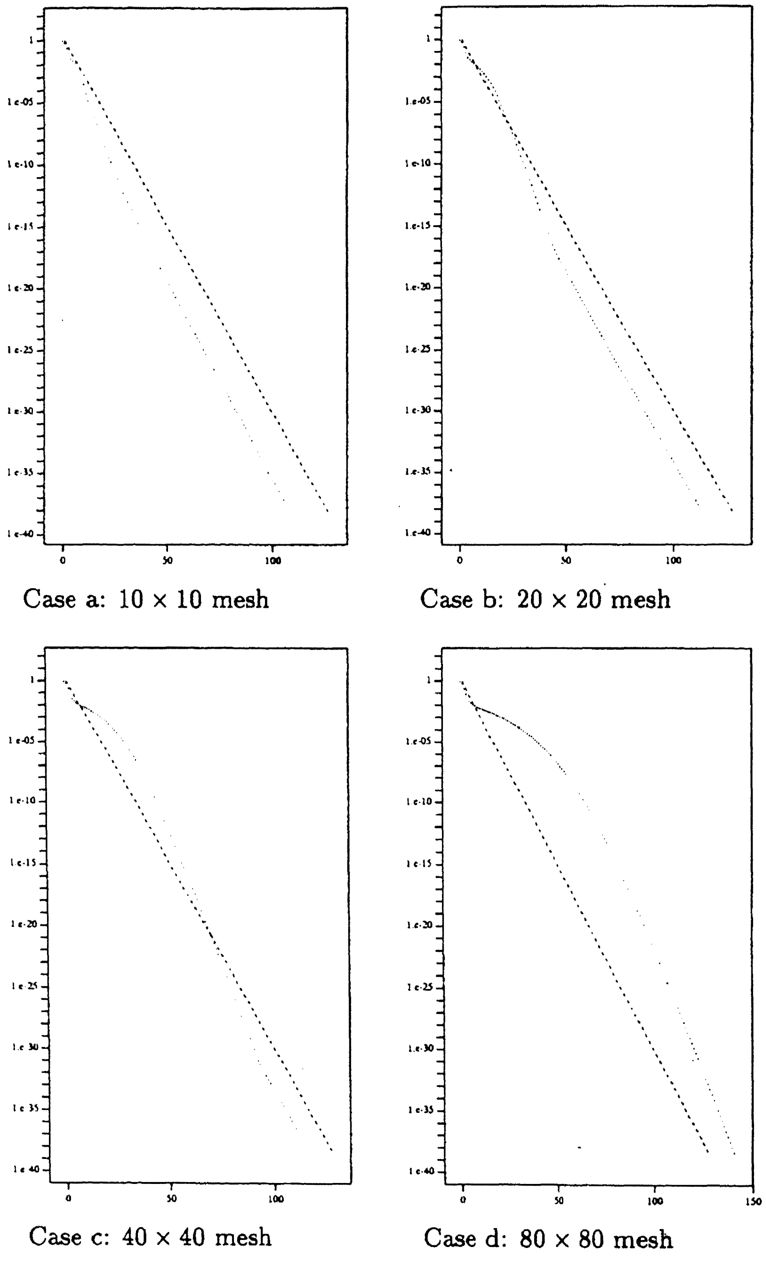
limiters). In this setting a parameter $\beta$ can be introduced that is completely analogous to the $\beta$ used in the previous sections. For details about the method we refer to [9].

The first-order discrete equations are solved by a nonlinear multigrid method. It employs a nonlinear symmetric Point–Gauss–Seidel relaxation as a smoother and a nested sequence of Galerkin discretisations for the coarse grid corrections. Experience has shown that a small number of iteration cycles of this multigrid method solve the discrete system to a high degree of accuracy. In the experiments shown, three FAS V-cycles were applied for each single defect-correction step. It was shown by experiments that the same results were obtained for multigrid iteration with two through five FAS V-cycles. All initial estimates were obtained by interpolation from a first-order accurate solution on a coarser grid.

In §§4.2 through 4.4 we show results for flow over the standard NACA0012 airfoil. Section 4.2 treats a subsonic flow with the Mach number at infinity $M_\infty = 0.63$ and the angle of attack $\alpha = 2.0^o$. This is a smooth flow where no special effects are to be expected except that, in contrast to the linear problem treated before, now the problem is described by a complex nonlinear system of equations and the domain is not simply connected.

The problem solved in §4.3 is an artificial problem that simulates an Euler flow with a pure contact discontinuity on a simply connected domain. This domain is the square $[0, 1] \times [0, 1]$, on which the boundary conditions are specified so that the contact discontinuity exists along the line $x = y$. The flow is from the bottom left to the top right and the boundary conditions are at left (inflow) $u = v = 0.5, c = \sqrt{2}$; at bottom (inflow) $u = v = 1.0, c = \sqrt{2}$; at outflow (right and top) $p = 1.0$ ($p$: presure, $c$: speed of sound, $u, v$: velocity components). The problems in §4.4 describe more or less standard transonic flow problems.

It will be clear that the problems solved in these sections deviate to a large extent from the linear model problems treated in the previous sections. The present examples use a non-linear system of equations and, moreover, the subsonic and transonic problems are not even fully hyperbolic. Nevertheless we show some of these results because we find that some of our experiences with such complex problems still sufficiently resemble the problems analysed.



Case a: 10 × 10 mesh

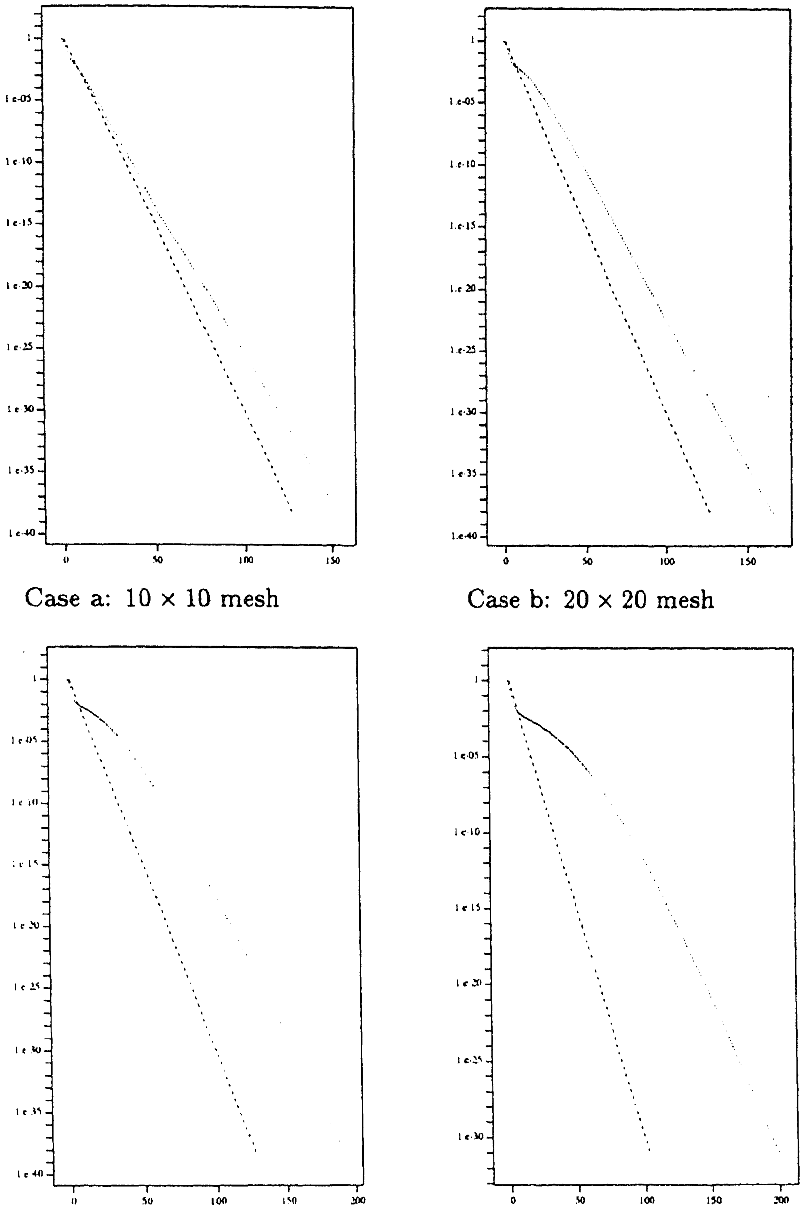
Case b: 20 × 20 mesh

Case c: 40 × 40 mesh

Case d: 80 × 80 mesh

FIG. 11. *Convergence histories of iteration with Fromm's scheme. The dashed line corresponds to a convergence rate of* 1/2 ($\beta = 1/2$, *random initial error*).

**4.2. Subsonic flow over an NACA0012 airfoil.** In Fig. 14 we see the convergence of the defect-correction iteration for subsonic flow and for different values of $\beta$. We see that the iteration does not converge for $\beta = 0$, as it does not for $\beta = 1$ (not shown). We obtain slow convergence for $\beta = 0.1$ and $\beta = 0.9$. Good convergence with a rate of approximately 0.5 per iteration step is obtained for $\beta = 1/3$, 1/2, and 2/3. The precise asymptotic rate cannot be observed because rounding error accuracy is obtained after approximately 40 iterations.



Case a: 10 × 10 mesh

Case b: 20 × 20 mesh

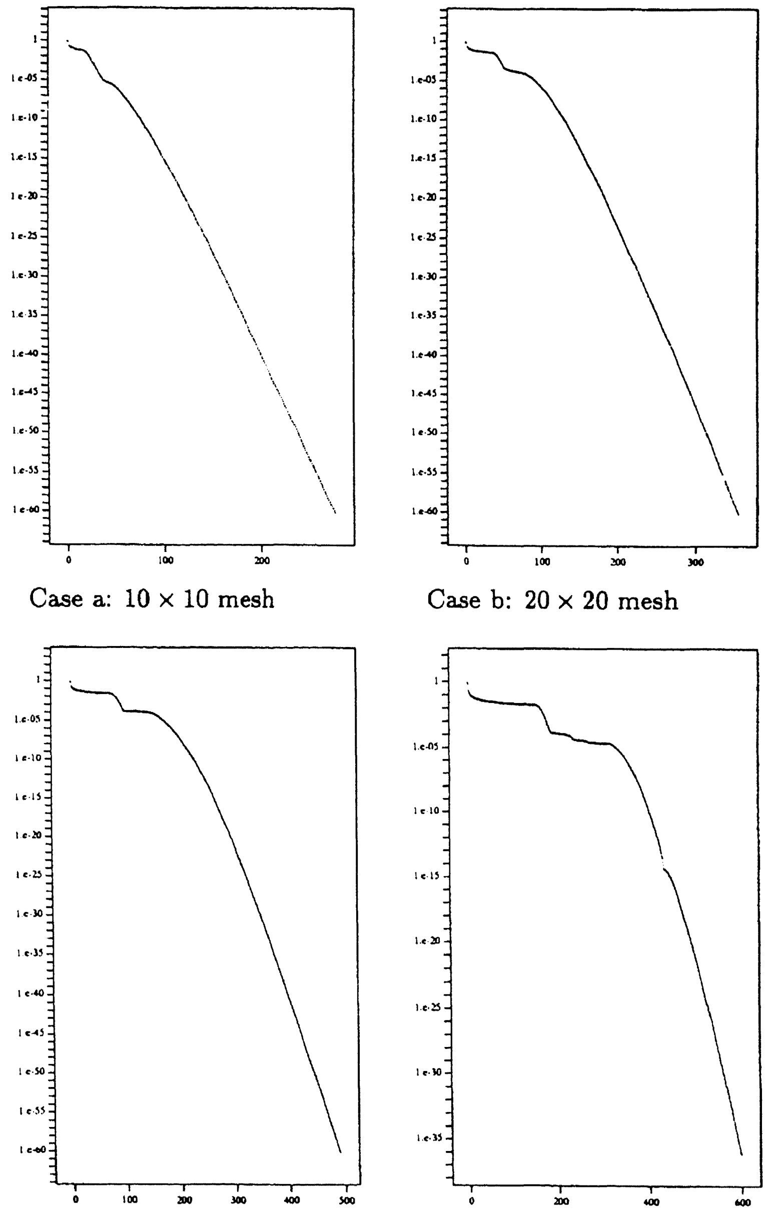
Case c: 40 × 40 mesh

Case d: 80 × 80 mesh

FIG. 12. *Convergence histories of iteration with upwind-biased scheme. The dashed line corresponds to a convergence rate of* 1/2 ($\beta = 1/3$, *random initial error*).

For $\beta = 1/3$ and $\beta = 2/3$ we see that after an initial phase with $\rho \approx 0.5$, we obtain another phase with a slightly slower convergence rate. Such effect is not (yet) seen for $\beta = 1/2$.

**4.3. Flow with a contact discontinuity.** For this flow with a contact discontinuity, convergence results are shown in Fig. 15: $\beta = 0$ gives a diverging process (not shown), and $\beta = 0.1$ shows worse convergence than $\beta = 0.9$. The asymmetry in the convergence behaviour with respect to $\beta > 1/2$ (better) and $\beta < 1/2$ (worse convergence) might be under-



Case a: 10 × 10 mesh

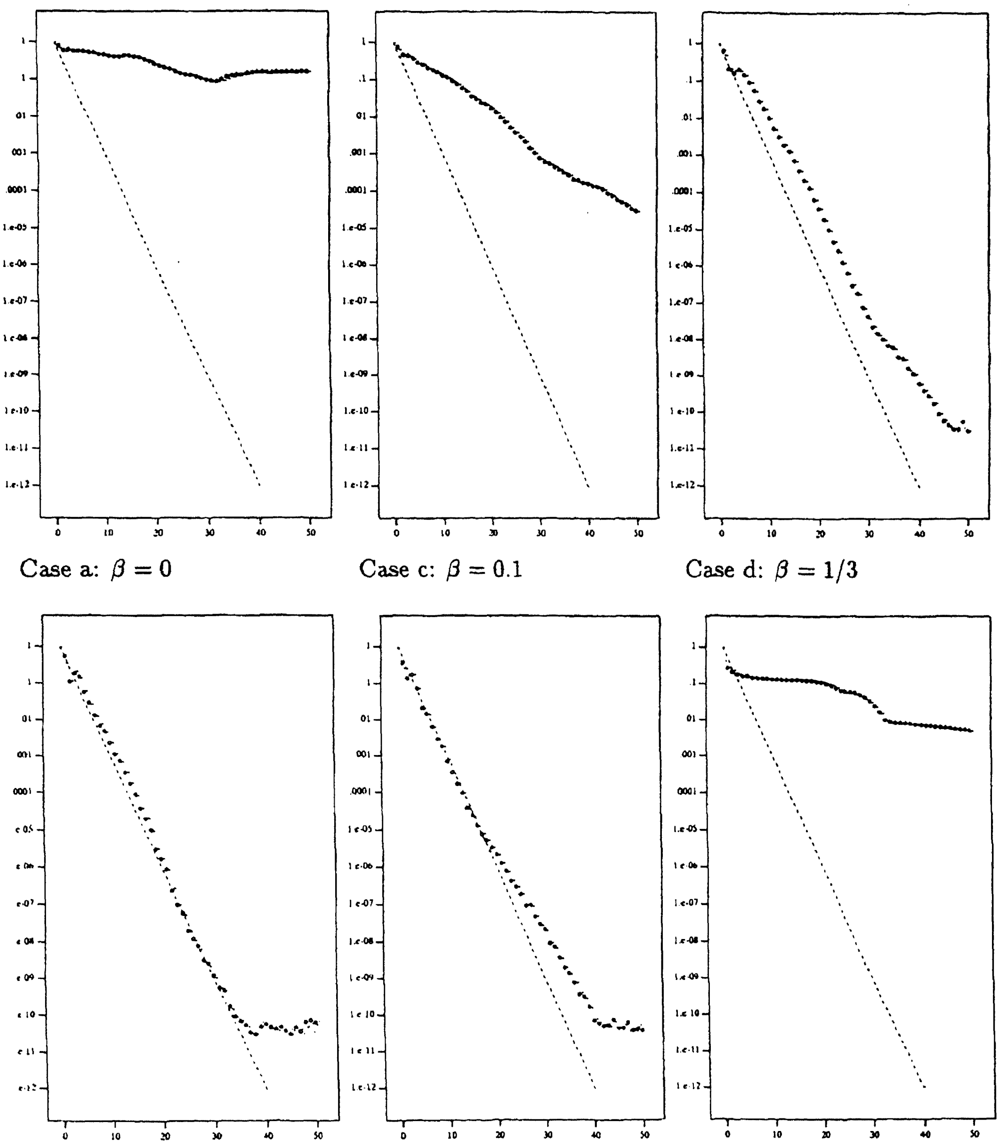
Case b: 20 × 20 mesh

Case c: 40 × 40 mesh

Case d: 80 × 80 mesh

FIG. 13. *Convergence histories of iteration with fully-upwind scheme.* ($\beta = 1$, *random initial error.*)

stood by the location of the eigenvalues in the complex plane (as shown in Fig. 9). There we see that more eigenvalues are located in the neighbourhood of the origin for $\beta > 1/2$ than for $\beta < 1/2$. This may be of greater importance for the nonlinear equations, where the corresponding eigenvectors are excited again and again, than for the linear problems, where the effect of these eigenvalues is no longer seen after a sufficient number of iterations.



Case a: $\beta = 0$

Case c: $\beta = 0.1$

Case d: $\beta = 1/3$

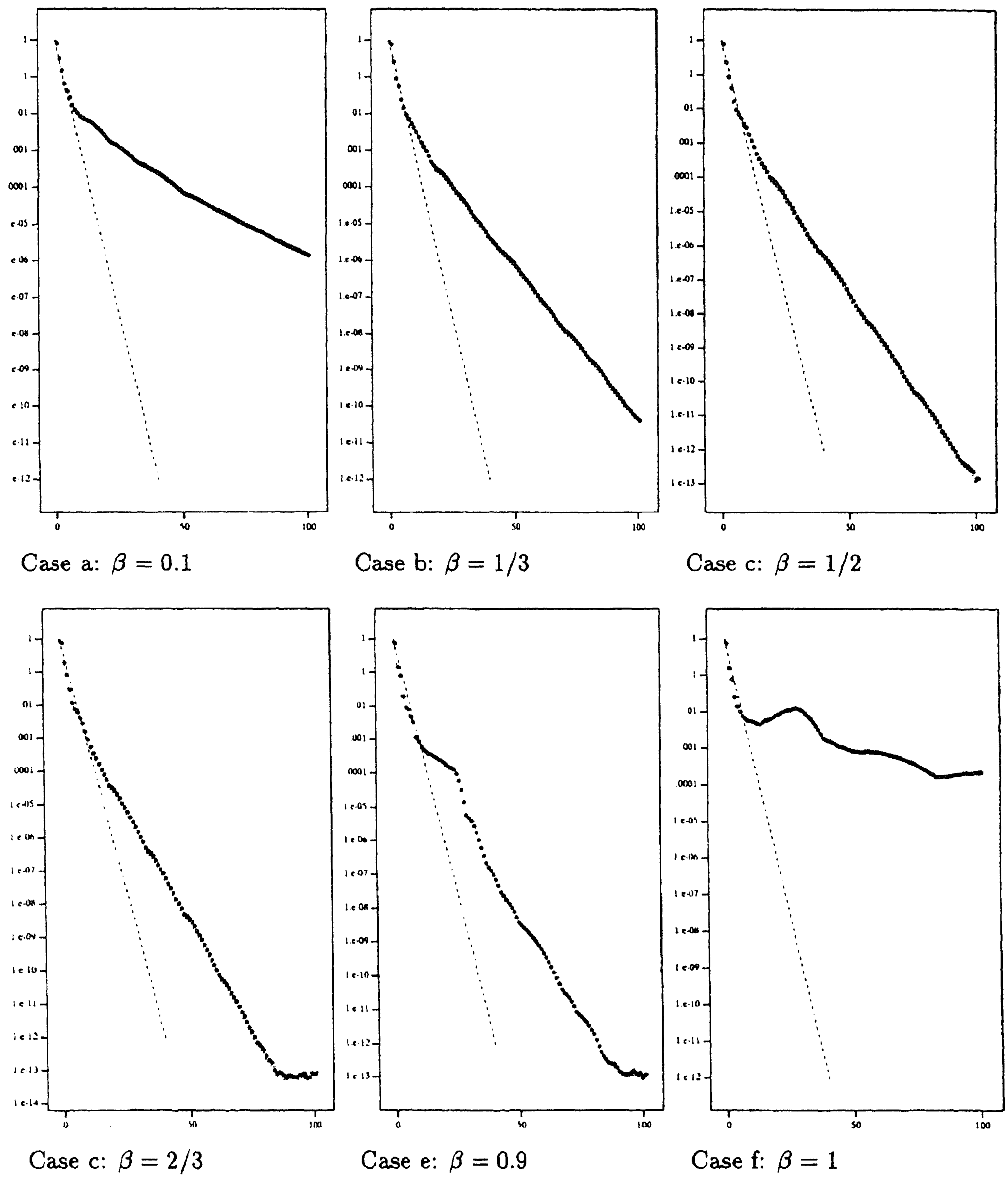
Case d: $\beta = 1/2$

Case e: $\beta = 2/3$

Case f: $\beta = 0.9$

FIG. 14. *Subsonic flow over an NACA0012 airfoil. Defect-correction method,* 20 × 32 *mesh. The dashed line corresponds to a convergence rate of* 1/2.

**4.4. Transonic flow over an NACA0012 airfoil.** In Fig. 16 we give results for a symmetric transonic flow, $M_\infty = 0.85$ and $\alpha = 0.0^o$. The flow in this problem has two shocks. In Fig. 17 we give results for an asymmetric transonic flow, $M_\infty = 0.85$ and $\alpha = 1.0^o$, in which problem we encounter an additional contact discontinuity. This last problem is also known from the GAMM workshop on the numerical simulation of compressible Euler flows (1986) [4]. The mesh used for the NACA airfoil is a 20 × 32 mesh (i.e., a level three mesh in a sequence of which the coarsest is a 5 × 8). Similar results, however, were obtained on the



Case a: $\beta = 0.1$

Case b: $\beta = 1/3$

Case c: $\beta = 1/2$

Case c: $\beta = 2/3$

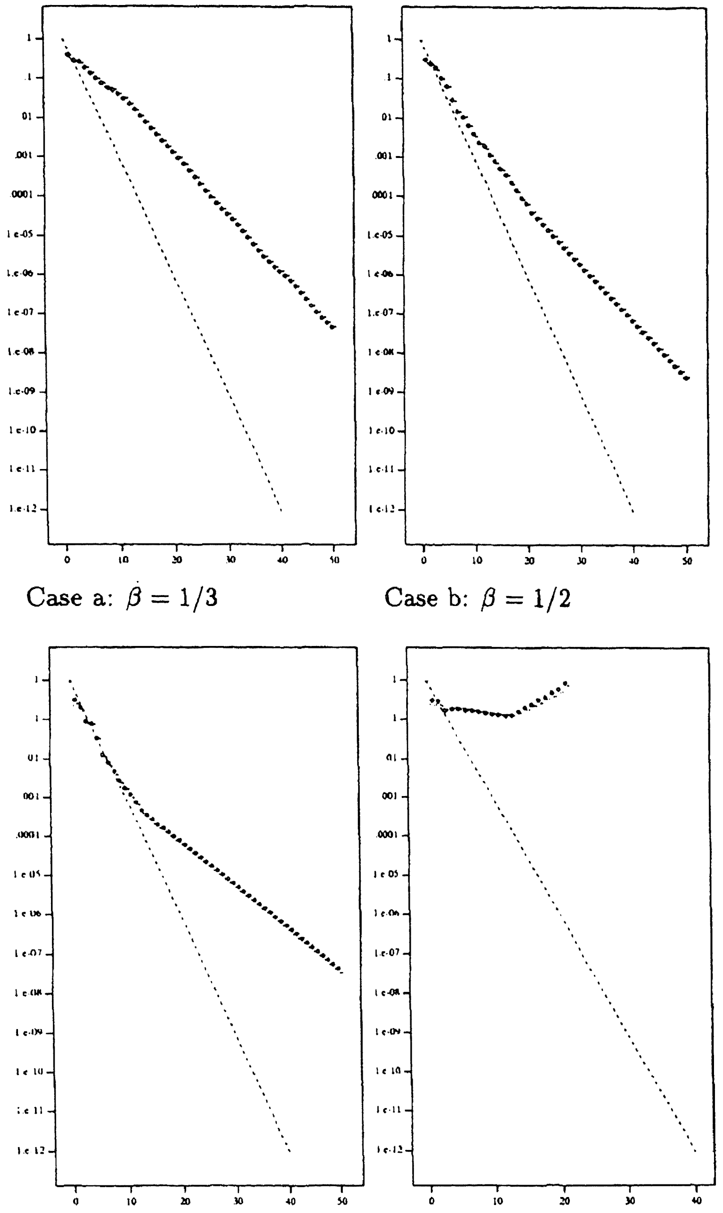
Case e: $\beta = 0.9$

Case f: $\beta = 1$

FIG. 15. *Flow with a* 45° *contact discontinuity. Defect-correction method,* 16 × 16 *mesh. The dashed line corresponds to a convergence rate of* 1/2.

40 × 64 mesh, with the only differences being that (i) some convergence effects were seen after a larger number of defect-correction iteration cycles, and (ii) some defect-correction convergence rates were slightly faster(!).

**Conclusions.** When very large timesteps are used, implicit time integration methods applied to the time-dependent Euler equations are identical to defect-correction methods applied to the steady discrete system. The convergence of such iterations was examined. Fourier and matrix analysis were applied for a linear model problem, both for the one- and two-dimensional



Case a: $\beta = 1/3$

Case b: $\beta = 1/2$

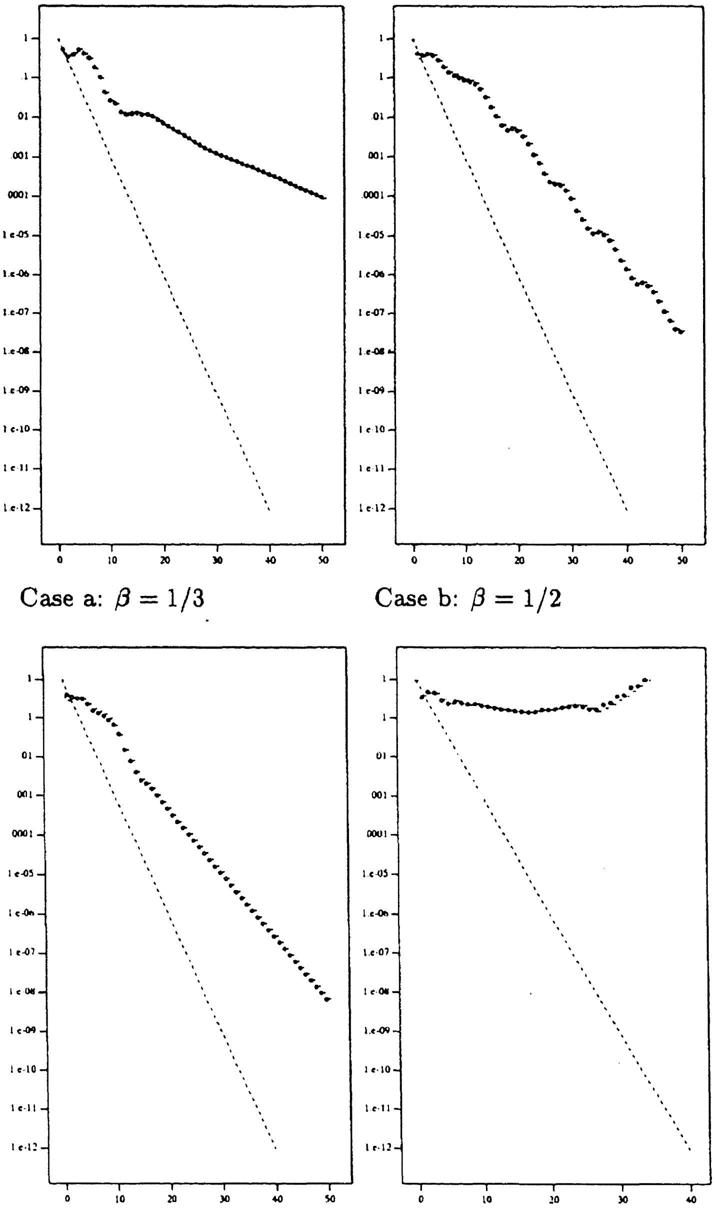
Case c: $\beta = 2/3$

Case d: $\beta = 0.9$

FIG. 16. *Symmetrical transonic flow over an NACA0012 airfoil. Defect-correction method,* 20 × 32 *mesh. The dashed line corresponds to a convergence rate of* 1/2.

cases. The convergence rate was evaluated, depending on a parameter $\beta \in [0, 1]$that determines the amount of upwinding present in the second-order discrete operator. The values $\beta = 0$ and $\beta = 1$ were shown to yield defective error amplification matrices.

The matrix analysis allowed us to understand the pathology of the schemes that were characterised by such defective amplification matrices. For the linear model problems, these schemes, before they achieved their asymptotic convergence, exhibited a *pseudo-convection*



Case a: $\beta = 1/3$

Case b: $\beta = 1/2$

Case c: $\beta = 2/3$

Case d: $\beta = 0.9$

FIG. 17. *Asymmetrical transonic flow over an NACA0012 airfoil. Defect-correction method,* $20 \times 32$ *mesh. The dashed line corresponds to a convergence rate of* $1/2$.

*phase* during which the norm of the residual may not have been reduced. The error, prior to being dissipated, was transferred over the mesh at a completely unphysical speed. This phase extended over a number of iterations equal to $N/(1-\rho)$ in which $N$ was the size of the largest defective Jordan block of the amplification matrix and $\rho$ was its spectral radius; thus this extent could have been very large. In one dimension, the non-pathological schemes asymptotically converged at the rate of the sequence $2^{-n}$. In two dimensions, schemes for which the upwinding parameter satisfied $\beta \geq 1/2$ also obeyed this, as a rule. However, after the initial impulsive start and before the asymptotic rate was reached, those schemes that were close to the pathological ones showed a *Fourier phase* during which the effect of boundary conditions was not yet felt. In this phase the spectral radius given by Fourier analysis did control the convergence rate.

In the last section, computations of two-dimensional Euler flows were presented. Calculations were shown for both smooth flows and flows with shocks and contact discontinuities. Several theoretical results for the model problem were confirmed in these more complex situations: (i) the recommended *half-upwind scheme* ($\beta = 1/2$) is shown to converge roughly[6] at a rate comparable with the sequence $2^{-n}$, whereas (ii) the *upwind-biased scheme* ($\beta = 1/3$) is slightly less efficient; (iii) the *central scheme* ($\beta = 0$) and the *fully upwind scheme* ($\beta = 1$) do not converge; and (iv) the schemes with $\beta$ close to 0 or 1 converge badly.

In summary, when using the implicit upwind scheme, in which the preconditioner is based on only first-order differencing while a second-order partially upwind approximation is constructed explicitly, use of the central-differencing scheme ($\beta = 0$), or the fully upwind scheme ($\beta = 1$) explicitly is not recommended because both result in defective methods with pathological iterative convergence. Preferably, one should use the Fromm scheme ($\beta = 1/2$) to realise the best separation of the eigenvalues. The more accurate upwind-biased scheme ($\beta = 1/3$) may be slightly less robust.

**Acknowledgments.** We want to thank Barry Koren, Marie-Hélène Lallemand, Walter Lioen, Hervé Stève, and Paul de Zeeuw for their helpful comments. Encouragement and advice by Alain Dervieux and Roger Peyret have been constant; we gratefully thank them.

## REFERENCES

[1] R. Beam and R. F. Warming, *An implicit finite-difference algorithm for hyperbolic systems in conservation-law-form*, J. Comput. Phys., 22 (1976), pp. 87–110.

[2] K. Böhmer, P. Hemker, and H. Stetter, *The defect correction approach*, Comput. Suppl., 5 (1984), pp. 1–32.

[3] A. Brandt, *The Weizmann institute research in multilevel computation*, in Proceedings of the Fourth Copper Mountain Conference on Multigrid Methods, J. Mandel, S. F. McCormick, J. E. Dendy, Jr., C. Farhat, G. Lonsdale, S. V. Parter, J. W. Ruge, and K. Stüben, eds., American Mathematical Society, Providence, RI, 1989, pp. 13–53.

[4] A. Dervieux, B. Van Leer, J. Périaux, and A. Rizzi, *Numerical Simulation of Compressible Euler Flows*, Vieweg Verlag, Braunschweig/Wiesbaden, 1989.

[5] J.-A. Désidéri and P. Hemker, *Analysis of the convergence of iterative implicit and defect correction algorithms for hyperbolic problems*, Institut National de Recherche en Informatique et en Automatique Report No. 1200, 1990.

[6] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, New York, 1985.

[7] P. Hemker and B. Koren, *A non-linear multigrid method for the steady Euler equations*, in Numerical Simulation of Compressible Euler Flows, A. Dervieux, B. Van Leer, J. Périaux, and A. Rizzi, eds., Vieweg Verlag, Braunschweig/Wiesbaden, 1989.

[8] P. Hemker and S. Spekreijse, *Multiple grid and Osher's scheme for the efficient solution of the steady Euler equations*, Appl. Numer. Math., 2 (1986), pp. 475–493.

[9] B. Koren, *Multigrid and defect correction for the steady Navier-Stokes equations: Application to aerodynamics*, CWI Tracts 74, Centrum voor Wiskunde en Informatica, Amsterdam, 1990.

[10] B. Van Leer and W. Mulder, *Relaxation methods for hyperbolic conservation laws*, in Proc. of the INRIA Workshop on Numerical Methods for the Euler Equations of Fluid Dynamics, Rocquencourt, France, December 7-9, 1983, F. Angrand, A. Dervieux, J. A. Désidéri, and R. Glowinski, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1985.

[11] R. MacCormack, *Current status of numerical solutions of the Navier-Stokes equations*, American Institute of Aeronautics and Astronautics Paper 85-0032, 1985.

[6] For the subsonic case the correspondence is almost perfect; in the presence of discontinuities the convergence reduces to roughly $\sqrt{2}^{-n}$.

[12] J. STEGER, *Implicit finite-difference simulation of flow about arbitrary geometries with application to airfoils*, American Institute of Aeronautics and Astronautics Paper 77-663, 1977.

[13] B. STOUFFLET, J. PÉRIAUX, F. FEZOUI, AND A. DERVIEUX, *Numerical simulation of 3-D hypersonic Euler flows around space vehicles using adapted finite elements*, American Institute of Aeronautics and Astronautics Paper 87-0560, 1987.

[14] J. THOMAS, B. VAN LEER, AND R. WALTERS, *Implicit flux-split schemes for the Euler equations*, American Institute of Aeronautics and Astronautics Paper 85-1680, 1985.

[15] B. VAN LEER, *Upwind difference methods for aerodynamic problems governed by the Euler equations*, in Large-Scale Computations in Fluid Mechanics, S. Osher, B. Engquist, and R. Somerville, eds., Lectures in Applied Mathematics, Vol. 22, American Mathematical Society, Providence, RI, 1984/1985, pp. 327–336.