Analysis of the convergence of iterative implicit and defect-correction algorithms for hyperbolic problems

Jean-Antoine Desideri, Pieter W. Hemker

INRIA

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 1200

Programme 7
Calcul Scientifique,
Logiciels Numériques et Ingénierie Assistée

ANALYSIS OF THE CONVERGENCE OF ITERATIVE IMPLICIT AND DEFECT-CORRECTION ALGORITHMS FOR HYPERBOLIC PROBLEMS

Jean-Antoine DESIDERI
Pieter W. HEMKER

Mars 1990

RR_1200

ANALYSIS OF THE CONVERGENCE OF ITERATIVE IMPLICIT AND DEFECT-CORRECTION ALGORITHMS FOR HYPERBOLIC PROBLEMS

ANALYSE DE LA CONVERGENCE ITERATIVE D'ALGORITHMES IMPLICITES DE RESIDU-CORRECTION POUR LES PROBLEMES HYPERBOLIQUES

Jean-Antoine DESIDERI
INRIA Centre de Sophia Antipolis
2004, Route des Lucioles
06560 Valbonne, France

and

Pieter W. HEMKER
CWI
Kruislaan 413
Amsterdam, The Netherlands

...to Harvard Lomax and Joseph L. Steger who,
some time ago,
have addicted me to matrix analysis.

J.-A. D.

Abstract

This paper studies the convergence of unfactored implicit schemes for the solution of the steady discrete Euler equations. In these schemes first and second order accurate discretisations are simultaneously used. The close resemblance of these schemes with iterative defect correction is shown.

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

In general, in the iteration, after an impulsive initial phase a slower pseudo-convective (or Fourier) phase can be distinguished, and finally again 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 β the pseudo-convection phase is less significant. Fromm's scheme ($\beta = 1/2$) or van Leer's third order scheme ($\beta = 1/3$) show a quite satisfactory convergence behaviour.

In this paper, first the linear convection problem in one and two dimensions is studied in detail. Differences between the various cases are signalized. In the last section experiments are shown for the Euler equations, including comments on how the theory is well or partially verified depending on the problem.

Résumé

Dans cet article, on étudie la convergence de schémas implicites de résolution discrète des équations d'Euler stationnaires. Dans ces schémas on utilise simultanément des discrétisations précises au premier et au second ordre. On met en évidence le lien étroit qu'il existe entre ces constructions et les méthodes itératives dites de "Defect-Correction" (Résidu-Correction).

On introduit des modèles théoriques linéaires pour les cas mono- et bi-dimensionnels. Les problèmes modèles sont analisés en détail par l'analyse de Fourier et par l'analyse matricielle. Il apparait que la nature de la convergence dépend très fortement d'un paramètre β qui contrôle le degré de décentrage introduit dans l'approximation du second ordre.

En général, au cours de l'itération, après la phase initiale du départ impulsif, on distingue une phase plus lente de pseudo-convection (ou phase de Fourier), enfin apparait la phase plus rapide de convergence asymptotique. Les deux valeurs extrèmes du paramètre de décentrage, $\beta = 0$ (aucun décentrage) et $\beta = 1$ (décentrage total), sont des cas particuliers pour lesquels le comportement itératif est dégénéré. En pratique, on recommende de ne pas utiliser ces valeurs. Pour les valeurs intermédiaires de β, la phase de pseudo-convection joue un rôle moins important. Le schéma de Fromm ($\beta = 1/2$) et le schéma du troisième ordre de van Leer ($\beta = 1/3$) démontrent une convergence satisfaisante.

Dans la rédaction, on étudie d'abord en détail le problème linéaire de convection en une et deux dimensions d'espace. Les différences entre les différents schémas sont mises en évidence. Dans la dernière partie, on présente des expériences sur les équations d'Euler, et on s'attache à démontrer comment la théorie s'applique plus ou moins bien suivant le problème.

Table of Contents

I. INTRODUCTION

1. Model Problems and Differencing Schemes

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 4×4 Jacobian matrices, that can be diagonalized explicitly either when x and y are Cartesian coordinates or generalized coordinates [WBHT, JLS1]. To allow a **linear analysis** of numerical schemes, one may **construct a model hyperbolic test equation by setting A and B to constant matrices**, subject to the condition that any linear combination of A and B should be diagonalizable.

In one dimension, and after diagonalization, the above system reduces to a set of **convection equations**, and another 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, $x - ct =$ constant. Nevertheless, the exact solution over the spatial interval $[0, X]$ (for some fixed X), is stationary for $t \geq \frac{X}{c}$ (and uniform): $u(x,t) \equiv 1$. The process of convergence (to steady state) is therefore distinct from that of a dissipative phenomenon (e.g. the heat equation). However, in the discrete models, dissipation may exist in the form of **artificial dissipation**.

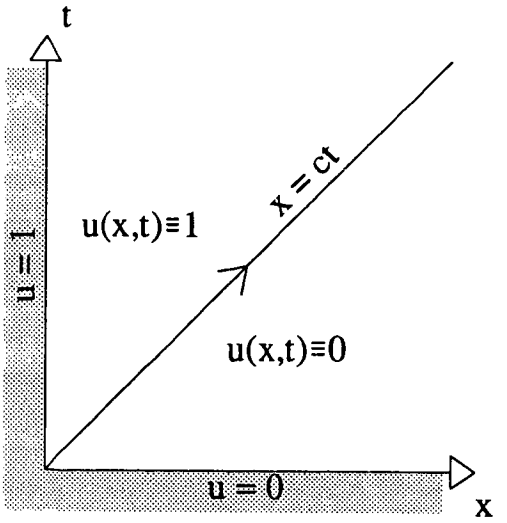


Figure I.1: Quarter-Plane Problem.

For the model problem, various differencing schemes may be employed to represent the spatial derivative u_x (multiplied by Δx): central differencing, δ_x^c, 1st-order backward differencing, $\delta_{x,1}^u$, 2nd-order backward differencing, $\delta_{x,2}^u$. [1] The matrix analogs of these operators are defined in detail in Appendix I in the periodic case, and in the case where some appropriate non-periodic boundary conditions are applied. The corresponding sets of eigenvalues are also shown. It should be stressed that the two cases, although represented by matrices that differ from one another by a few elements only, are very distinct in nature. The periodic model is adequate to carry out a Fourier analysis, yielding L2-stability conditions, and phase-error and wavespeed evaluations. However, the non-periodic model is more appropriate for (asymptotic) iterative convergence rate estimations.

The following general statements can be made:
(1) All the **eigenvalues** λ_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 **diagonalized**. In the periodic case, the eigenvalues are **purely imaginary** which indicates the absence of artificial viscosity. In the non-periodic case, all but one eigenvalues can be shown to satisfy $\Re(\lambda_m) = O(\log(N)/N)$.
(3) In the non-periodic case, **upwind schemes** are represented by **defective (i.e. non diagonalizable) triangular** 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.

The first of these statements is a **well-posedness requirement** for the spatial-differencing scheme, since it is equivalent to the stability of the time-continuous, space-discrete equation

$$u_t + D_h u = 0 \quad , \tag{4}$$

in which $D_h u$ is the employed discrete approximation of cu_x ($D_h = \frac{c}{\Delta x}\delta_x^c$, or $\frac{c}{\Delta x}\delta_{x,1}^u$, or $\frac{c}{\Delta x}\delta_{x,2}^u$, $\cdots$), since the solution of (4) is given by:

$$\mathbf{u}(t) = (u_1, u_2, \cdots, u_N)^T = \sum_m c_m e^{-\frac{c\lambda_m t}{\Delta x}} \mathbf{v}_m \quad , \tag{5}$$

in which $\mathbf{v}_m$ is the mth eigenvector of D_h, λ_m the associated eigenvalue, and c_m the component along $\mathbf{v}_m$ of the initial solution. Note that the stability of the space differencing is usually a necessary condition for the time-discrete scheme, but certainly not a sufficient one, since the time integration of (4) has been performed exactly, and involves no timestep limitation.

We now turn our attention to upwind schemes. The real parts of their eigenvalues are of order 1. This corresponds to the important amount of artificial dissipation inherent in this type of approximation. Equation (5) implies that in the convergence to steady state, along with convection of errors away across and away from the domain, dissipation at finite rate plays a role. The fact that these operators are represented in the model problem by defective (triangular) matrices has important consequences on the nature and properties of the iterative convergence.

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

2. Unfactored implicit schemes

We now turn to the time-discretisation method, sometimes referred to as the **solver**. For the solution of steady problems, implicit schemes are attractive since 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 **linearized backward Euler scheme** also known as the **fully implicit method**. It can be written in the following **delta form**:

$$M_h \left(w^{n+1} - w^n\right) = -\Delta t D_h w^n \tag{6}$$

where Δt is the timestep and the operator M_h is defined by

$$M_h = I + \Delta t \left(D_h w^n\right)_w \tag{7}$$

in which the subscript w indicates that the Jacobian is formed. To evaluate the stability of this method we consider again the linear hyperbolic model, for which D_h and M_h can be thought as matrices constant during the iteration and satisfying

$$M_h = I + \Delta t D_h \tag{8}$$

so that, an amplification matrix $G_{\Delta t}$ can be defined by

$$w^{n+1} = G_{\Delta t} w^n + b \tag{9}$$

in which b is a constant vector containing prescribed boundary terms, and turns out to be

$$G_{\Delta t} = I - \Delta t M_h^{-1} D_h = I - \left(I + (\Delta t D_h)^{-1}\right)^{-1}. \tag{10}$$

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

$$g_m(\Delta t) = 1 - \frac{1}{1 + \frac{1}{z_m}} = \frac{1}{z_m + 1} \tag{11}$$

where $z_m = \lambda_m \Delta t$. Since for all m,

$$\Re(z_m) \geq 0 \tag{12}$$

as established in the previous section, it follows that for all Δt

$$|g_m(\Delta t)| \leq 1 \tag{13}$$

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

$$\lim_{\Delta t \to \infty} g_m(\Delta t) = 0 \tag{14}$$

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 Δt that are **not limited by the CFL condition** which restricts the usual explicit schemes, and becomes **more dissipative with larger timesteps** while the steady-state solution, which is independent of Δt, is only determined by the differencing operator D_h appearing explicitly on the right-hand side. (This in contrast to e.g. the Lax-Wendroff type schemes, for which the steady state depends on the time step used.)

Note that if one lets $\Delta t \to \infty$ and defines $\Phi(w) = D_h w$, then (6) becomes

$$\Phi_w(w^n)\left(w^{n+1} - w^n\right) = -\Phi\left(w^n\right). \tag{15}$$

In this formulation, we recognize **Newton's method** whose convergence is **quadratic** [ORTE]. (For a linear problem, the amplification matrix G_∞ is then equal to the null matrix and the process converges in one iteration.) This confirms that being able to use stably very large time-steps is a highly desirable feature for the solver. Again this property falls for usual factored schemes.

We now examine the algorithmic standpoint. The application of the algorithm defined in (6) is performed in three steps:
(1) **Physical phase**:
computation of the **right-hand side vector** $R = -\Delta t D_h w^n$
(2) **Mathematical phase**:
solution of the system $M_h \Delta w^n = R$, in which the unknown is the vector Δw^n;
(3) **Update**:

$$w^{n+1} = w^n + \Delta w^n \quad . \tag{16}$$

The implicit mathematical 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, however, defines alone 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** [BMWG, JLS2, STOU], or instead a **second-order upwind scheme**. This alternative has gained some popularity in recent papers [JLS3, VLMU, TVLW, RMCK, FEM3], because it yields schemes having better monotonicity properties and thus producing more physically relevant solutions near discontinuities. Another alternative is to combine (linearly) a central discretisation with an upwind discretisation. In any case, the requirement of obtaining a second-order accurate space discretisation is generally not very difficult to meet because the step is explicit; moreover, in many (but not all) simple schemes, no Jacobians need be calculated but only linear combinations of the flux vectors; that is, the approximation $D_h w^n$ is computed without having to explicitly evaluate the operator D_h. In contrast, the mathematical phase is far more complicated to realize, and this for several reasons:
• For almost all solution methods, it is imperative to evaluate the operator M_h itself. If splitting is not used, the constituent blocks of M_h are linear combinations of the 4x4 Jacobians (in 2-d), that involve 4 times more functions to compute than in the flux vectors themselves. If splitting is applied, the true Jacobian $(D_h w^n)_w$ may be difficult to express in closed-form, while an approximate linearization may already be quite involved computationally. In addition, the precise linearization of the boundary procedure in a 2nd-order scheme is also difficult.
• The system $M_h \Delta w^n = b$ needs to be solved. This is always a computationally difficult and expensive task when the system is large, particularly when the mesh is

not regular in structure, or when the discretisation is sophisticated and complex.
• Inversion processes necessitate the storage of the constituent blocks of the matrix M_h. When this matrix is too large, the limit of the computer's storage capability can be attained and there is no alternative to overwriting certain blocks and reevaluating them every time they reappear in a subsequent calculation. Thus some of the Jacobians are evaluated more than once, and more work is required than would be expected solely by inspection of the mathematical equations.

For all of these reasons, many authors are using simplified versions of the implicit method, in which the Euler equations are approximated only to first-order in the mathematical phase, while the physical phase remains the same. This **reduces the bandwidth** of the matrix M_h and thus significantly lessens the amount of computation done in the mathematical phase. In addition, there is an important advantage in using the first-order scheme implicitly: in doing so, at least for the model problem, the matrix system to be solved at each iteration is **diagonally dominant**. Therefore the inversion can be performed by an iterative procedure that does not require the explicit Jacobian matrix, e.g. **multigrid** ([MCCM], [HACK], [BKOR], [MHL1]) or **relaxation**: either **Gauss-Seidel** iteration [CHAK, FZBS], or **Point-Jacobi** iteration [FARL, STV1, STV2] which has regained interest in **vectorized** computations.

However with a first-order implicit **preconditioner**, some inconsistency in the formulation is introduced, 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 paper, in particular for a class of implicit schemes in which a parameter β $(0 < \beta < 1)$ controls the degree of upwinding introduced in the 2nd-order differencing scheme.

3. The model Implicit Upwind Schemes

In preparation of the next Section, we introduce notations to analyze the iterative convergence of the implicit delta scheme applied to the model problem in the case where a **second-order** difference operator of **adjustable upwinding** is employed in the explicit phase,

$$D_h = \frac{c}{\Delta x}\delta^{\beta}_{x,2} \tag{17}$$

where $\delta^{\beta}_{x,2}$ combines the fully-upwind scheme with the central differencing scheme,

$$\delta^{\beta}_{x,2} = \beta\delta^{u}_{x,2} + (1-\beta)\delta^{c}_{x} \tag{18}$$

$(0 \leq \beta \leq 1)$, and a **first-order** upwind scheme is applied in the implicit phase:

$$M_h = I + \frac{c\Delta t}{\Delta x}\delta^u_{x,1} \tag{19}$$

Thus the iteration is defined by

$$\left(I + \frac{c\Delta t}{\Delta x}\delta^u_{x,1}\right)(\mathbf{u}^{n+1} - \mathbf{u}^n) = -\frac{c\Delta t}{\Delta x}\delta^\beta_{x,2}\mathbf{u}^n \tag{20}$$

This allows us to again define the amplification matrix, $G_{\Delta t}$, by

$$\mathbf{u}^{n+1} = G_{\Delta t}\mathbf{u}^n + b \quad . \tag{21}$$

Consequently, when $\Delta t \to \infty$, the amplification matrix $G_{\Delta t}$ approaches

$$G_\infty = I - \left(\delta^u_{x,1}\right)^{-1}\delta^\beta_{x,2}. \tag{22}$$

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

4. The Defect-Correction Method

The iteration (22), derived in the previous section, can be seen as a particular application of the **Defect Correction Method** [DCM1]. In a **Defect Correction Iteration** the solution $\mathbf{u}^*$ of a linear or nonlinear equation

$$\Phi_2(\mathbf{u}) = \mathbf{f} \tag{23}$$

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

$$\Phi_1(\mathbf{u}^1) = \mathbf{f} \tag{24}$$

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

$$\Phi_1(\mathbf{u}^{n+1}) = \Phi_1(\mathbf{u}^n) - \Phi_2(\mathbf{u}^n) + \mathbf{f}. \tag{25}$$

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 (23).

If the operators Φ_1 and Φ_2 are differentiable, with nonsingular Jacobian matrices $D\Phi_1$ and $D\Phi_2$, then a small error $\mathbf{e}^{n+1} = \mathbf{u}^{n+1} - \mathbf{u}^*$ approximately satisfies the equation

$$D\Phi_1(\mathbf{e}^{n+1}) = D\Phi_1(\mathbf{e}^n) - D\Phi_2(\mathbf{e}^n). \tag{26}$$

Hence, the linear error amplification operator is given by

$$G_\infty = I - (D\Phi_1)^{-1} D\Phi_2. \tag{27}$$

In this way, second-order approximations are evaluated only to form right-hand sides in (25). Inversions only involve the simpler first-order approximation scheme. In our case, for the steady state (i.e. $\Delta t \to \infty$)

$$\Phi_1 = \frac{c}{\Delta x}\delta^u_{x,1}, \qquad \Phi_2 = \frac{c}{\Delta x}\delta^\beta_{x,2}, \tag{28}$$

and the identification is obvious. In a similar way, for finite Δt, the fully implicit method is identified with a defect correction iteration for which

$$\Phi_1 = M_h, \qquad \Phi_2 = \Delta t D_h. \tag{29}$$

Defect correction iteration has interesting implications from the point of view of stability and accuracy. In the linear case, because only the operator Φ_1 is inverted, for a stable approximate solution $\mathbf{u}^n$ only stability of the operator Φ_1 is required. 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 Φ_1 is stable, and if Φ_i is a p_i-th order discretisation of a continuous operator Φ, (i=1,2), $p_1 < p_2$, then $\mathbf{u}^n$ is an $O(h^{\min(ip_1,p_2)})$ 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, as described in Section 1.1.3, with a factor $O(h)$ per iteration step.

It is interesting to know that the *high-frequency components* that are slowly converging to the solution of $\Phi_2(\mathbf{u}) = \mathbf{f}$ are essentially the same that give a poor approximation to the continuous problem. This is illustrated e.g. by the results of B.Koren ([BKOR], Figure 5.7).

Although these arguments (as well as reasons of efficiency) suggest that a small number of iterations is advantageous above iteration to convergence, and that -in general- it will be unwise to iterate (25) until convergence, we want to study in general the convergence behaviour of the above iteration procedure for hyperbolic problems.

II. ONE-DIMENSIONAL ANALYSIS

1. Fourier type Analysis

1.1. *The interior domain*

In this section we first give the analysis of the defect correction iteration process, 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. Therefore, we consider the operators $\delta_{x,1}$ and $\delta^{\beta}_{x,2}$ working on a uniform discretisation of the line $(-\infty, +\infty)$. Then $\delta_{x,1}$ and $\delta^{\beta}_{x,2}$ can be represented by infinite Toeplitz matrices of the form

$$\delta_{x,1} = \text{Trid}\,(-1, 1, 0), \tag{30}$$

and

$$\begin{aligned}\delta^{\beta}_{x,2} &= (1-\beta)\ \text{Trid}\,(-\frac{1}{2}, 0, \frac{1}{2}) + \beta\ \text{Pentad}\,(\frac{1}{2}, -2, \frac{3}{2}, 0, 0) \\ &= \frac{1}{2}\ \text{Pentad}\,(\beta, -3\beta - 1, 3\beta, 1-\beta, 0).\end{aligned} \tag{31}$$

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

$$u_h(jh) = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{i\omega hj} F(u_h)(\omega) d\omega, \tag{32}$$

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

$$F(\delta_{x,1})(\omega) = -e^{-i\omega h} + 1 = 2ie^{-i\omega h/2} \sin(\omega h/2). \tag{33}$$

Similarly

$$\begin{aligned}F(\delta^{\beta}_{x,2})(\omega) = 2ie^{-i\omega h/2}\ [\ &\sin(\omega h/2) + \\ &i\ \cos(\omega h/2)\sin^2(\omega h/2) + (2\beta - 1)\sin^3(\omega h/2)],\end{aligned} \tag{34}$$

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

$$\begin{aligned}F(G_\infty)(\omega) &= 1 - (F(\delta_{x,1})(\omega))^{-1}\ F(\delta^{\beta}_{x,2})(\omega) \\ &= i\sin(\omega h/2)\cos(\omega h/2) + \kappa \sin^2(\omega h/2),\end{aligned} \tag{35}$$

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

$$|F(G_\infty)(\omega)| = |\sin(\omega h/2)|\sqrt{\cos^2(\omega h/2) + \kappa^2 \sin^2(\omega h/2)} \tag{36}$$

and

$$\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)| = \sup_{t\in(0,1)} \sqrt{\kappa^2 t^2 + t(1-t)}. \tag{37}$$

As upperbounds we find

$$\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)| = \frac{1}{2}\frac{1}{\sqrt{1-\kappa^2}} \quad \text{for } \kappa^2 \leq 1/2, \tag{38}$$

and

$$\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)| = |\kappa| \quad \text{for } 1/2 \leq \kappa^2 \leq 1. \tag{39}$$

Special cases are

$$\begin{aligned} &\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)| = 1 \quad \text{for } \beta = 0 \quad \text{or} \quad \beta = 1, \\ &\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)| = 1/2 \quad \text{for } \beta = 1/2, \\ &\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)| = \frac{3}{8}\sqrt{2} \approx 0.530 \quad \text{for } \beta = 1/3. \end{aligned} \tag{40}$$

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 Figure (II.1) we give a picture of this amplification operator, $\sup_{\omega\in(-\pi/h,\pi/h)} |F(G_\infty)(\omega)|$, as a function of β.

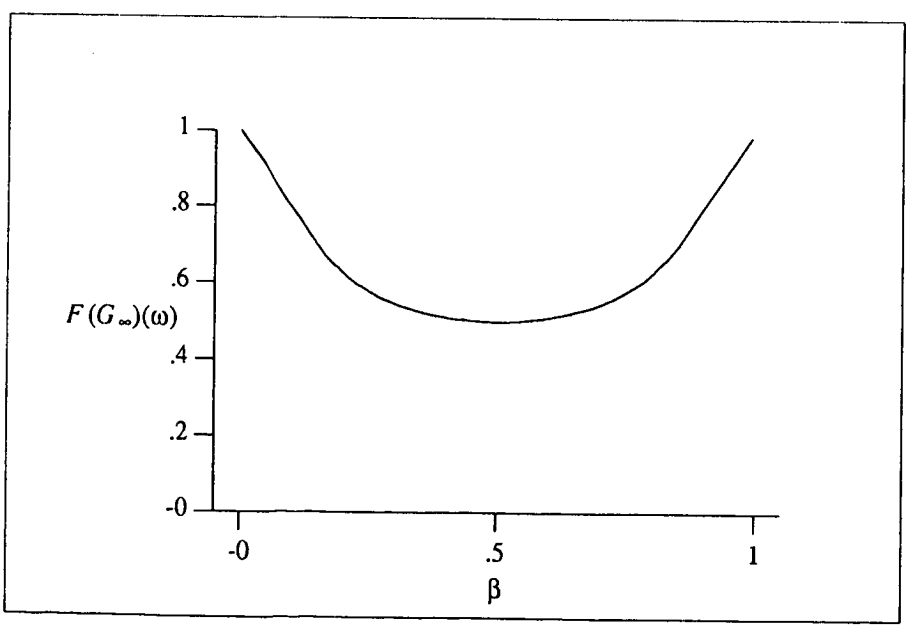


Figure II.1: Fourier Amplification Factor (1D).

From the analysis we see that convergence can be expected for $\beta \in (0,1)$. 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. The values $\beta = 0$ and $\beta = 1$ are special cases for which no convergence can be expected.

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 Section 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.

1.2. *The boundary domain*

To obtain an impression of the influence of the inflow Dirichlet boundary, we consider grid functions on a uniform partition $\{x_i = ih;\ i = 0,1,2,...\}$ of the half-line $[0,\infty[$ and we restrict ourselves to error components that vanish for large x_i. (I.e. we consider errors that live only in some neighbourhood of the boundary.) The operators $\delta_{x,1}$ and $\delta^{\beta}_{x,2}$ are again described by (30, 31), except for the first two equations (those related to the boundary) that are determined by the boundary discretisation (see e.g. appendix II).

The amplification operator G_∞ of the defect correction iteration is given by $G_\infty = I - (\delta_{x,1})^{-1}\delta^{\beta}_{x,2}$. Eigenfunctions u_λ of G_∞ and corresponding eigenvalues λ satisfy the relation

$$\delta^{\beta}_{x,2}\, u_\lambda = (1-\lambda)\,\delta_{x,1}\, u_\lambda \tag{41}$$

and from (30, 31) it follows that u_λ has the form

$$u_\lambda(jh) = A_0 + A_1\mu_1^j + A_2\mu_2^j\ , \tag{42}$$

where μ_1 and μ_2 are roots of the equation

$$(1-\beta)\mu^2 + (2\beta + 2\lambda - 1)\mu - \beta = 0\ . \tag{43}$$

The constants A_i, $i = 0,1,2$ are determined by the boundary condition and the discretisation scheme used at the points near the boundary. Since we are only interested in real errors that live in the neighbourhood of the boundary, the relevant eigenfunctions are restricted to those with $|\mu| \leq 1$. Further, because the eigenfunctions should be real, we see from (42) that either $\mu_1, \mu_2 \in I\!R$ or $|\mu_1| = |\mu_2|$, $\mu_1 \neq \mu_2$.

In the case of real μ, only those $\lambda \in \mathbb{R}$ for which $|\mu_1| \leq 1$ and $|\mu_2| \leq 1$ are acceptable. This implies $D = (2\beta + 2\lambda - 1)^2 + 4\beta(1-\beta) \geq 0$ and

$$-2(1-\beta) \leq -2\beta - 2\lambda + 1 \pm \sqrt{D} \leq 2(1-\beta) \; . \tag{44}$$

The left inequality implies $\lambda \leq 1 - 2\beta$, the right one $\lambda \geq 0$, which combines to the requirement $0 \leq \lambda \leq 1 - 2\beta$. In fact, eigenfunctions vanishing at infinity (i.e. as $x_i \to \infty$) are such that $A_0 = 0$, and they satisfy the Dirichlet condition $u_\lambda(x_0) = 0$ if, in addition, $A_1 + A_2 = 0$ holds, which gives:

$$u_\lambda(x_1) = A_1\mu_1 - A_1\mu_2, \tag{45}$$

$$u_\lambda(x_2) = A_1\mu_1^2 - A_1\mu_2^2, \tag{46}$$

which implies

$$u_\lambda(x_2)/u_\lambda(x_1) = \mu_1 + \mu_2 = -(2\beta + 2\lambda - 1)/(1-\beta) \; \in \mathbb{R} \; . \tag{47}$$

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

$$(1-\beta)\; u_\lambda(x_2) = (2(1-\lambda) - 2\beta)\; u_\lambda(x_1) \; . \tag{48}$$

In the case of complex μ 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,

$$\frac{-\beta}{1-\beta} = \mu_1\mu_2 = \mu_1^2 \; e^{2i\theta}, \tag{49}$$

and we obtain

$$\mu_{1,2} = i\sqrt{\frac{\beta}{1-\beta}}\; e^{\pm i\theta} \; . \tag{50}$$

Further $\mu_1 + \mu_2 = \frac{1-2\beta-2\lambda}{1-\beta}$ yields $2\lambda = 1 - 2\beta - (1-\beta)(\mu_1 + \mu_2)$, and hence

$$\lambda = 1/2 - \beta \pm i\sqrt{\beta(1-\beta)}\,\cos(\theta), \quad \theta \neq 0 \bmod(\pi). \tag{51}$$

Here it follows that $|\lambda| \leq \sqrt{(1/2-\beta)^2 + \beta(1-\beta)} = 1/2$. The same result is obtained if a finite interval is considered and Neumann boundary conditions are assumed for the central difference scheme at the right hand boundary (see appendix II).

2. Matrix Analysis

Two types of iterative schemes appear: the *standard* schemes ($0 < \beta < 1$) characterized by an amplification matrix that can be diagonalized and this is done in Appendix II, and the *pathological* schemes (β =0 or 1) associated with a defective amplification matrix, analyzed in Appendix III.

2.1. *Standard Schemes*

For $0 < \beta < 1$, the eigenvalues of the amplification operator are distinct (see Appendix II):

$$\begin{cases} \lambda_0 = 0, \\ \lambda_m = \dfrac{1}{2} - \beta + i\sqrt{\beta(1-\beta)}\cos\dfrac{m\pi}{N}, \qquad (m = 1, 2, \cdots, N-1), \end{cases} \tag{52}$$

A diagram of these eigenvalues, in which β is a parameter, is given by Figure II.2. The eigenvectors can be expressed explicitly, each one being a simple function of the corresponding eigenvalue; the matrix G_∞ is diagonalizable [1] 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:

$$\begin{aligned} \rho = \rho_{1D}(\beta) &= |\lambda_1| \\ &= \frac{1}{2}\sqrt{1 - 4\beta(1-\beta)\sin^2\frac{\pi}{N}} < \frac{1}{2}, \quad \text{and} \ \approx \frac{1}{2} \end{aligned} \tag{53}$$

[1] 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 consequence on the convergence rate since only one eigenvector is missing, as it will be explained in the next section.

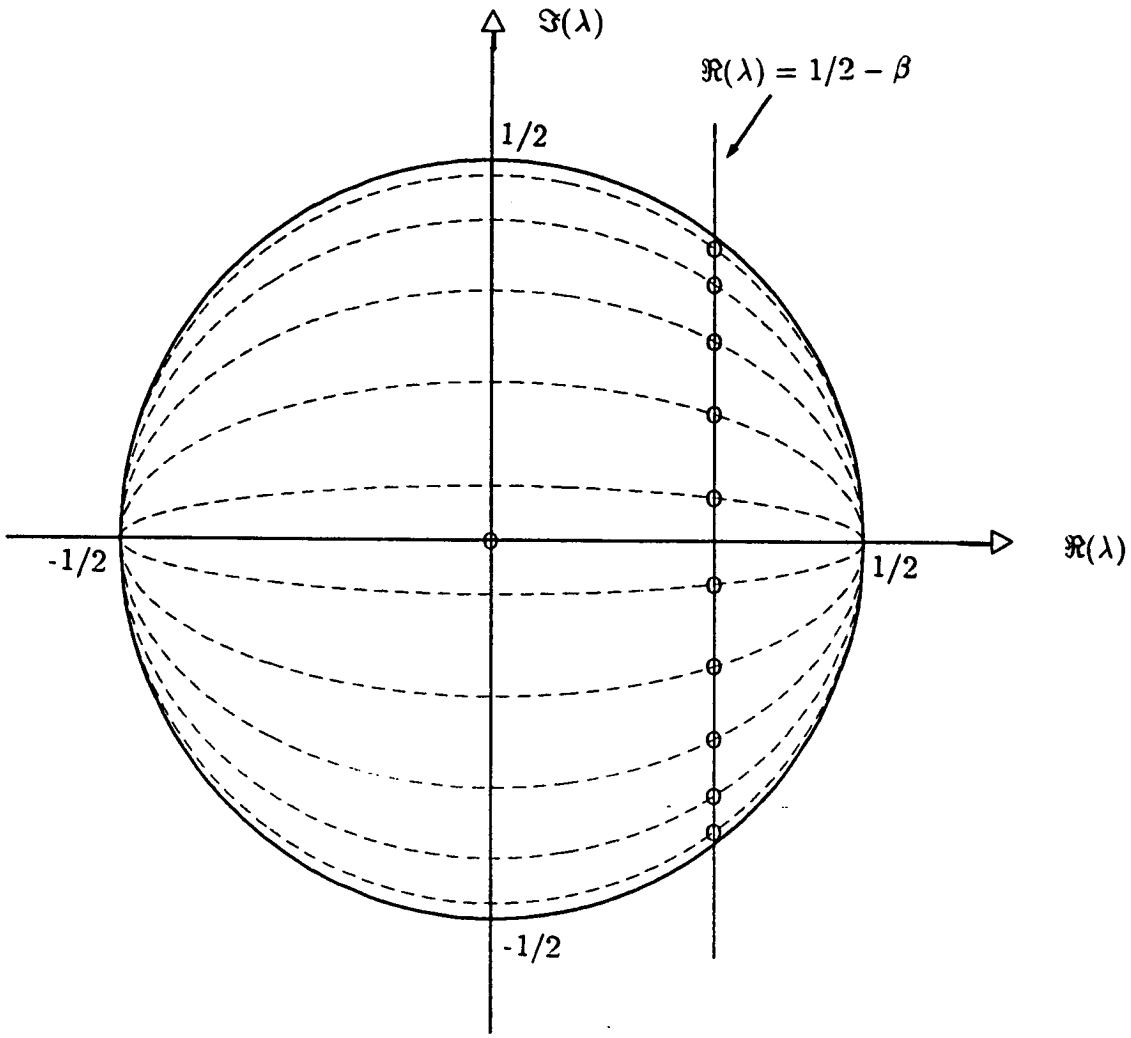


Figure II.2: Eigenvalues of the amplification operator for the 1-D non-periodic model problem with explicitly, 2nd-order β-fully upwind scheme, implicitly, first-order upwind scheme, and infinite timestep

This result is remarquable since it implies that the iterative convergence rate is strictly less than one in all cases, and essentially **independent of meshsize**. The best separation of the eigenvalues and the best condition system of eigenvectors are realized by the **half-fully upwind** scheme ($\beta = 1/2$), presumably the most robust scheme. (However, see section 3 for the two-dimensional case)

2.2. *The Pathological Schemes*

We begin this section by examining the application of the **simple explicit** first-order upwind scheme to (2):

$$u_j^{n+1} = u_j^n - c\Delta t \frac{u_j^n - u_{j-1}^n}{\Delta x} \tag{54}$$

with a Courant number,

$$\nu = \frac{c\Delta t}{\Delta x} \tag{55}$$

set equal to 1, the method reduces to the method of characteristics (which is exact):

$$u_j^{n+1} = u_{j-1}^n \tag{56}$$

If we let

$$\mathbf{u}^n = \begin{pmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_N^n \end{pmatrix} \tag{57}$$

where N denotes the number of mesh-intervals ($\Delta x = \frac{x}{N}$),

$$\mathbf{u}^{n+1} = G\mathbf{u}^n + b \tag{58}$$

where

$$G = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{pmatrix} \quad , \qquad b = \begin{pmatrix} u_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{59}$$

The amplification matrix G is defective, and although the spectral radius $\rho = 0$, the steady-state solution is not found in 1 iteration only, but in $N-1$. 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 a generalized eigenvector**

to the next. This phase extends over $N-1$ iterations, that is according to the analysis of Appendix II, the ratio of the number of missing eigenvectors, $N-1$, to $1-\rho$, where ρ is the spectral radius (here $\rho = 0$). Then only, after the error content is "flipped" into the only true eigenvector, the dissipative phase begins. Here, this phase reduces to immediate annihilation (in 1 iteration) since $\rho = 0$. In this case, the convective phase corresponds to an exact integration of the P.D.E. (with exact finite wave-propagation speed) and is no surprise.

The sketch below illustrates the convection of an error signal initially concentrated at the left-boundary point by the simple explicit method when the Courant number is equal to unity. The signal travels one meshinterval at each iteration.

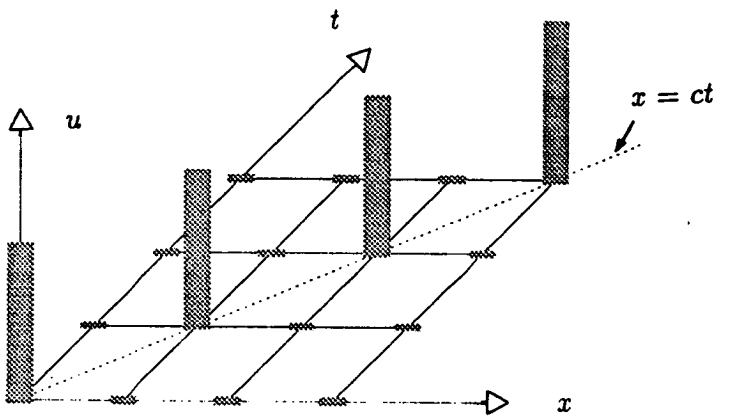


Figure II.3: Convection of an error Spike by an explicit method (CFL=1).

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

This is the case in particular for the implicit methods under study when the timestep is infinite and the upwinding parameter β is set to either limit 0 or 1. To see this, return to Figure II.2. In either limit, the spectral radius is equal to $\frac{1}{2}$; however, $N-1$ eigenvalues are identical and the corresponding eigenvectors that can be expressed in closed-form as $\mathbf{v}^m = \chi(\lambda_m)$, for the same known vector-valued function $\chi(\lambda)$, coalesce. 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, for N large, to $\frac{N}{1-\rho} = 2N$.

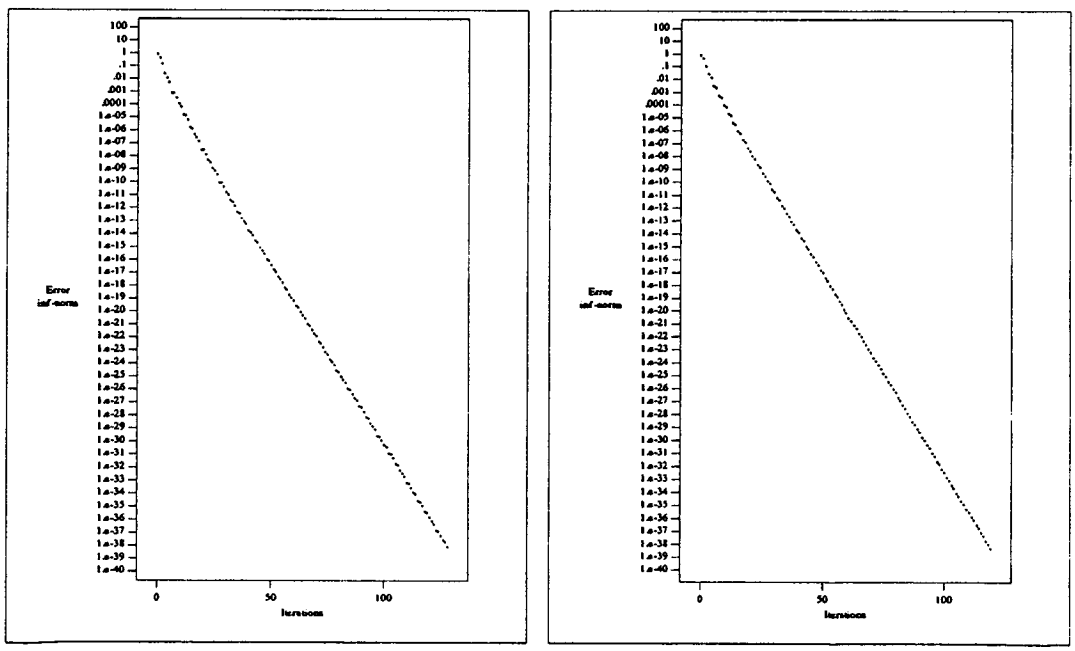
Here, the **propagation speed has no physical meaning**, the wave travelling 1 mesh-interval every 2 iterations, while the Courant number is infinite. The phenomenon is therefore a **numerical pathology**.

Finally, note that the phenomenon being related to the defective nature of the amplification matrix can only appear if some non-periodic boundary conditions are assumed. Indeed, in the periodic case, all linear operators that are constant from point-to-point are represented by circulant matrices, and all such matrices can be simultaneously diagonalized by the discrete Fourier transform.

3. Numerical Experiments

3.1. *Standard Iterative Methods*

To illustrate these results, we show some experiments made for the simple linear model problem.



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

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

Figure II.4: Convergence History of Standard Methods.

In figure II.4 we see results for iterations applied with $\beta = 1/3$ and $\beta = 1/2$ on a mesh with 100 intervals. We notice that in both cases the asymptotic rate

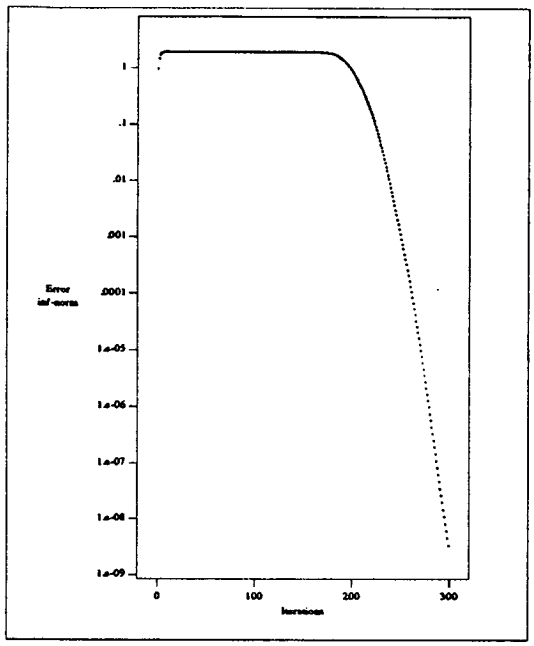
is $\rho = \lim_{n\to\infty}\rho_n = 0.5$, where $\rho_n = \|e_n\|_\infty/\|e_{n-1}\|_\infty$ and e_n is the error after n iterations. For these examples all components of the initial error e_0 were chosen randomly from the interval$(0,1)$ with a uniform distribution. This convergence rate corresponds with what is expected from the analysis.

3.2. *Iteration with Central Scheme*

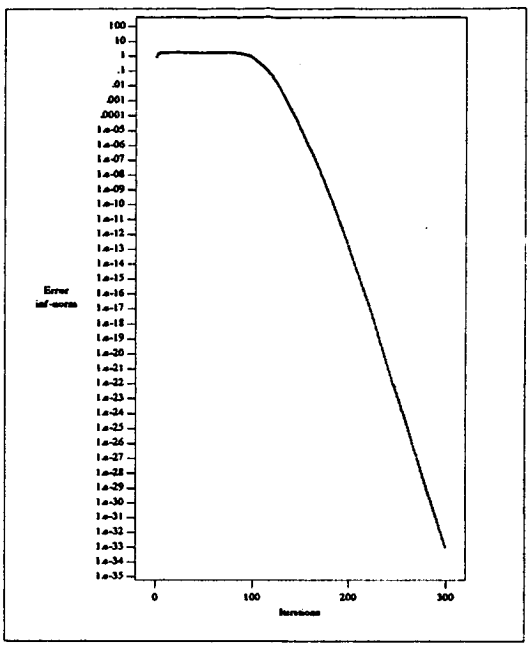
In figure II.5 we show results for an iteration with $\beta = 0$. The discretisation is on $N = 50$ or $N = 100$ nodes, and different types of initial error are used. The **oscillating initial error** e_0 is defined by the element values $e_{0,i} = (-1)^i$, $i = 1, 2, ...$; the **spike initial error** is $e_{0,1} = 1.0$, $e_{0,i} = 0.0$ for $i = 2, 3, ...$; in the **random initial error** the error at all nodes is randomly chosen, uniformly distributed in the interval $(0, 1)$.

We see that the highly oscillating error, for which $\sin(\omega h/2) \approx 1$, is the most persistent indeed: the convergence factor is approximately 1.0 for the first $2N$ iteration steps. Only after the 2N-th iteration it is the spectral radius of the amplification operator that starts to determine the convergence rate. The asymptotic convergence rate is again 0.5.

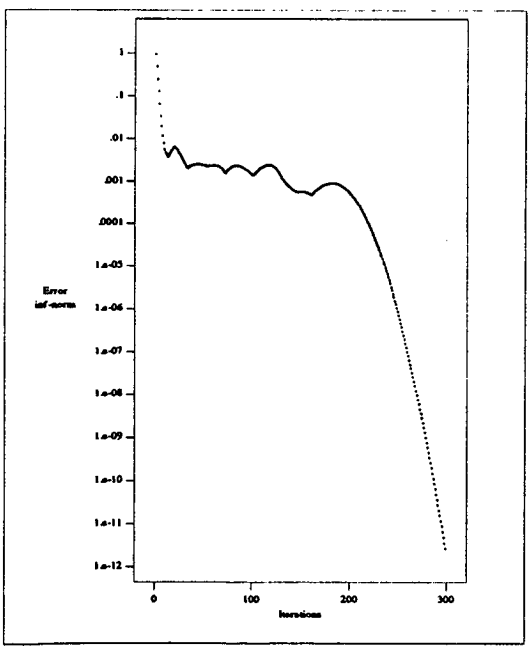
Further we see that the actual convergence is strongly dependent on the initial error components. The random initial error, which contains low as well as high frequencies, shows an impulsive start: a sharp decrease of the error in the beginning. For this pathological scheme ($\beta = 0$, as well as for the other with $\beta = 1$) we clearly recognize the pseudoconvection phase in the beginning, and after approximately 2N iterations the parabolic asymptotic phase.



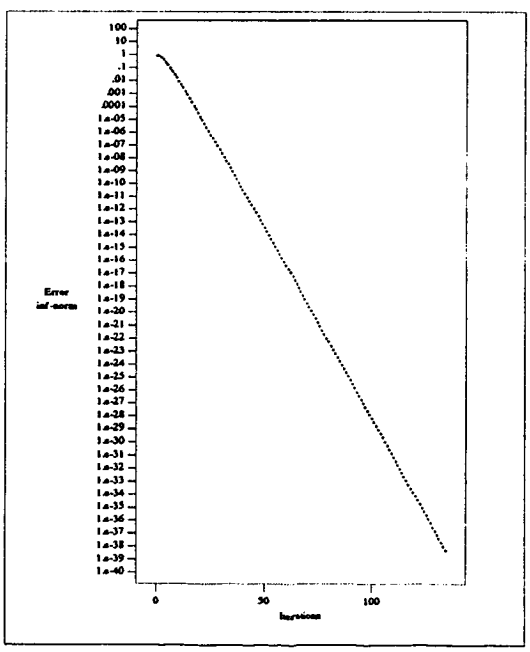
Case a: $N = 100$, $\beta = 0$, oscillating initial error.



Case b: $N = 50$, $\beta = 0$, oscillating initial error.



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



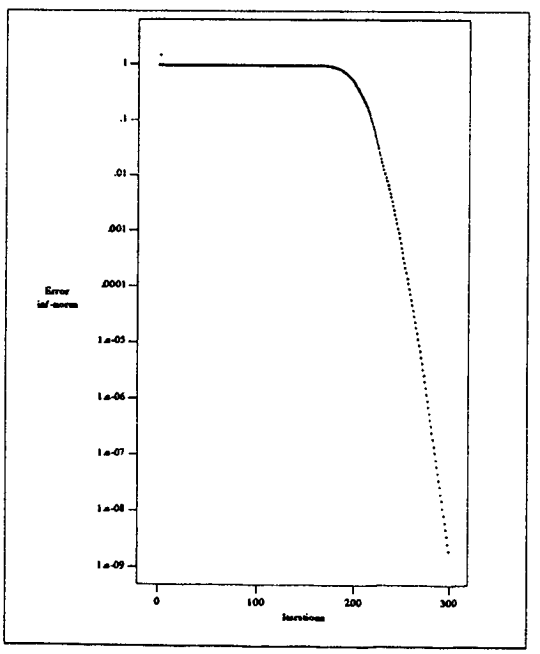
Case d: $N = 100$, $\beta = 0$, spike initial error.

Figure II.5: Convergence Histories of Iteration with Central Scheme.

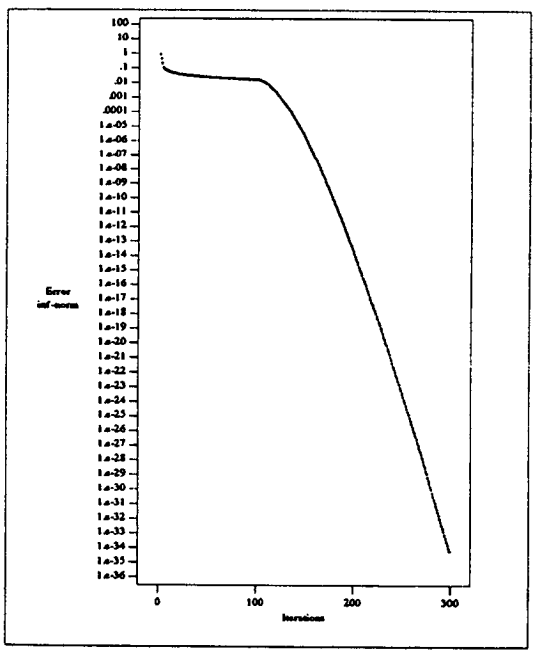
3.3. *Iteration with Fully-Upwind Scheme*

In the figures II.6, for the fully upwind scheme ($\beta = 1$), we observe a similar behaviour as for the central scheme ($\beta = 0$). But now also the spike initial error takes $2N$ iterations before it starts to converge with approximately the asymptotic convergence rate. This can be understood if we consider the evolution of the error during the iteration process.

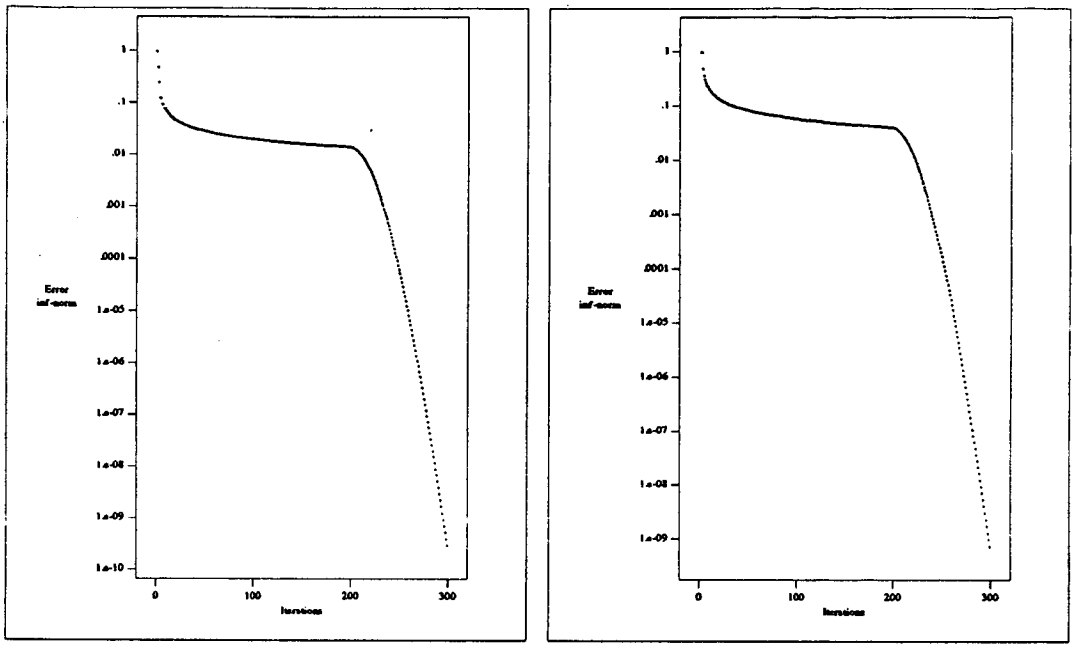
As an example of this evolution, we give in figure II.7 the behaviour of an initial oscillating error for $\beta = 0$, $\beta = 1/2$, and $\beta = 1$ on 10 nodes. The Dirichlet boundary condition was taken at the left hand side. We see that for $\beta = 0$ or 1 it takes $2N$ iterations before the error 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 with the corresponding eigenvalue is -1.



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



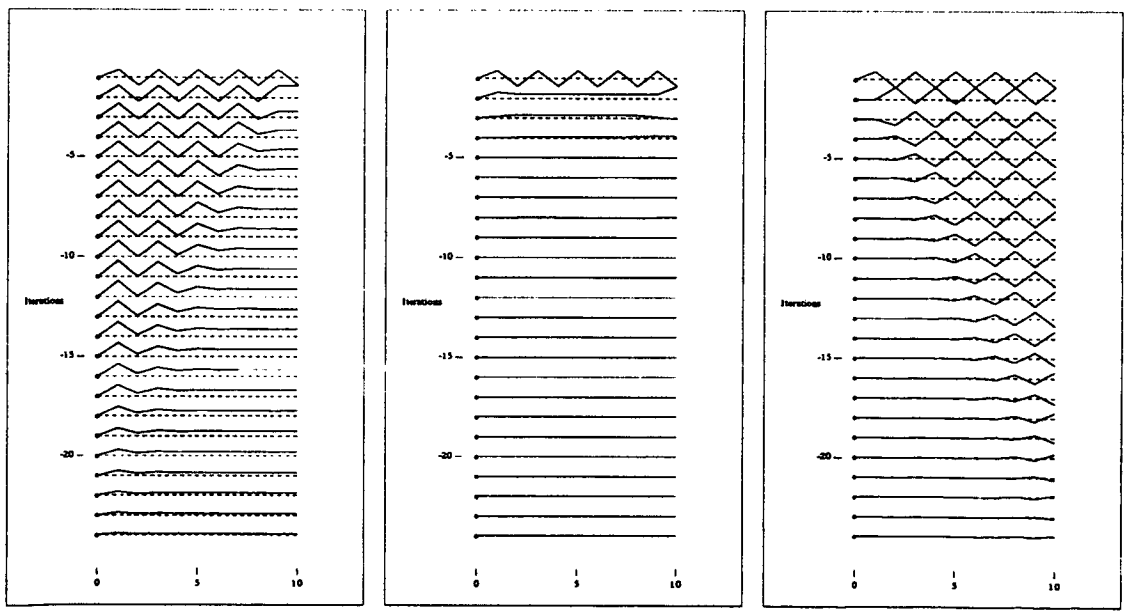
Case b: $N = 50$, $\beta = 1$, random initial error.



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

Case d: $N = 100$, $\beta = 1$, spike initial error.

Figure II.6: Convergence Histories of Iteration with Fully-Upwind Scheme.

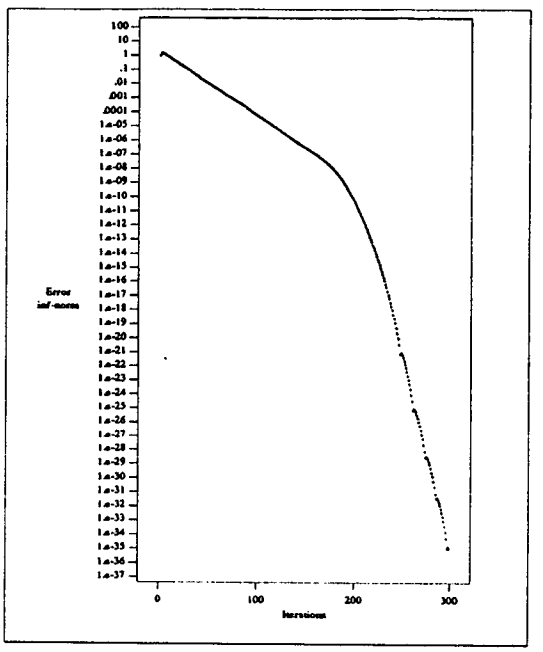


Case a: $\beta = 0$ Case b: $\beta = 1/2$ Case c: $\beta = 1$

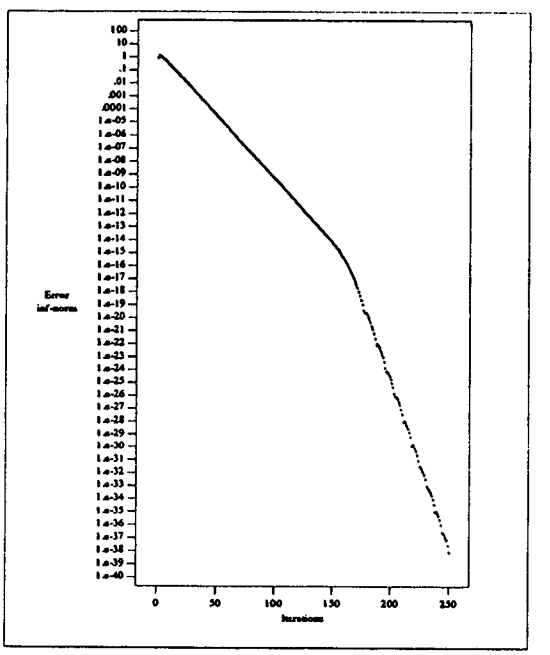
Figure II.7: Evolution of an Initially Oscillating Error ($N = 10$).

3.4. *Near-Pathological Methods*

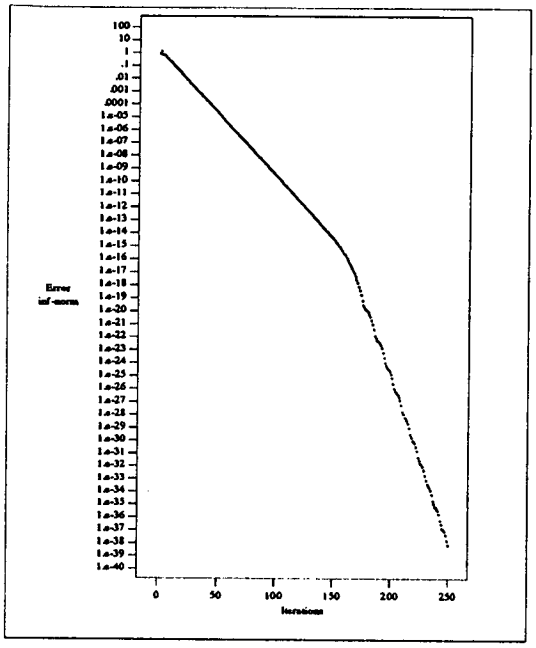
In the figures II.8 we see the convergence of the defect correction iteration for different values of β close to either $\beta = 0$ or $\beta = 1$. For these near-pathological cases, again, we clearly distinguish the pseudo-convection phase, in which the convergence rate $\rho_n = |1 - 2\beta|$ is predicted by Fourier analysis. After 2N iterations, the convergence behaviour is dominated by the asymptotic convergence rate 1/2.



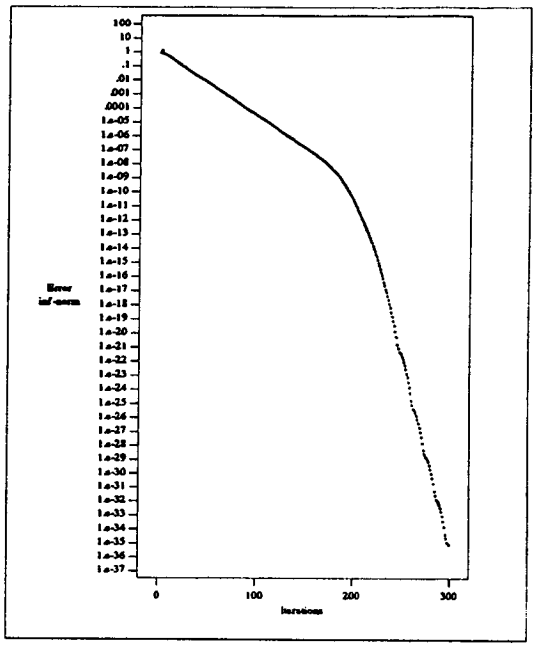
Case a: $N = 100$, $\beta = 0.05$, oscillating initial error.



Case b: $N = 100$, $\beta = 0.10$, oscillating initial error.



Case c: $N = 100$, $\beta = 0.9$, oscillating initial error.



Case d: $N = 100$, $\beta = 0.95$, oscillating initial error.

Figure II.8: Convergence Histories of Near-Pathological Methods.

3.5. *Conclusion*

In 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 with small eigenvalues are damped. For the regular schemes (β different from 0 or 1) soon an asymptotic rate of 1/2 is obtained. For the (near) pathological cases (β 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 recognize a Fourier (pseudo-convection) phase, where the error doesn't decrease for 2N iterations, and the parabolic asymptotic rate due to the large Jordan block in the eigenvalue decomposition.

III. TWO-DIMENSIONAL ANALYSIS

1. Fourier Analysis

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

$$u_t + au_x + bu_y = f \quad , \tag{60}$$

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 ([HACK, MCCM]) for the discrete central and upwind operators are given by

$$\delta_1^u = \begin{bmatrix} & 0 & \\ -a & a+b & 0 \\ & -b & \end{bmatrix}, \tag{61}$$

$$\delta_2^c = \frac{1}{2}\begin{bmatrix} & b & \\ -a & 0 & a \\ & -b & \end{bmatrix}, \tag{62}$$

$$\delta_2^u = \frac{1}{2}\begin{bmatrix} & & 0 & & \\ & & 0 & & \\ a & -4a & 3(a+b) & 0 & 0 \\ & & -4b & & \\ & & b & & \end{bmatrix}. \tag{63}$$

Similar to the previous section we define

$$\begin{aligned} \delta_2^\beta &= (1-\beta)\delta_2^c + \beta\delta_2^u = \\ &= \frac{1}{2}\begin{bmatrix} & & 0 & & \\ & & (1-\beta)b & & \\ \beta a & -(1+3\beta)a & 3\beta(a+b) & (1-\beta)a & 0 \\ & & -(1+3\beta)b & & \\ & & \beta b & & \end{bmatrix}. \end{aligned} \tag{64}$$

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

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

$$F(\delta_1) = 2ia\ e^{-i\omega_1 h_1/2}\sin(\omega_1 h_1/2) + 2ib\ e^{-i\omega_2 h_2/2}\sin(\omega_2 h_2/2) \tag{65}$$

and

$$\begin{aligned} F(\delta_2^\beta) =& 2iae^{-i\omega_1 h_1/2}S_1(C_1^2 + iS_1C_1 + 2\beta S_1^2) + \\ & 2ibe^{-i\omega_2 h_2/2}S_2(C_2^2 + iS_2C_2 + 2\beta S_2^2) \quad , \end{aligned} \tag{66}$$

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

$$\frac{F(\delta_1) - F(\delta_2^\beta)}{F(\delta_1)} = \frac{a_1 e^{-i\omega_1 h_1/2}S_1^2[\kappa S_1 - iC_1] + a_2 e^{-i\omega_2 h_2/2}S_2^2[\kappa S_2 - iC_2]}{a_1 e^{-i\omega_1 h_1/2}S_1 + a_2 e^{-i\omega_2 h_2/2}S_2}, \tag{67}$$

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

$$\begin{aligned} g(\omega) = & \left\| \frac{F(\delta_1) - F(\delta_2^\beta)}{F(\delta_1)} \right\| = \\ & \sqrt{\frac{(a_1 S_1^2(1 - 2\beta S_1^2) + a_2 S_2^2(1 - 2\beta S_2^2))^2 + 4\beta^2(a_1 S_1^3 C_1 + a_2 S_2^3 C_2)^2}{(a_1 S_1^2 + a_2 S_2^2)^2 + (a_1 S_1 C_1 + a_2 S_2 C_2)^2}}. \end{aligned} \tag{68}$$

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

In the neighborhood of the origin, for small ω_1 and ω_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

$$g(\omega) \approx \sqrt{\frac{(a_1\omega_1^2 h_1^2 + a_2\omega_2^2 h_2^2)^2}{(a_1\omega_1^2 h_1^2 + a_2\omega_2^2 h_2^2)^2 + 4(a_1\omega_1 h_1^2 + a_2\omega_2 h_2^2)^2}}. \tag{69}$$

We notice that the amplification factor becomes 1 in the neighborhood of the origin where $a_1\omega_1 h_1^2 + a_2\omega_2 h_2^2 = 0$. Introducing

$$z = \frac{1}{2}(a_1\omega_1 h_1 + a_2\omega_2 h_2) \tag{70}$$

and

$$w = \frac{1}{2}\sqrt{a_1 a_2}\ (\omega_1 h_1 - \omega_2 h_2) \tag{71}$$

yields in the neighborhood of the origin

$$g(\omega) \approx \sqrt{\frac{(z^2+w^2)^2}{(z^2+w^2)^2+z^2}} \quad . \tag{72}$$

This implies that the level curves for $g(\omega)$ are a family of circles in the (z,w)-plane through the origin, that all are tangent to the line $z = 0$, (see Figure III.1). This means that the origin is a singular point for the function $g(\omega)$, and

$$\lim_{(\omega \to 0, a_1\omega_1 + a_2\omega_2 = 0)} g(\omega) = 1, \tag{73}$$

and

$$\lim_{(\omega \to 0, a_1\omega_1 + a_2\omega_2 = c \neq 0)} g(\omega) = 0. \tag{74}$$

To further study the behaviour of $g(\omega)$ in the neighborhood of the origin, we consider the line $\omega_1 = \omega_2$. Then we find, as in section 2.1,

$$g(\omega) = |\sin(\omega h/2)| \sqrt{\cos^2(\omega h/2) + \kappa^2 \sin^2(\omega h/2)} \quad , \tag{75}$$

which formula describes for instance the levels of the equilevel ellipses of $g(\omega)$ near the origin in the ω-plane. This formula also shows that, for a given β, we never can expect a better convergence rate in the 2-dimensional case than in the 1-D case.

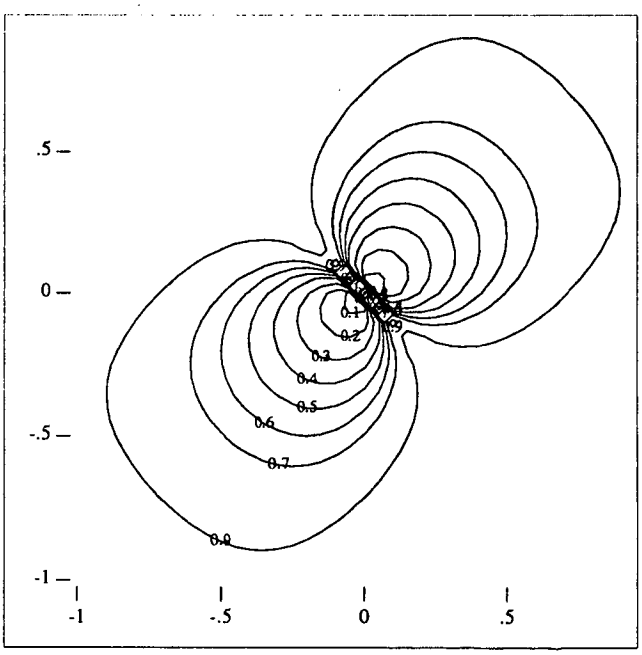
This analysis shows that for the hyperbolic problem there are always low-frequency modes u_ω for which $g(\omega) = 1$. These modes correspond with low frequencies ω for which $F(\delta_1)(\omega) = 0$, i.e. $a_1\omega_1 + a_2\omega_2 \approx 0$. These correspond with functions that are constant in the characteristic direction of the hyperbolic equation. Such modes form also the kernel 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 difference operators, and thus -in particular- for δ_1 and δ_2.

In fact, the same situation was found in the one-dimensional case, where -in contrast to 2D- the exceptional situation could not occur along the line $a_1\omega_1 + a_2\omega_2 = 0$, but only at $\omega = 0$, i.e. the origin itself.

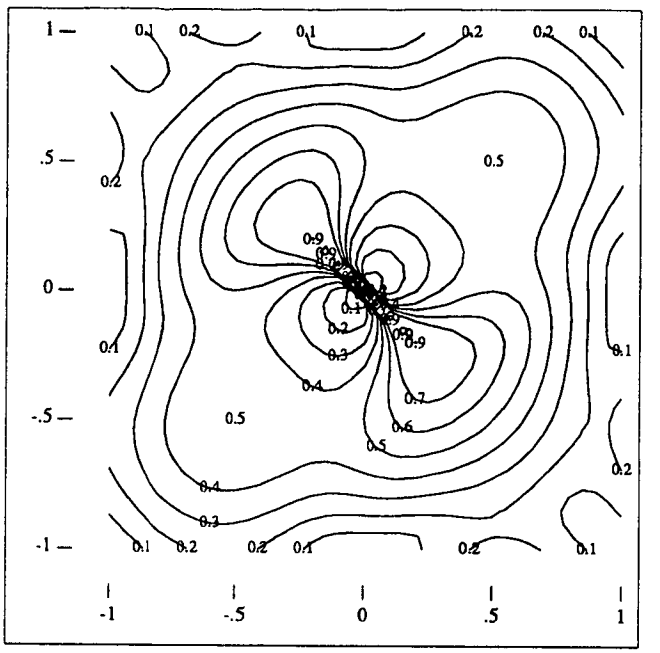
For the treatment of the the two-dimensional case, the consequence is that we cannot use the quantity $\sup_{\omega, \omega \neq 0} g(\omega)$ as a reasonable convergence factor for the iterative defect correction process. Because of the continuity of $g(\omega)$ (except at the origin), neither can we use $\sup_{\omega, a_1\omega_1 + a_2\omega_2 \neq 0} g(\omega)$. Nevertheless, the function $g(\omega)$ gives good qualitative information about the convergence behaviour. It is only more difficult to derive a useful quantitative measure, as a convergence factor. In Figure (III.1) we give equilevel plots for $g(\omega)$ for some special cases. We took $h_1 a_1 = h_2 a_2 =$

1, i.e. a convection direction of 45°, and $\beta = 0$, $\frac{1}{2}$ or 1. To get a rough quantitative impression, we can consider $\sup_{\omega,\omega_1-\omega_2=0} g(\omega)$ which, -as mentioned before- gives the same convergence rates as the one dimensional case.

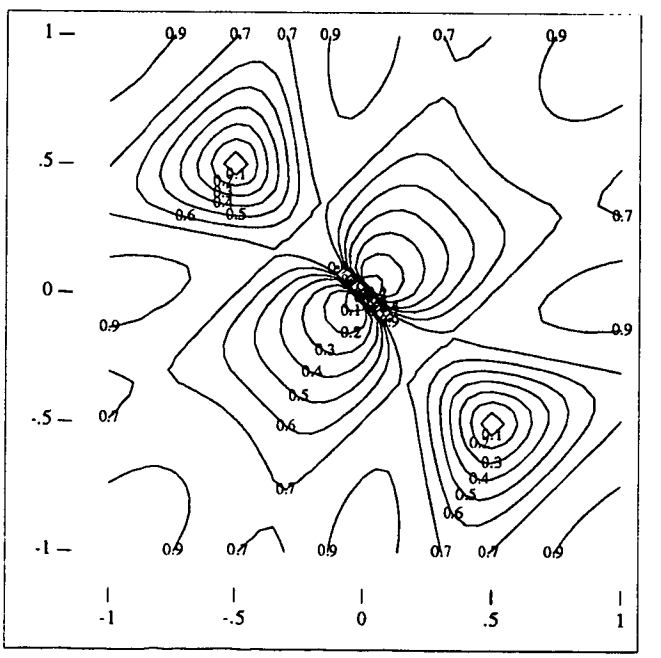
For the special case $\beta = 0$, we see that $|g(\omega)| = 1$ for all ω for which $a_1 \sin(\omega_1 h_1) + a_2 \sin(\omega_2 h_2) = 0$, and there are two branches of frequencies, both with high and with low frequencies, that will not converge. (These branches are found along the line $\omega_1 + \omega_2 = 0$ and in the left-top and the right-bottom corner of figure III.1.a.)



Case a: $\beta = 0$



Case b: $\beta = 1/2$



Case c: $\beta = 1$

Figure III.1: Level Curves of Fourier Amplification Factor ($a = b$).

2. Matrix Analysis

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

$$\begin{cases} u_t + a\,u_x + b\,u_y = 0 \quad (a > 0,\ b > 0;\ x, y, t \geq 0) \\ u(x, y, 0) = u_0(x, y) \quad \text{(specified)} \\ u(0, y, t) = u(x, 0, t) = 0 \quad (\forall\, x, y, t \geq 0) \end{cases} \tag{76}$$

Assuming a mesh of $N_x \times N_y$ gridpoints, and employing the Kronecker product notation (see APPENDIX, Section 4), a finite-difference analog of the operator

$$\mathcal{D} = a\,\frac{\partial}{\partial x} + b\,\frac{\partial}{\partial y} \tag{77}$$

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

$$\mathcal{D}_h = \mathcal{D}_x \oplus \mathcal{D}_y \tag{78}$$

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 analogs of the operators $a\,\frac{\partial}{\partial x}$ and $b\,\frac{\partial}{\partial y}$ respectively (boundary conditions included).[1] In particular, if first-order backward differences are employed we can form

$$\mathcal{D}_{h,1} = (\nu_x\,\delta^u_{x,1}) \oplus (\nu_y\,\delta^u_{y,1}) \tag{79}$$

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

$$\mathcal{D}_{h,2} = (\nu_x\,\delta^\beta_{x,2}) \oplus (\nu_y\,\delta^\beta_{y,2}) \tag{80}$$

if the same value of the upwinding parameter β is used in the two directions. From these definitions the expression of the limiting amplification matrix G_∞ (as $\Delta t \to \infty$) follows

$$G_\infty = I - (\mathcal{D}_{h,1})^{-1}\,\mathcal{D}_{h,2}. \tag{81}$$

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

$$G_\infty\,u = \lambda\,u, \tag{82}$$

[1] The subscript h refers to $N_x N_y \times N_x N_y$ matrices representing finite-difference analogs of differential operators in the two-dimensional problem.

or equivalently

$$\mathcal{D}_{h,2}\, u = (1-\lambda)\, \mathcal{D}_{h,1}\, u. \tag{83}$$

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

$$\left(A_x(\lambda) \oplus A_y(\lambda)\right) u = 0, \tag{84}$$

where

$$\begin{aligned} A_x(\lambda) &= \nu_x \left(\delta^{\beta}_{x,2} + (\lambda - 1)\, \delta^{u}_{x,1}\right), \\ A_y(\lambda) &= \nu_y \left(\delta^{\beta}_{y,2} + (\lambda - 1)\, \delta^{u}_{y,1}\right). \end{aligned} \tag{85}$$

We were not able to solve this (generalized) 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 β be arbitrary.
This is the case in particular if $N_x = N_y$ and $p = q = 1, 2, \ldots, N_x - 1$. According to (52),

$$\lambda_{x,p} = \lambda_{y,q} = \lambda \tag{86}$$

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

$$\begin{aligned} A_x(\lambda)\, u_x &= 0, \\ A_y(\lambda)\, u_y &= 0, \end{aligned} \tag{87}$$

then (84) 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 β is assumed to be the same in both directions), and all the eigenvalues of the one-dimensional problems can be considered as 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_∞ is greater or equal to the one-dimensional value for the same β (given by (53)):

$$\forall\, \beta, \qquad \rho_{2D}(\beta) \geq \rho_{1D}(\beta), \tag{88}$$

One can expect that for some values of β, 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. When 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 (88). Before this, we observe the following fact:

Remark 2: With varying λ, let the eigenvalues of the matrices $A_x(\lambda)$ and $A_y(\lambda)$ be denoted by $\left(\alpha_x(\lambda)\right)_j$,$j = 1, 2, \ldots, N_x$, and $\left(\alpha_y(\lambda)\right)_k$,$k = 1, 2, \ldots, N_y$, respectively. The eigenvalues of the Kronecker sum $A_x(\lambda) \oplus A_y(\lambda)$ are the numbers (see APPENDIX, Section 4):

$$\left(\alpha_{x\oplus y}(\lambda)\right)_{j,k} = \left(\alpha_x(\lambda)\right)_j + \left(\alpha_y(\lambda)\right)_k \tag{89}$$

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

$$\left(\alpha_x(\lambda)\right)_j + \left(\alpha_y(\lambda)\right)_k = 0 \quad , j = 1, 2, \ldots, N_x\,,\ k = 1, 2, \ldots, N_y. \tag{90}$$

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

Remark 3: The case $\beta = 1$.
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:

$$\begin{aligned} &\left(\alpha_x(\lambda)\right)_1 = \lambda\,,\ \left(\alpha_x(\lambda)\right)_2 = \left(\alpha_x(\lambda)\right)_3 = \ldots = \left(\alpha_x(\lambda)\right)_{N_x} = \lambda + \frac{1}{2} \\ &\left(\alpha_y(\lambda)\right)_1 = \lambda\,,\ \left(\alpha_y(\lambda)\right)_2 = \left(\alpha_y(\lambda)\right)_3 = \ldots = \left(\alpha_y(\lambda)\right)_{N_y} = \lambda + \frac{1}{2} \end{aligned} \tag{91}$$

This gives the following equations for λ:

$$\begin{cases} \lambda + \lambda = 0 & \text{once} \\ \lambda + (\lambda + \frac{1}{2}) = 0 & (N_x - 1) + (N_y - 1) \text{ times} \\ (\lambda + \frac{1}{2}) + (\lambda + \frac{1}{2}) = 0 & (N_x - 1)(N_y - 1) \text{ times} \end{cases} \tag{92}$$

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

$$\rho_{2D}(1) = \rho_{1D}(1) = \frac{1}{2} \tag{93}$$

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 by a routine of the NAG Library and for different combinations of the parameters (N_x, N_y), (ν_x, ν_y) and β.

According to (83), computing the eigenvalues of the matrix G_∞ if it is diagonalizable,[1] is equivalent to solving the following generalized eigenproblem

$$\mathcal{D}_{h,2}\, U = \mathcal{D}_{h,1}\, U\, (I - \Lambda) \tag{94}$$

where U is the eigenvector matrix and Λ the diagonal eigenvalue matrix. In the one-dimensional case, the matrix δ_1 was *lower* bidiagonal and the matrix δ_2 in *lower* Hessenberg form. In two dimensions, these structures are not preserved since Kronecker sums are formed; however, the nonzero elements are more numerous *below* the main diagonal, and for $\beta = 1$ these matrices are *lower* triangular. Since the first task realized by the employed numerical routine is to recast the generalized eigenproblem into one of the form $A\, x = \mu\, B\, x$, where A is in *upper*-Hessenberg form and B *upper*-triangular, it appeared more appropriate and less subject to numerical inaccuracies to transpose the problem into:

$$\mathcal{D}_{h,2}^T\, V = \mathcal{D}_{h,1}^T\, V\, (I - \Lambda) \tag{95}$$

In this formulation, the eigenvalues are the same, but the eigenvectors different. It can easily be shown that the matrices U and V are related by

$$U = \left(V^T\, \mathcal{D}_{h,1}\right)^{-1}\, N, \tag{96}$$

where N is a normalization diagonal matrix.[2] This formulation was employed to obtain the results reported on Figure III.2 where the locus of the eigenvalues of the amplification matrix G_∞ 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 β, and convection directions corresponding with $\nu_x = \nu_y$ 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, but it is represented on the Figure III.2 by a vertical dashed segment for $\beta = 1/3$ and 2/3.

In the case most different from the one-dimensional case, a convection direction of 45° with the grid ($\nu_x = \nu_y$), the cloud of new eigenvalues appears around the 1-D eigenvalue spectrum mostly to the right, and it shifts with it to the left as β increases. For small values of β, the eigenvalue of largest modulus is real positive and exterior to the disk of radius 1/2. For β larger than some value between 1/3 and

[1] If it is defective, replace the matrix Λ by a Jordan form J in the equations.

[2] Normalization is essential to define the condition number uniquely.

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 largest imaginary part, and since it belongs to the 1-D 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 contrasting with the one-dimensional case, the eigenvalue spectra of two schemes defined by values of β symmetrical with respect to 1/2 are not symmetrical with respect to the origin, and the corresponding spectral radii and condition numbers are different.

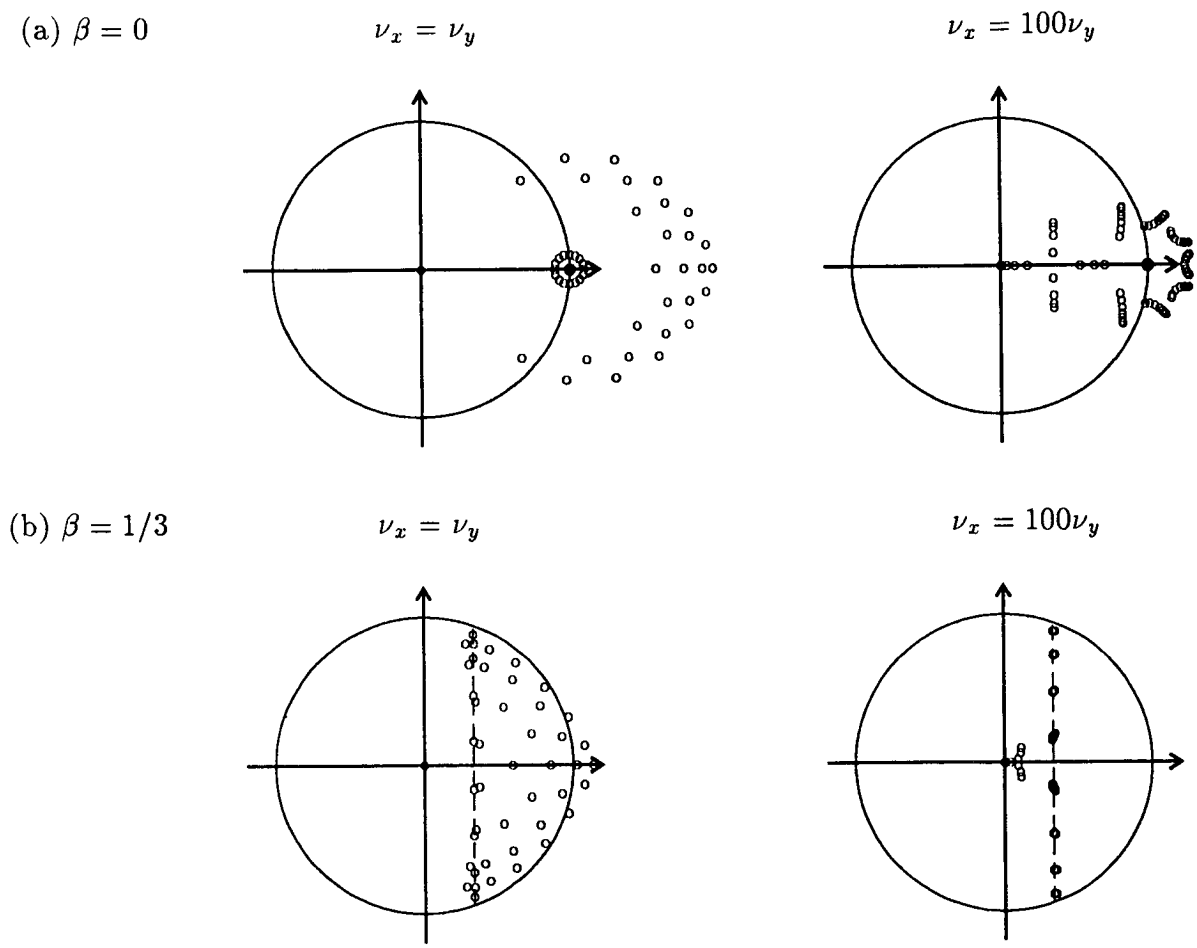
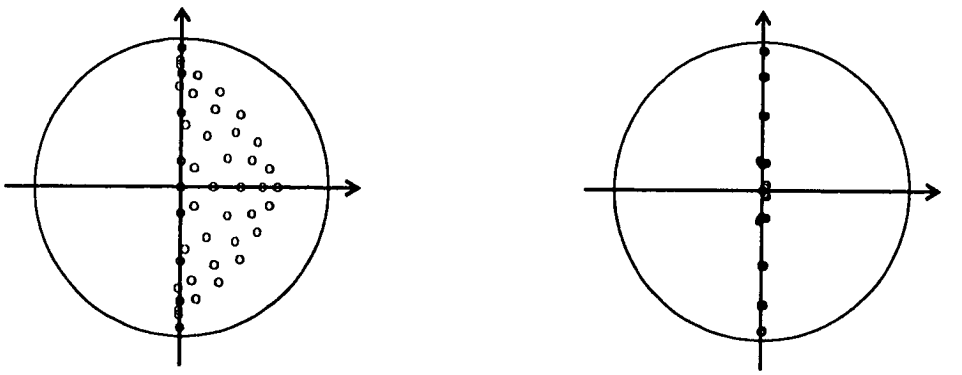
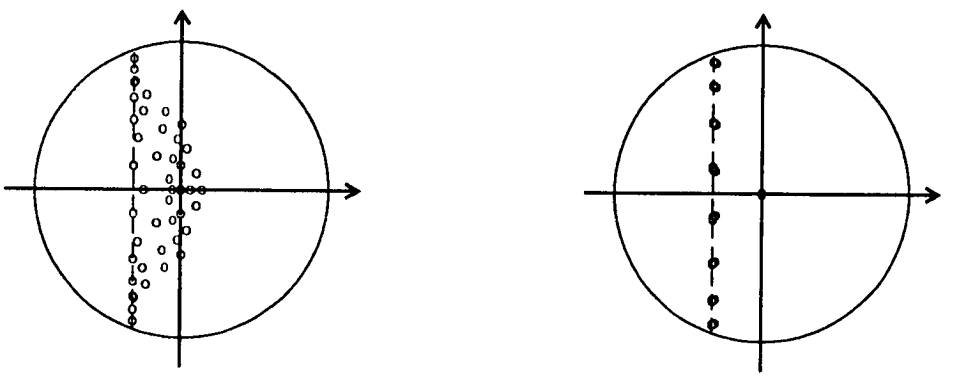


Figure III.2: Locus of the eigenvalues of the amplification matrix G_∞ of the two-dimensional model problem relatively to the circle of radius 1/2.

(c) $\beta = 1/2$ $\nu_x = \nu_y$ $\nu_x = 100\nu_y$



(d) $\beta = \frac{2}{3}$ $\nu_x = \nu_y$ $\nu_x = 100\nu_y$



(e) $\beta = 1$ $\nu_x = \nu_y$ $\nu_x = 100\nu_y$

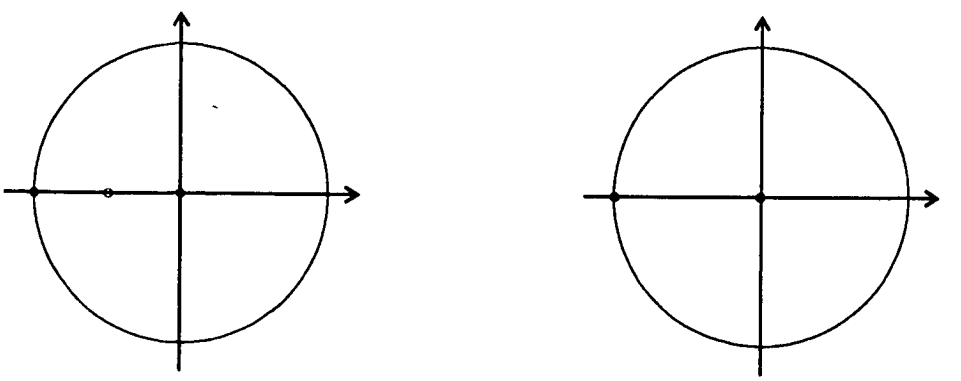


Figure III.2: Cases (c), (d) and (e).

For the cases shown on the right side of Figure III.2, $\nu_x = 100\nu_y$, and convection in the x direction dominates convection in the y direction. As a result, if $\epsilon = \nu_y/\nu_x$,

$$\begin{aligned} G_\infty &= I_x \otimes I_y - \left(\delta^u_{1,x} \otimes I_y + \epsilon I_x \otimes \delta^u_{1,y}\right)^{-1} \left(\delta^\beta_{2,x} \otimes I_y + \epsilon I_x \otimes \delta^\beta_{2,y}\right) \\ &= I_x \otimes I_y - \left(\left(\delta^u_{1,x}\right)^{-1} \otimes I_y\right) \left(\delta^\beta_{2,x} \otimes I_y\right) + O(\epsilon) \\ &= I_x \otimes I_y - \left(\left(\delta^u_{1,x}\right)^{-1} \delta^\beta_{2,x}\right) \otimes I_y + O(\epsilon) \\ &= (G_\infty)_x \otimes I_y + O(\epsilon) \end{aligned} \tag{97}$$

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 the picture, the eigenvalue spectrum is found much closer to the 1-D spectrum particularly for values of β 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.

2.2. *Spectral Radius, ρ*

We now examine more closely the behaviour of the spectral radius ρ which is given in Tables 1 to 5 for various combinations of the parameters. In Tables 1-2, all possible situations with respect to the parities of the numbers N_x and N_y were examined; evidently, these parities have no significant effect on the spectral radius.

In Tables 1, 2 and 3 experiments with equal, comparable and very different parameters ν_x, ν_y, and for meshes of comparable sizes are reported.

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

$$\rho_{2D}(\beta) \begin{cases} > \rho_{1D}(\beta) & \text{if } \beta < \bar{\beta} < 1/2 \\ = \rho_{1D}(\beta) & \text{if } \beta \geq \bar{\beta} \end{cases} \tag{98}$$

We conjecture that $\bar{\beta}$ should converge to $1/2$ as $N_x = N_y \to \infty$, but no experiment was made to confirm this point. Consequently, the maximum asymptotic convergence rate is achieved for $\beta = 1/2$:

$$\min_{\beta \in [0,1]} \rho_{2D}(\beta) = \rho_{2D}(1/2) = \rho_{1D}(1/2) = \frac{1}{2} \cos \frac{\pi}{N_x} \tag{99}$$

However, since the spectral radius is nearly invariant as β increases from $1/2$, it might be judicious in practice to set β to a slightly larger value to cluster the spectrum of eigenvalues around 0.

Table 1: Spectral Radius ($\nu_x = \nu_y$)			
	$N_x \times N_y = 9 \times 9$	10×9	10×10
$\beta = 0$	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*

* 1-D theoretical value

Table 2: Spectral Radius ($\nu_x = 2\nu_y$)				
	$N_x \times N_y = 9 \times 9$	9×10	10×9	10×10
$\beta = 0$	0.93387	0.93596	0.93750	0.93938
0.1	0.83216	0.83417	0.83346	0.83548
1/3	0.56350	0.56456	0.56635	0.56741
1/2	0.46985*	0.47175	0.47364	0.47553*
2/3	0.47329*	0.47497	0.47665	0.47831*
0.9	0.48936*	0.49001	0.49067	0.49133*
1	0.5*	0.5*	0.5*	0.5*
0.49	0.46986*	0.47176	0.47365	0.47554*
0.51	0.46986*	0.47176	0.47365	0.47554*

* 1-D theoretical value

In the cases of Table 3, $\nu_x = 100\nu_y$ and convection is preponderant in the x direction. As a result, the algebraic system behaves more like in one dimension, and for $0 \leq \beta < \bar{\beta}$, since ϵ is small but finite, the spectral radius although different from the 1-D theoretical value, is found closer to it than the analogous value in the first column of Table 1 or 2.

Table 3: Spectral Radius ($\nu_x = 100\nu_y$, $N_x \times N_y = 9 \times 9$)	
β	ρ
0	0.64278
0.1	0.49869
1/3	0.47653
1/2	0.46985*
2/3	0.47329*
0.9	0.48936*
1	0.5*
0.49	0.46986*
0.51	0.46986*

* 1-D theoretical value

In the one-dimensional case, we have found that as $N_x \to \infty$, the spectral radius ρ tends to 1/2 for all β. Contrasting with 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 Tables 4 and 5, as N_x and N_y increase, the spectral radius for $\beta = 1/2$ remains equal to the 1-D theoretical value which is bounded by 1/2, while this bound is violated by the third-order method, particularly if the parameters are of comparable sizes (e.g. as in Table 4), since for $\nu_x \gg$ or $\ll \nu_y$ (e.g. as in Table 5) the problem is more nearly one-dimensional.

Table 4: Spectral Radius ($\nu_x = \nu_y$)		
	$\beta = 1/3$	$\beta = 1/2$
$N_x \times N_y = 5 \times 5$	0.52253	0.40451*
10×10	0.57235	0.47553*
20×20	0.58423	0.49384*
30×30	0.58633	0.49726*

* 1-D theoretical value

Table 5: Spectral Radius ($\nu_x = 100\nu_y$)		
	$\beta = 1/3$	$\beta = 1/2$
$N_x \times N_y = 5 \times 5$	0.42037	0.40451*
10×10	0.48145	0.47553*
20×20	0.49726	0.49384*
30×30	0.50119	0.49726*

* 1-D theoretical value

2.3. *Condition number, κ*

In the last set of experiments of this section, an attempt was made to compute along with the spectral radius ρ, the condition number κ of the eigenvector matrix U:

$$\kappa = \kappa_U = \| U \|_2 \| U^{-1} \|_2 = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}} (U^* U)} \tag{100}$$

where the matrix U^* is the adjoint of the matrix U. This parameter becomes infinite when approaching a case where the matrix G_∞ is defective. Hence we expect that such cases can be detected by the increase of this parameter.

The previous formulation could provide the matrix V easily, κ could have been computed by forming the matrix U according to (96). Such computation which involves an inversion, would have probably been inaccurate, and it was rejected. Instead two other methods were employed.

The first procedure consisted of solving the eigenproblem directly as originally formulated. This route is simple but bears some risk, as previously mentioned, since for β near 1, attempt is made to transform a near defective *lower* triangular matrix into a near *upper* triangular matrix.

In the second procedure, one first considers the following permutation matrix

$$\Pi = \begin{pmatrix} & & & 1 \\ & & 1 & \\ & \cdot^{\cdot^{\cdot}} & & \\ & 1 & & \\ 1 & & & \end{pmatrix} \tag{101}$$

The matrix Π is symmetric and involutive ($\Pi^{-1} = \Pi^T = \Pi$). Substituting

$$U = \Pi W \tag{102}$$

into (94) and premultiplying by the matrix Π yield the following equivalent formulation:

$$\tilde{\mathcal{D}}_{h,2} W = \tilde{\mathcal{D}}_{h,1} W (I - \Lambda) \tag{103}$$

where the matrices

$$\begin{cases} \tilde{\mathcal{D}}_{h,1} = \Pi \mathcal{D}_{h,1} \Pi \\ \tilde{\mathcal{D}}_{h,2} = \Pi \mathcal{D}_{h,2} \Pi \end{cases} \tag{104}$$

are similar to the matrices $\mathcal{D}_{h,1}$ and $\mathcal{D}_{h,2}$ respectively, and obtained from them by reversing the order of both rows and columns.[1] Since the matrix Π is orthogonal ($\Pi^T \Pi = \Pi^2 = I$), it preserves the condition number; hence κ is computed as

$$\kappa = \kappa_W \tag{105}$$

The results are given for a case of convection across the grid ($\nu_x = \nu_y$) and a case of convection almost along the grid ($\nu_x = 100\nu_y$) on Tables 6 and 7 respectively. In each column, the numbers given by the two methods are separated by a hyphen, and '-id.' indicates the two methods gave the same answer.

In all the cases except $\beta = 1$, the same value of the spectral radius was obtained by the two methods, confirming the third column of Tables 1 and 3.

We now examine the condition numbers. The symbol '∞' indicates that the smallest eigenvalue found by the routine is either 0 or a very small negative number due only to loss of precision.

In the two cases (Tables 6 and 7), the number $N_x = N_y = 9$ is odd[1] and the results tend to indicate that only $\beta = 0$ and $\beta = 1$ result in a defective amplification matrix. A minimum condition number is achieved near $\beta = 1/2$, at $\beta \approx 0.46$. (Presumably, this value of β distinct from $\bar{\beta}$ should also converge to $1/2$ as $N_x = N_y \to \infty$, but no experiment was made to confirm this point.) We note that the two methods produce exactly the same figures in some cases, in other cases they only yield the same order of magnitude. This indicates that the evaluation of κ is itself ill-conditioned. In view of the structure of the matrices involved and of the values found for the spectral radius for $\beta = 1$, the more reliable method is believed to be the second for $\beta \geq 1/2$, and because of symmetry, the first for $\beta \leq 1/2$.

---

[1] These matrices are (the negative of) the forward-difference operators that would be formed to study the *left* quarter-plane problem, defined over $x < 0$ with the Dirichlet condition at $x = 0$. Hence, not surprisingly, this problem is completely symmetrical, and the same eigenvalues and condition number are associated with it.

[1] For the one-dimensional problem, it was observed that for N_x even and $\beta = 1/2$, the matrix G_∞ 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 ν_x/ν_y sufficiently large, the same situation should be approached according to (97) for $\beta = 1/2$ (this was verified experimentally for $N_x = N_y = 10$) with no more consequence on the iterative process, the 2×2 sub-block $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ appearing N_y times in the Jordan reduced form of the matrix G_∞.

For $\nu_x = 100\nu_y$, the two methods produce the same condition numbers in all but one cases: in this nearly one-dimensional case the computation of κ is easier.

Table 6: Spectral Radius and Condition Number
($\nu_x = \nu_y$, $N_x \times N_y = 9 \times 9$)

β	ρ	κ
0	0.98693-id.	∞*-id.
0.1	0.87353-id.	0.1225×10^{9}-0.1038×10^{9}
1/3	0.56854-id.	0.3697×10^{5}-id.
1/2	0.46985*-id.	0.3550×10^{4}-0.8928×10^{4}
2/3	0.47329*-id.	0.6089×10^{5}-id.
0.9	0.48936*-id.	0.1281×10^{8}-0.5998×10^{8}
1	0.54632†-0.5*	∞*-id.
0.45	0.47016*-id.	0.6126×10^{4}-id.
0.46	0.47005*-id.	0.5983×10^{4}-id.
0.47	0.46996*-id.	0.6008×10^{4}-id.

* 1-D theoretical value
† evidently erroneous value

Table 7: Spectral Radius and Condition Number
($\nu_x = 100\nu_y$, $N_x \times N_y = 9 \times 9$)

β	ρ	κ
0	0.64278-id.	∞*-id.
0.1	0.49869-id.	0.5853×10^{7}-id.
1/3	0.47653-id.	0.4063×10^{5}-id.
1/2	0.46985*-id.	0.1299×10^{5}-id.
2/3	0.47329*-id.	0.4085×10^{5}-id.
0.9	0.48936*-id.	0.4922×10^{8}-0.4933×10^{8}
1	0.51344†-0.5*	∞*-id.
0.45	0.47016*-id.	0.7948×10^{4}-id.
0.46	0.47005*-id.	0.7787×10^{4}-id.
0.47	0.46996*-id.	0.7989×10^{4}-id.

* 1-D theoretical value
† evidently erroneous value

We close this section by the following remark:

Remark 4: The case of wavespeeds of arbitrary signs.
Suppose that instead of (76) we had treated the case of wavespeeds of opposite signs, as for example in the following alternate model problem:

$$\begin{cases} u_t - a\,u_x + b\,u_y = 0 \quad (a > 0,\ b > 0;\ x \le 0,\ y,t \ge 0) \\ u(x,y,0) = u_0(x,y) \quad \text{(specified)} \\ u(0,y,t) = u(x,0,t) = 0 \quad (\forall\, x \le 0,\ y,t \ge 0) \end{cases} \tag{106}$$

This continuous problem is trivially similar to the original one, since it identifies to it if x is changed in $-x$. Naturally the corresponding algebraic formulations are also similar.[1]

Hence, the analysis of this section applies to cases of wavespeeds a and b of arbitrary signs, so long as the side of the applied Dirichlet conditions and the directions of upwinding are defined consistently to construct a well-posed algebraic formulation.

3. Numerical Experiments on 2-D Wave Equation

Numerical experiments were made for the two-dimensional linear wave equa-

[1] For this let ϖ be the permutation matrix having the same structure as the matrix Π but the smaller dimension $N_x \times N_x$. A well-posed algebraic formulation is now constructed by employing forward differences in the x direction (and still backward differences in the y direction). This is equivalent to replacing the matrices $\mathcal{D}_{h,1}$ and $\mathcal{D}_{h,2}$ by the following ones:

$$\mathcal{D}'_{h,1} = (\nu_x\, \varpi\, \delta^u_{x,1}\, \varpi) \oplus (\nu_y\, \delta^u_{y,1})$$
$$\mathcal{D}'_{h,2} = \left(\nu_x\, \varpi\, \delta^\beta_{x,2}\, \varpi\right) \oplus \left(\nu_y\, \delta^\beta_{y,2}\right)$$

to approximate $-a\,\frac{\partial}{\partial x} + b\,\frac{\partial}{\partial y}$. Clearly, since

$$(\varpi \otimes I_y)^{-1} = \varpi^{-1} \otimes I_y^{-1} = \varpi \otimes I_y$$

the matrices

$$\mathcal{D}'_{h,1} = (\varpi \otimes I_y)^{-1}\, \mathcal{D}_{h,1}\, (\varpi \otimes I_y)$$
$$\mathcal{D}'_{h,2} = (\varpi \otimes I_y)^{-1}\, \mathcal{D}_{h,2}\, (\varpi \otimes I_y)$$

are similar to the original matrices $\mathcal{D}_{h,1}$ and $\mathcal{D}_{h,2}$ respectively. Consequently, the eigenvalues and the spectral radius ρ are the same, and since the permutation matrix $\varpi \otimes I_y$ is orthogonal, the condition number κ also.

tion, for a range of angles for the convection direction $\arctan(a_1/a_2)$, 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.

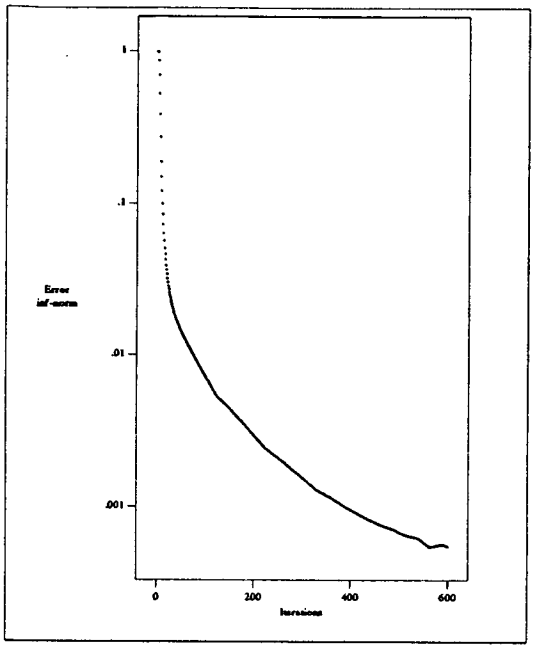
In this section we show the convergence behaviour for the linear wave equation for which the convection direction is skew to the grid, $a_1 = a_2 = 1$ (a convection angle of 45^o). We use a square grids with $N \times N$ ($N = 10, 20, 40, 80$) gridpoints. Similar initial errors were used as for the one-dimensional case. The **oscillating initial error** is defined by $e_{ij} = (-1)^{ij}$, $i, j = 1, 2, ...$; the **spike initial error** is $e_{11} = 1.0$, $e_{ij} = 0.0$ for $i + j > 2$; in the **random initial error** the error at all nodes is randomly chosen, uniformly distributed in the interval $(0, 1)$.

First, in figure III.3, we show results for an iteration applied with the central scheme ($\beta = 0$). We see that the iteration doesn't converge. For a random error and the spike 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 with 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.

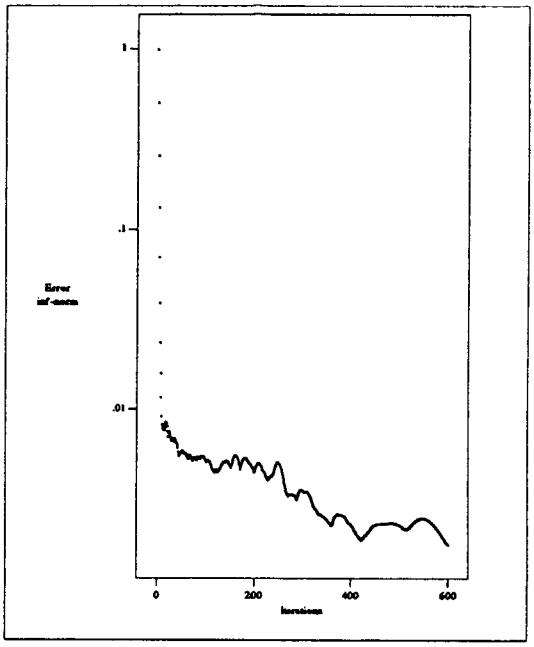
The highly oscillating error increases in the first few steps (operator norm greater than one) and then remains constant for $O(N)$ iterations. Thereafter the error first decreases a few orders of magnitude before it behaves like e.g. the random error (Figure III.3.c).

For the non-pathological cases $\beta = 1/3$ or $\beta = 1/2$ we observe an asymptotic rate of convergence corresponding with the values as given in section 3.2.2 table 4. For $\beta = 1/2$ (Figure III.5) the true rate is hard to observe on finer meshes because an additional effect is seen. Viz. 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.

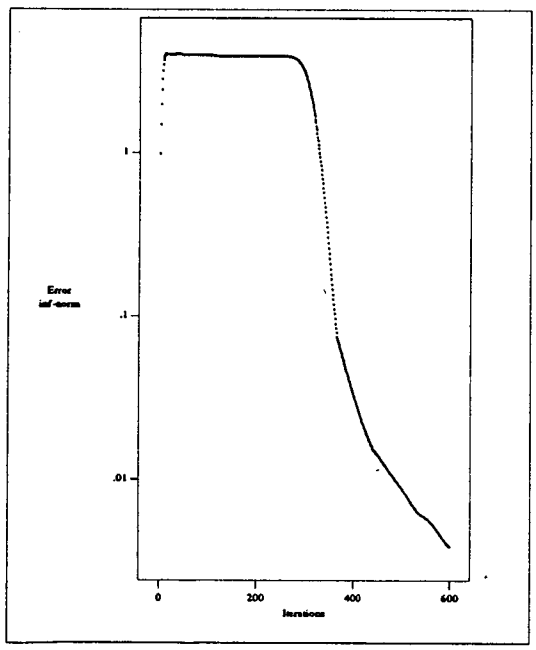
For the fully upwind scheme ($\beta = 1$), in the figures III.6 and III.7 we see that there is again a pseudo convection phase for $O(N)$ iterations. In figure III.7, for the simpler initial errors, this phase extends over approximately $4N$ iterations. For the random initial error (figure.III.6) we see that the situation is more complex. Here we also recognize a component with a pseudo convection phase of $2N$, and for the 80×80 mesh even a few additional components of which the length of such phase cannot clearly be identified. For the coarser meshes we clearly recognize the parabolic asymptotic phase at the end.



Case a: spike initial error.

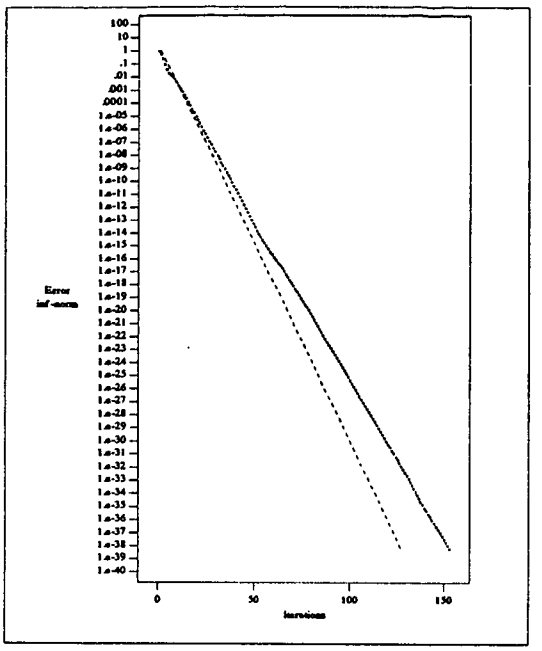


Case b: random initial error.

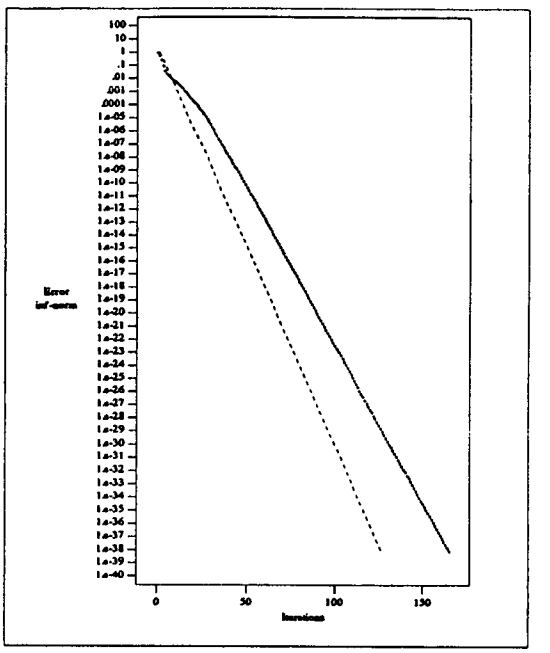


Case c: oscillating initial error.

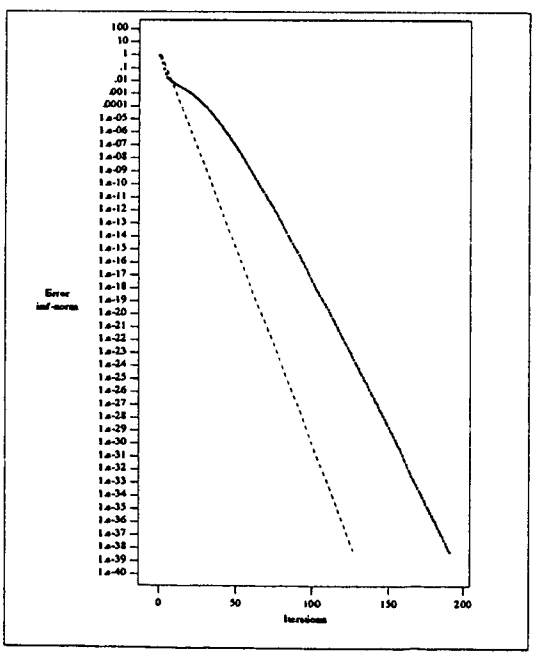
Figure III.3: Convergence Histories of Iteration with Central Scheme
The dashed line corresponds with a convergence rate 1/2.
($\beta = 0$, 2D computation over 80×80 mesh).



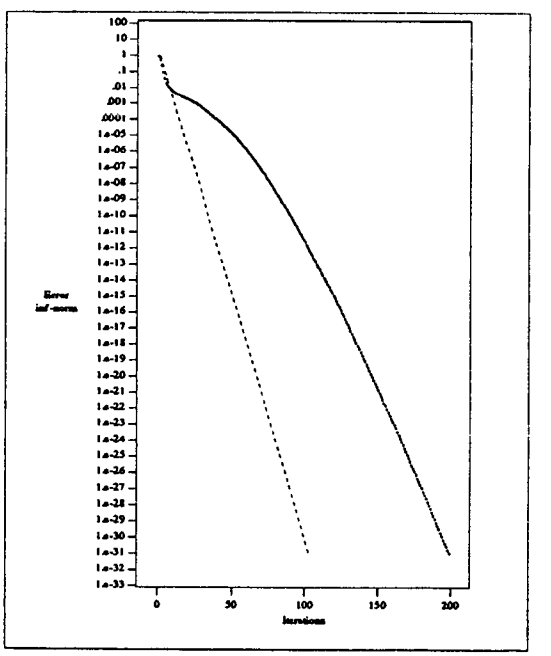
Case a: 10 × 10 mesh



Case b: 20 × 20 mesh

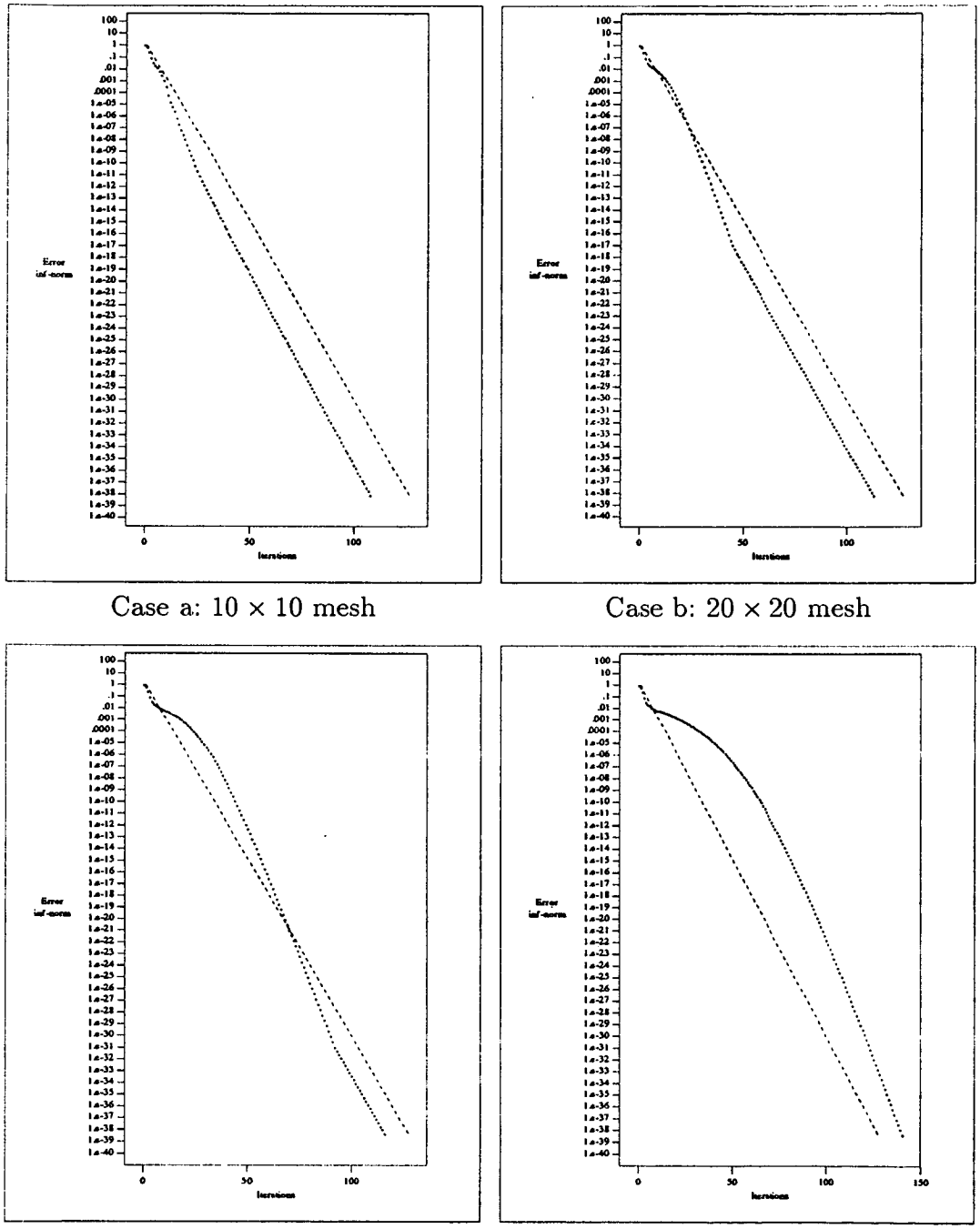


Case c: 40 × 40 mesh



Case d: 80 × 80 mesh

Figure III.4: Convergence Histories of Iteration with Upwind-Biased Scheme
The dashed line corresponds with a convergence rate 1/2
($\beta = 1/3$, random initial error).



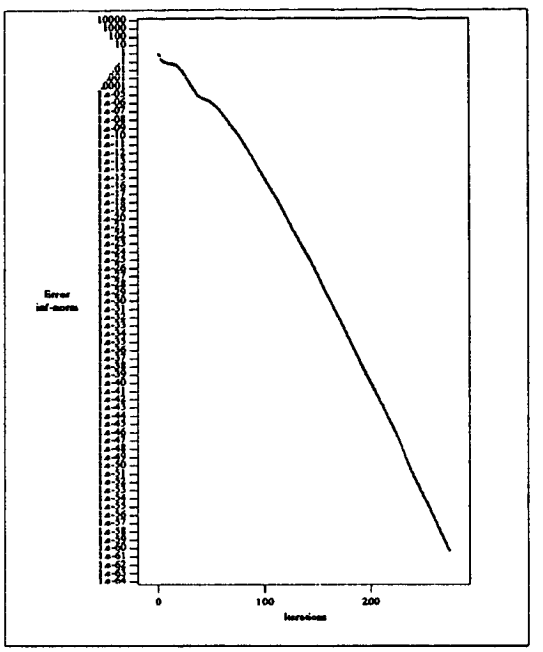
Case a: 10×10 mesh

Case b: 20×20 mesh

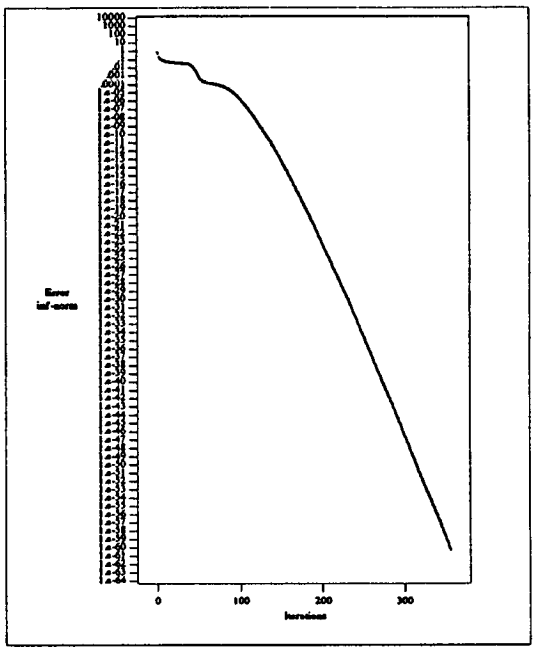
Case c: 40×40 mesh

Case d: 80×80 mesh

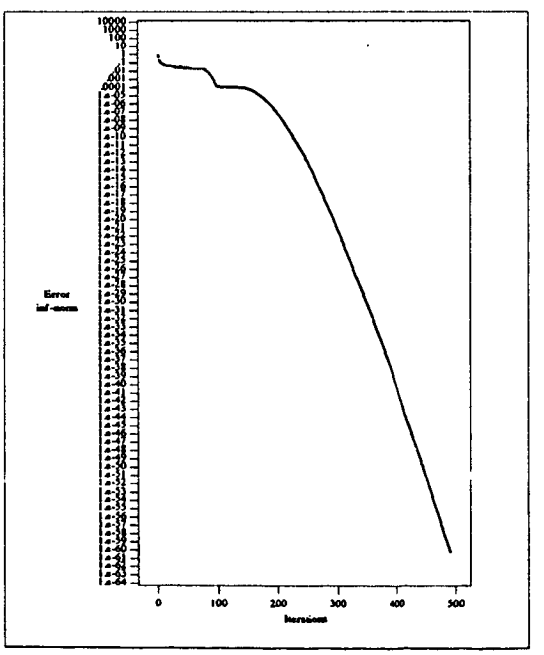
Figure III.5: Convergence Histories of Iteration with Fromm's Scheme
The dashed line corresponds with a convergence rate 1/2
($\beta = 1/2$, random initial error).



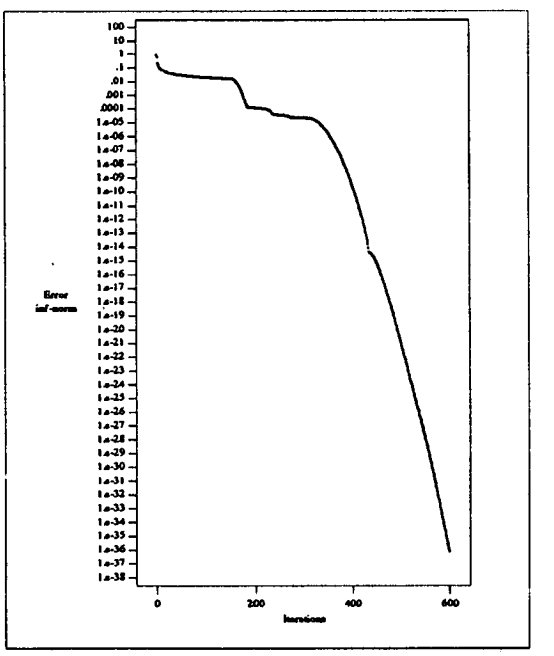
Case a: 10 × 10 mesh



Case b: 20 × 20 mesh

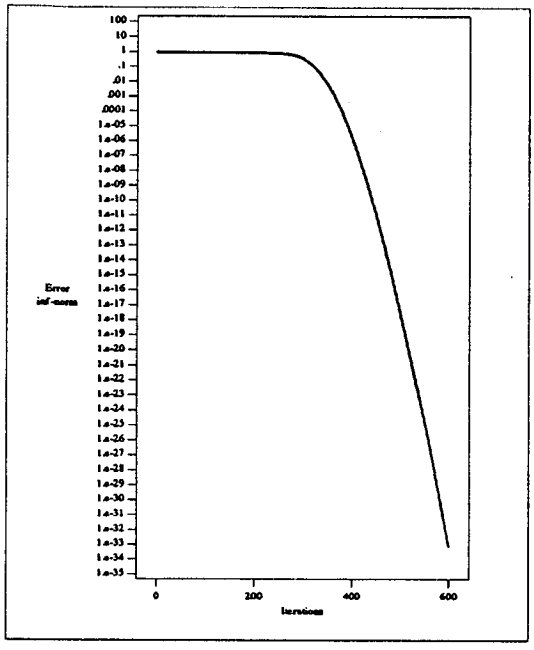


Case c: 40 × 40 mesh

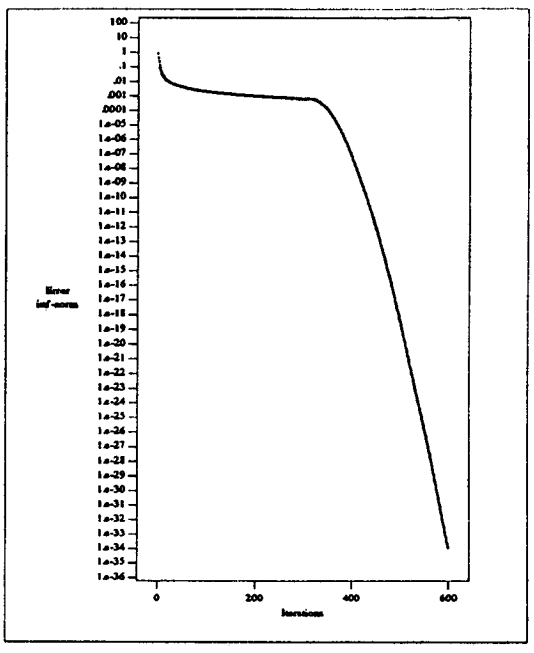


Case d: 80 × 80 mesh

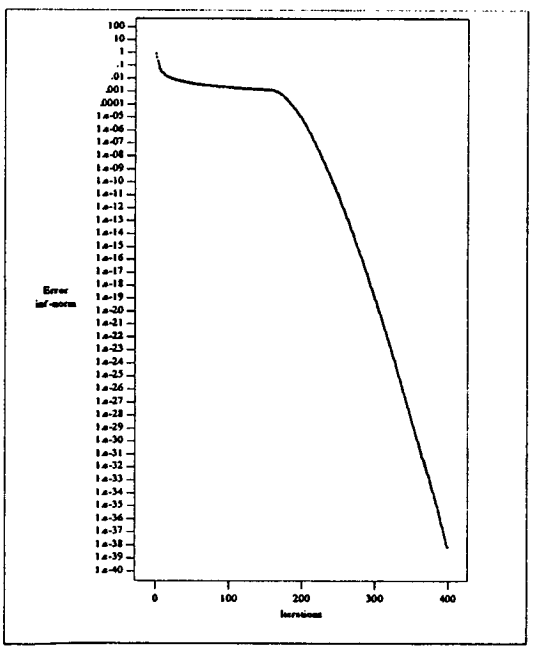
Figure III.6: Convergence Histories of Iteration with Fully-Upwind Scheme ($\beta = 1$, random initial error).



Case a: oscillating initial error, 80×80 mesh.



Case b: spike initial error, 80×80 mesh.



Case c: spike initial error, 40×40 mesh.

Figure III.7: Convergence Histories of Iteration with Fully-Upwind Scheme ($\beta = 1$, other initial errors).

In summary, when using the slightly inconsistent formulation of the implicit upwind scheme, in which the implicit preconditioner is based on only first-order differencing while a second-order partially-upwind approximation is constructed explicitly, it is not recommended to use the central-differencing scheme ($\beta = 0$) explicitly, or the fully-upwind scheme ($\beta = 1$) explicitly because both result in defective methods with pathological iterative convergence. Preferably, one should use the half-fully upwind scheme explicitly ($\beta = 1/2$) to realize the best separation of the eigenvalues, and presumably the least condition number of the matrix of eigenvectors. The upwind biased scheme ($\beta = 1/3$) will achieve a higher steady-state accuracy. However, it may be slightly less robust.

IV. EULER FLOW EXPERIMENTS

1. Introduction

Several illustrative experiments have been carried out in a much more complex context than that of the analysis. In this chapter, the Euler equations are solved in 2-D by a method of Hemker-Spekreijse-Koren [HRSP, HRKO]. It is a finite volume method on a structured quadrilateral grid. It makes use of Osher's approximate Riemann solver for the numerical flux function, both in the a first order method, and in the second order method. The second order approximation is computed by the MUSCLE approach. In this setting a parameter β can be introduced, completely analogous to the β used in the previous chapters. In the experiments in the following sections, no flux limiters were applied. For more details about the method, we also refer to [BKOR] To remain as close to the theory as possible, we did not apply a flux-limiter in these experiments.

The first order discrete equations are solved by a nonlinear multigrid method. It employs of 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 solves the discrete system to a high degree of accuracy. In the experiments shown, 3 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 2 through 5 FAS V-cycles. All initial estimates were obtained by interpolation from a first order accurate solution on a coarser grid.

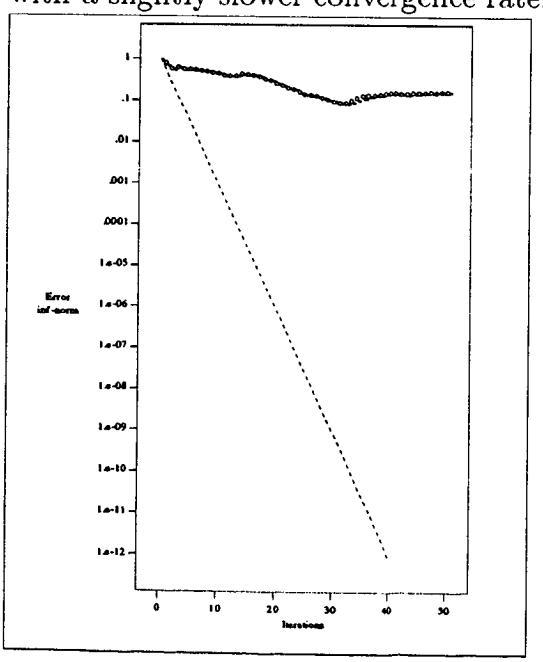
In the sections IV.2 through IV.5 we show results for flow over the standard NACA0012 airfoil. Section IV.2 treats a subsonic flow with the Mach number at infinity $M_\infty = 0.63$ and the angle of attack $\alpha = 2.0^o$. This is a smooth flow where no special effects are to be expected, except that, in contrast with 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 section IV.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 bottom-left to 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).

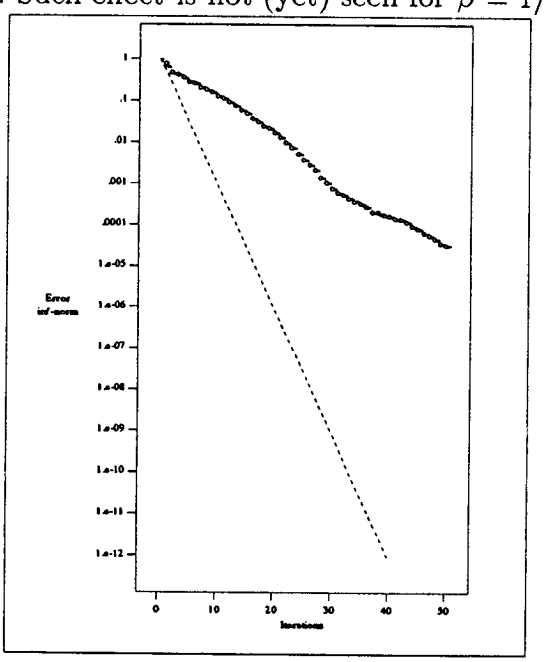
In section IV.4 we give results for a symmetric transonic flow, $M_\infty = 0.85$ and $\alpha = 0.0^o$. The flow in this problem has two shocks. In section IV.5 we give results for a asymmetric transonic flow, $M_\infty = 0.85$ and $\alpha = 1.0^o$, in which problem we encounter an additional contact discontinuity. This last problem is also known from the GAMM workshop on the Numerical Simulation of Compressible Euler Flows (1986) [DVPR]. The mesh used for the NACA airfoil is a 20×32 mesh (i.e. a level 3 mesh in a sequence of which the coarsest is a 5×8). Similar results, however, were obtained on the 40×64 mesh, with as only differences that (1) some convergence effects were seen after a larger number of ItDeC iteration cycles, and (2) some ItDeC convergence rates were slightly faster(!).

2. Subsonic Flow over a NACA0012 Airfoil

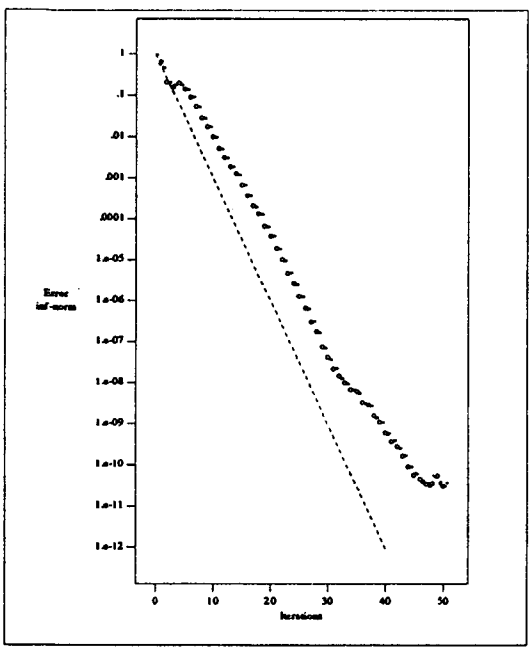
In figure IV.1 we see the convergence of ItDeC iteration for subsonic flow and for different values of β. We see that the iteration doesn't converge for $\beta = 0$, as it doesn't for $\beta = 1$ (not shown). We obtain slow convergence for $\beta = 0.1$ and $\beta = 0.9$. Good convergence with a rate of approximately 0.5 per iteration step is obtained for $\beta = 1/3, 1/2$ and $2/3$. Probably the asymptotic rate cannot be observed because rounding error accuracy is obtained after approximately 40 iterations. 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 $\beta = 1/2$.



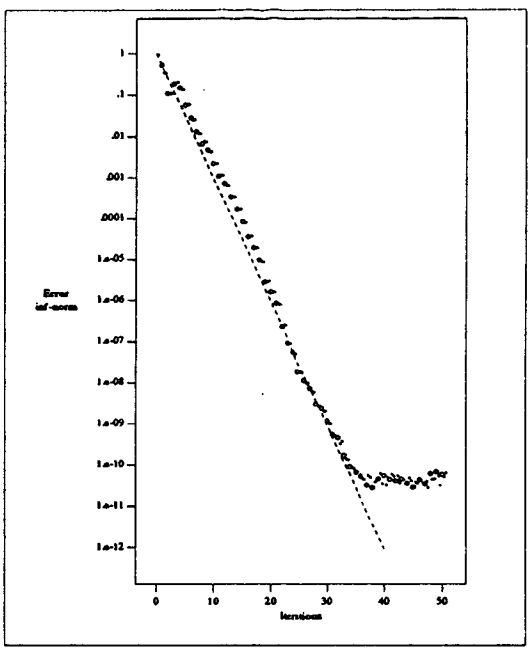
Case a: $\beta = 0$



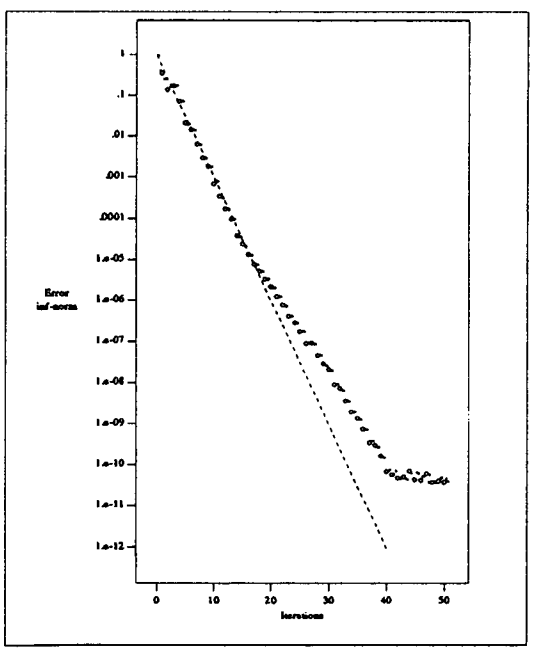
Case b: $\beta = 0.1$



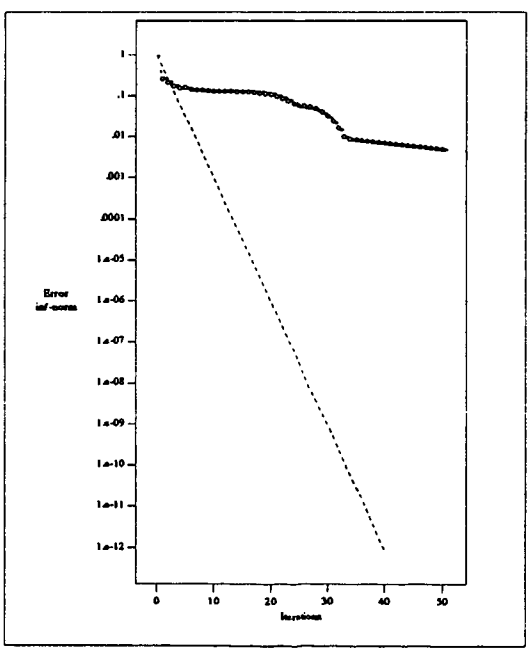
Case c: $\beta = 1/3$



Case d: $\beta = 1/2$



Case e: $\beta = 2/3$

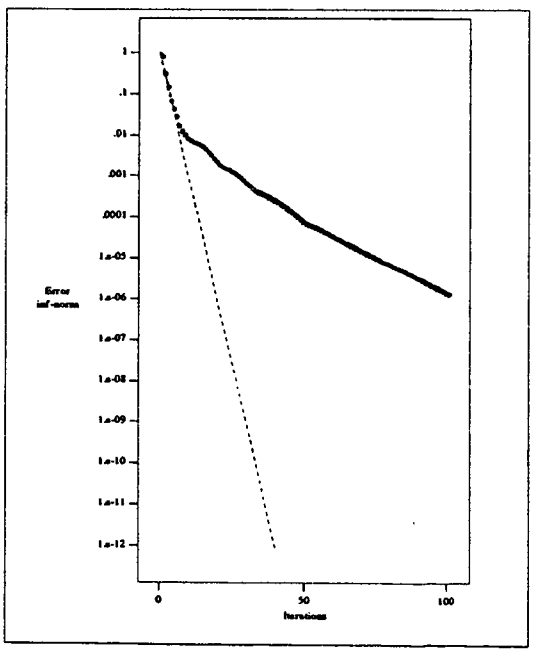


Case f: $\beta = 0.9$

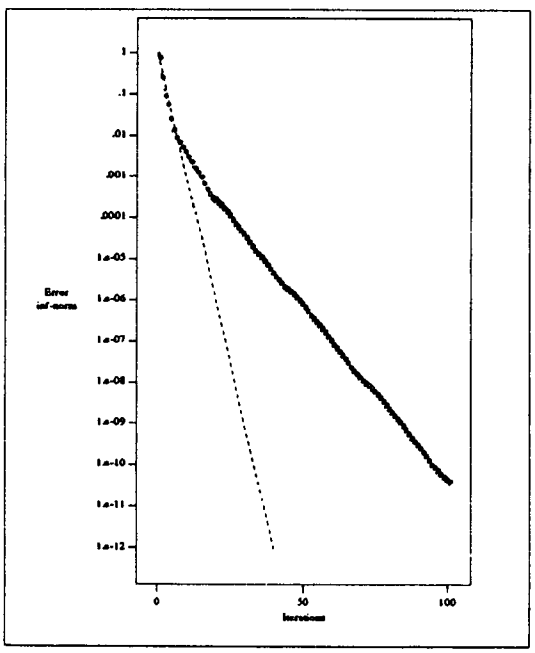
Figure IV.1: Subsonic Flow over a NACA0012 Airfoil
Defect-Correction Method, 20 × 32 mesh.
The dashed line corresponds with a convergence rate 1/2.

3. Flow with a Contact Discontinuity

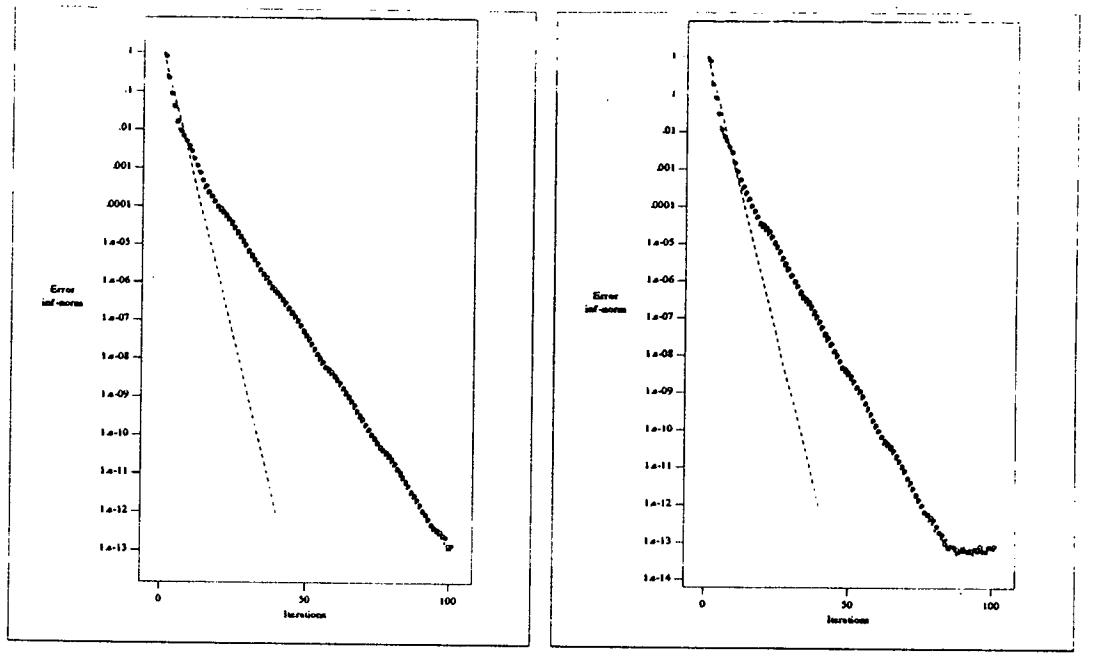
For this flow with a contact discontinuity, $\beta = 0$ gives a diverging process (not shown), and $\beta = 0.1$ shows worse convergence than $\beta = 0.9$. The asymmetry in the convergence behaviour with respect to $\beta > 1/2$ (worse) and $\beta < 1/2$ (better convergence) might be understood by the location of the eigenvalues in the complex plane (as shown in figure II.2). There we see that more eigenvalues are located in the neigbourhood 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.



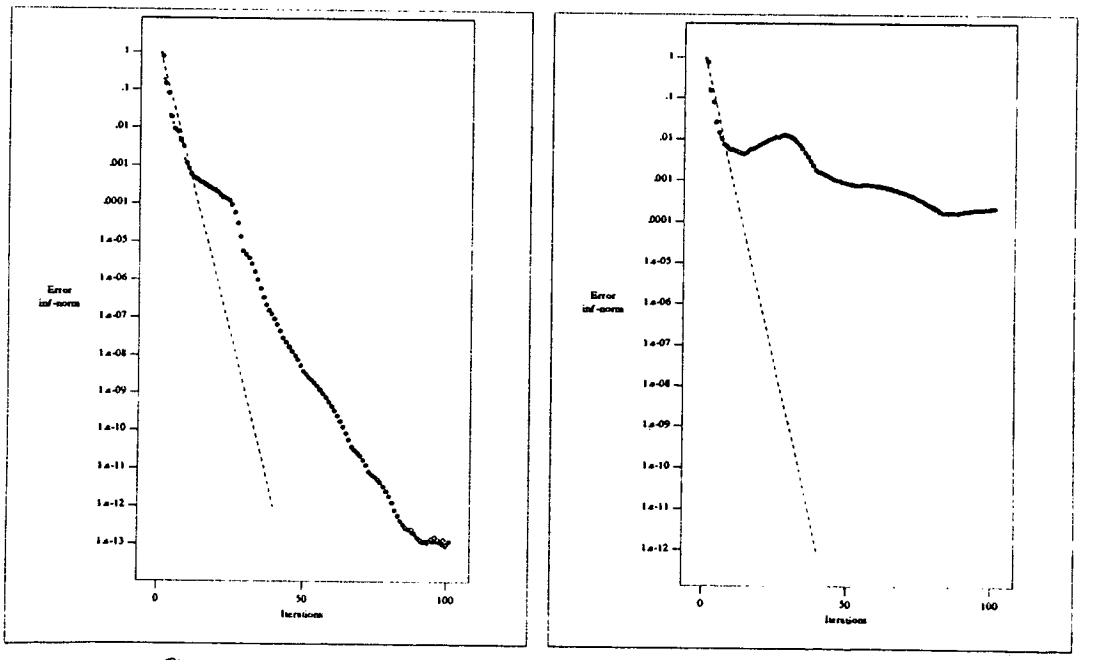
Case a: $\beta = 0.1$



Case b: $\beta = 1/3$



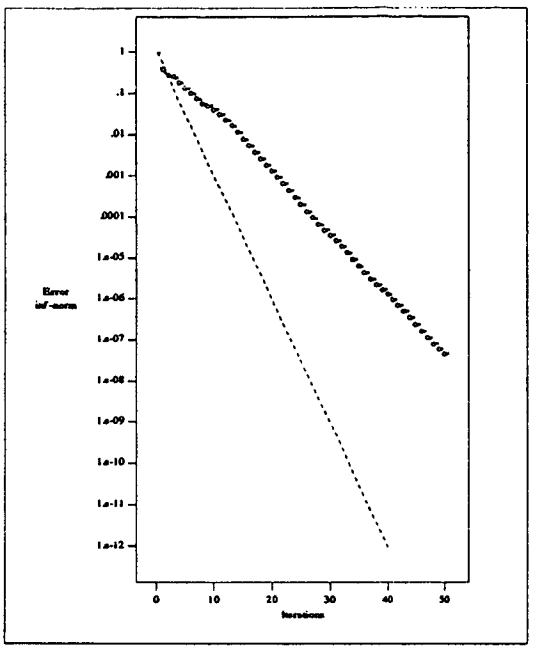
Case c: $\beta = 1/2$ Case d: $\beta = 2/3$



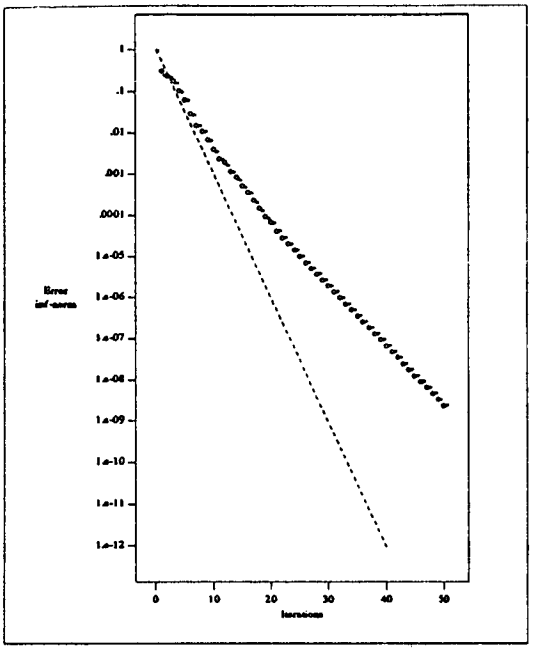
Case e: $\beta = 0.9$ Case f: $\beta = 1$

Figure IV.2: Flow with a 45° Contact Discontinuity.
Defect-Correction Method, 16 × 16 mesh.
The dashed line corresponds with a convergence rate 1/2.

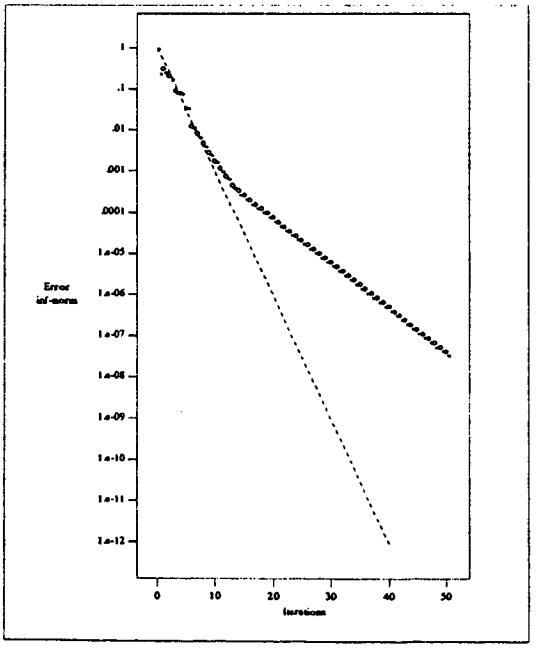
4. Symmetrical Transonic Flow over a NACA0012 Airfoil



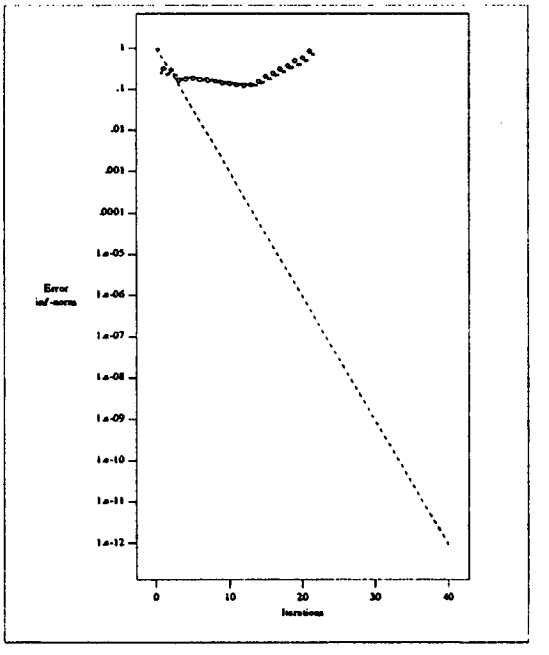
Case a: $\beta = 1/3$



Case b: $\beta = 1/2$



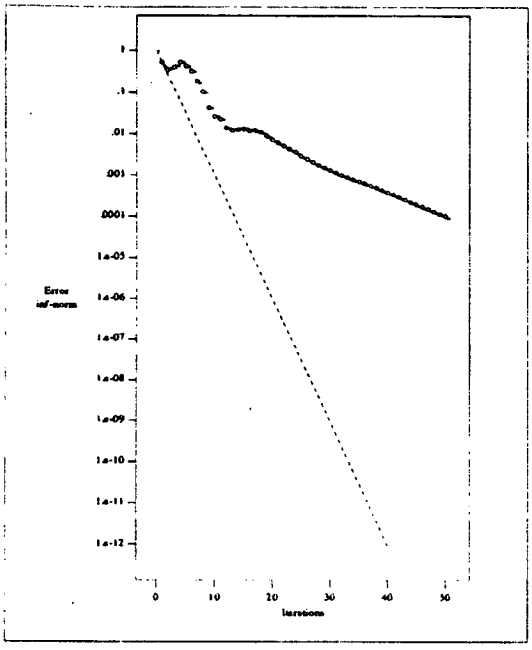
Case c: $\beta = 2/3$



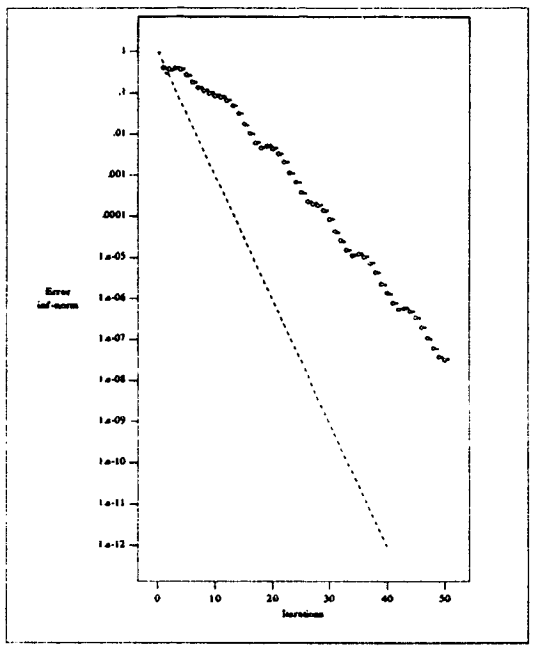
Case d: $\beta = 0.9$

Figure IV.3: Symmetrical Transonic Flow over a NACA0012 Airfoil
Defect-Correction Method, 20 × 32 mesh.
The dashed line corresponds with a convergence rate 1/2.

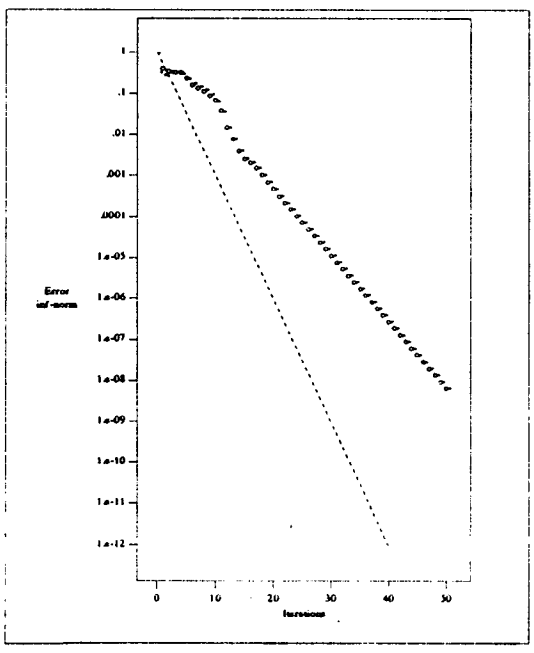
5. Asymmetrical Transonic Flow over a NACA0012 Airfoil



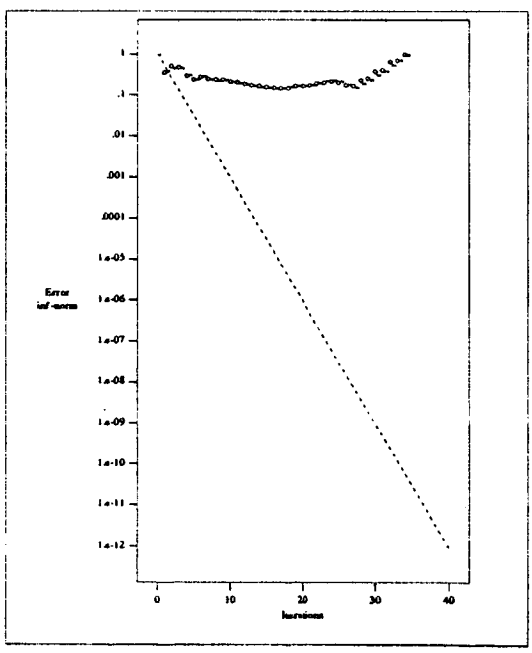
Case a: $\beta = 1/3$



Case b: $\beta = 1/2$



Case c: $\beta = 2/3$



Case d: $\beta = 0.9$

Figure IV.4: Asymmetrical Transonic Flow over a NACA0012 Airfoil
Defect-Correction Method, 20 × 32 mesh.
The dashed line corresponds with a convergence rate 1/2.

V. CONCLUSIONS

Unfactored implicit second order schemes for the solution of the steady state Euler equations are identified as iterative defect correction methods, and an analysis is given for the convergence of such iterations. For a linear model problem, both for the one- and for the two-dimensional case, Fourier and matrix analysis have been used. The convergence rate is evaluated, depending on a parameter $\beta \in [0,1]$, that determines the amount of upwinding that is present in the second order discrete operator. The values $\beta = 0$ and $\beta = 1$ are shown to yield defective error amplification matrices.

The matrix analysis allows us to understand the pathology of the schemes that are characterized by such defective amplification matrices. These schemes, before achieving their asymptotic convergence, exhibit a **pseudo-convection phase** during which the norm of the residual may not be reduced. The error -prior to being dissipated- is being transferred over the mesh at a completely unphysical speed. This phase extends over a number of iterations equal to $N/(1-\rho)$ in which N is the size of the largest defective Jordan block of the amplification matrix, and ρ its spectral radius; thus this extent can be very large. In one dimension, the non-pathological schemes asymptotically converge at the rate of the sequence 2^{-n}. In two dimensions, schemes for which the upwinding parameter satisfies $\beta \geq 1/2$ also obey this, as a rule. However, after the initial impulsive start and before the asymptotic rate is reached, those schemes that are close to the pathological ones, show a **Fourier phase** during which the effect of boundary conditions have not yet been felt. In this phase the spectral radius given by Fourier analysis does control the convergence rate.

In the last Section, computations of two-dimensional Euler flows are presented. Calculations are shown both for smooth flows and for flows with shocks and contact discontinuities. transonic symmetric/unsymmetric airfoil flows. Several theoretical results for the model problem are confirmed in these more complex situations:
- (i) the recommended **half-upwind scheme** ($\beta = 1/2$) is shown to converge approximately at the rate of 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.
- (iv) the schemes with β close to 0 or 1 converge badly.

VI. ACKNOWLEDGEMENTS

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 have been constant; we gratefully thank him. We also want to acknowledge the boards of directors of INRIA and CWI for their stimulation of the cooperation between our two institutes.

VII. REFERENCES

[BKOR] B. KOREN, "Multigrid and Defect Correction for the Steady Navier-Stokes Equations, Application to Aerodynamics", Doctoral Thesis, Centrum voor Wiskunde en Informatica, Amsterdam, 1989.

[BMWG] R. BEAM, and R. F. WARMING, "An Implicit Finite-Difference Algorithm for Hyperbolic Systems in Conservation-Law-Form", Journal of Computational Physics, Vol. 22, Sept. 1976, pp. 87-110.

[CHAK] S. R. CHAKRAVARTHY, "Relaxation Methods for Unfactored Implicit Upwind Schemes", AIAA Paper 84-0165 (also in AIAA Journal), AIAA 2nd Aerospace Sciences Meeting, January 9-12, 1984/Reno, Nevada.

[DCM1] K. BÖHMER, P. W. HEMKER, and H. J. STETTER, "The Defect Correction Approach", in: Defect Correction Methods. K. Böhmer, H. J. Stetter, eds., Comp. Suppl. 5, 1-32, Springer - Verlag, Wien, New York, 1984.

[DVPR] A. DERVIEUX, B. Van LEER, J. PERIAUX and A. RIZZI eds, "Numerical Simulation of Compressible Euler Flows ", Vieweg Verlag, Braunschweig/Wiesbaden, 1989.

[FARL] F. ANGRAND, J. ERHEL, "Vectorized Finite Element codes for compressible flows", Proc. of "Finite Element in Flow Problems", Antibes (France), June 16-20, 1986, Wiley and Sons.

[FEM3] B. STOUFFLET, J. PERIAUX, F. FEZOUI, A. DERVIEUX, "Numerical Simulation of 3-D Hypersonic Euler Flows Around Space Vehicles Using Adapted Finite Elements", AIAA-87-0560, AIAA 25th Aerospace Sciences Meeting, January 12-15, 1987/Reno, Nevada.

[FZBS] F. FEZOUI, B. STOUFFLET, "A Class of Implicit Upwind Schemes for Euler Simulation with Unstructured Meshes", J. of Comp. Phys. 84, No. 1, Sept. 1989.

[HACK] W. HACKBUSCH, "Multigrid Methods and Applications", Springer Verlag, 1985.

[HRKO] P. W. HEMKER, and B. KOREN, "A non-linear multigrid method for the steady Euler equations ", In:[DVPR]

[HRPW] P.W. HEMKER, "Computation of Layers in Eulerian Gas Flow", In: BAIL IV,

Proceedings of the Fourth International Conference on Boundary and Interior Layers (S.K.Godunov, J.J.H.Miller and V.A.Novikov eds), Boole Press, Dublin, 1986.

[HRSP] P. W. HEMKER, and S. P. SPEKREIJSE, "Multiple grid and Osher's scheme for the efficient solution of the steady Euler equations ", Appl. Num. Math. 2, 475-493 (1986).

[JAD1] J-A. DESIDERI, "Preliminary results on the iterative convergence of a class of implicit schemes", rapport INRIA No. 490, 1986.

[JAD2] J-A. DESIDERI, A. DERVIEUX, "Compressible Flow Solvers Using Unstructured Grids", Von Karman Institute for Fluid Dynamics, Lecture Series 1988-05, Computational Fluid Dynamics, March 7-11, 1988.

[JLS1] J. L. STEGER, "Coefficient Matrices for Implicit Finite Difference Solution of the Inviscid Fluid Conservation Law Equations", Computer Methods in Applied Mechanics and Engineering 13 (1978), pp. 175-188.

[JLS2] J. L. STEGER, "Implicit Finite-Difference Simulation of Flow about Arbitrary Geometries with Application to Airfoils", AIAA Paper 77-663 (1977).

[JLS3] J. L. STEGER, and R. F. WARMING, "Flux Vector Splitting of the Inviscid Gasdynamic Equations with Application to Finite-Difference Methods", J. of Comp. Phys. 40, 263-293 (1981).

[KRON] F. W. BYRON, and R. W. FULLER, "Mathematics of Classical and Quantum Physics", Vol. 1, Addison-Wesley Publishing Company, 1969.

[MCCM] "Multigrid Methods" (S. McCormick, Ed.), SIAM Frontier Series in Applied Mathematics, vol. III, SIAM, Philadelphia, 1987.

[MHL1] M-H. LALLEMAND, "Schémas Décentrés Multigrilles pour la Résolution des Equations d'Euler en Eléments Finis", Doctoral Thesis, Université de Provence, Centre Saint Charles, March 1988.

[ORTE] J. M. ORTEGA, and W. C. RHEINBOLDT, "Iterative Solution of Nonlinear Equations in Several Variables", Academic Press, New York, 1970.

[RMCK] R. W. MACCORMACK, "Current Status of Numerical Solutions of the Navier-Stokes Equations", AIAA Paper 85-0032, 1985.

[STOU] B. STOUFFLET, "Résolution Numérique des Equations d'Euler des Fluides Parfaits Compressibles par des Schémas Implicites en Eléments Finis", thesis,

Université Pierre et Marie Curie, Paris VI, 1984, and

in Proceedings, *INRIA Workshop on Numerical Methods for Compressible Inviscid Fluids, Rocquencourt, France, 1983*, edited by F. Angrand *et al.* (SIAM, Philadelphia, 1984), pp. 409-434.

[STRA] G. STRANG, "Linear Algebra and its Applications", Second Edition, Academic Press, New York, 1980.

[STV1] H. STEVE, "Méthodes Implicites Efficaces pour la Résolution des Equations d'Euler en Eléments Finis", INRIA Report No. 779, December 1987.

[STV2] H. STEVE, "Schémas Implicites Linéarisés Décentrés pour la Résolution des Equations d'Euler en Plusieurs Dimensions", Doctoral Thesis, Université de Provence Aix-Marseille I, July 1988.

[TVLW] J. L. THOMAS, B. van LEER, and R. W. WALTERS, "Implicit Flux-Split Schemes for the Equations", AIAA Paper 85-1680, 1985.

[VARG] R. S. VARGA, "Matrix Iterative Analysis", Prentice Hall, Inc., Enlewoods Cliffs, N. J., 1962.

[VLMU] B. van LEER, and W. A. MULDER, "Relaxation Methods for Hyperbolic Conservation Laws", Proc. of the INRIA Workshop on Numerical Methods for the Euler Equations of Fluid Dynamics, Rocquencourt, France, December 7-9, 1983, Angrand et al. Eds., SIAM Philadelphia/1985.

[WBHT] R. F. WARMING, R. M. BEAM, and B. J. HYETT, "Diagonalization and Simultaneously Symmetrization of the Gas-Dynamic Matrices", Mathematics of Computation, Vol. 29, 132 (1975), pp. 1037-1045.

VIII. APPENDIX

1. Model Differencing Schemes

In this section, the various differencing schemes employed in the one-dimensional matrix analysis are defined explicitly for both periodic and non-periodic boundary conditions. In the latter case, the assumed left (LB) and right (RB) boundary conditions are indicated.

Three operators are considered:
(a) the central-difference operator,
(b) the first-order upwind operator, and
(c) the second-order upwind operator. In the implicit schemes under study, the explicit part is a linear combination of the operators (a) and (c), while the implicit preconditioner is (b).

The eigenvalues of these operators are known analytically, except for the central difference operator in the non-periodic case for which a numerical computation has been made for the illustration.

Contrasting with the rest, the matrices representing upwind operators when non-periodic boundary conditions are assumed are defective. The linearly independent eigenvectors are only a few independently of the dimension N.

Model Differencing Schemes

matrices are $N \times N$ ($N = 25$ for plots)
$\theta_m = \frac{2m\pi}{N}$, $m = 0, 1, \cdots, N-1$
LB/RB: left/right boundary procedure

(1) central-difference operator

(a) periodic case (o)

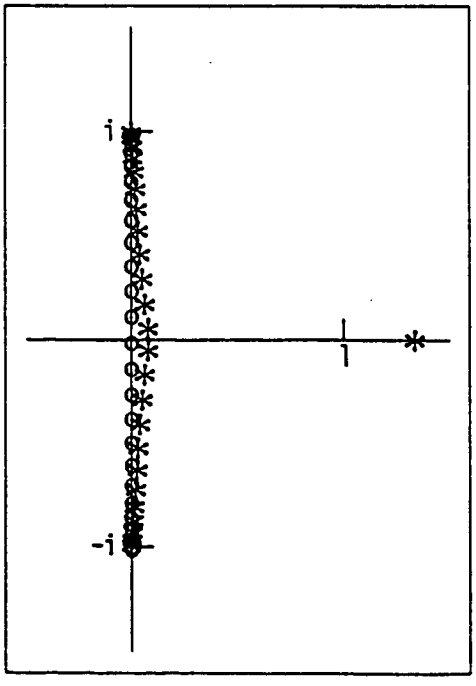
$$\delta_x^c = \frac{1}{2}\begin{pmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{pmatrix}$$

skew-symmetric (diagonalizable) matrix
purely imaginary eigenvalues
$\lambda_m = i \sin\theta_m$
$Re(\lambda_m) = 0$

(b) non periodic case (*)
LB: Dirichlet (u_0 given): RB: 1st-order upwind

$$\delta_x^c = \frac{1}{2}\begin{pmatrix} 0 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ & & & & -2 & 2 \end{pmatrix}$$

(diagonalizable matrix)
$\lambda_0 \in I\!R^+$
and for $m = 1, 2, \cdots, N-1$
$Re(\lambda_m) > 0$, and ≈ 0



(2) 1st-order upwind (backward-difference operator)

(a) periodic case (o)

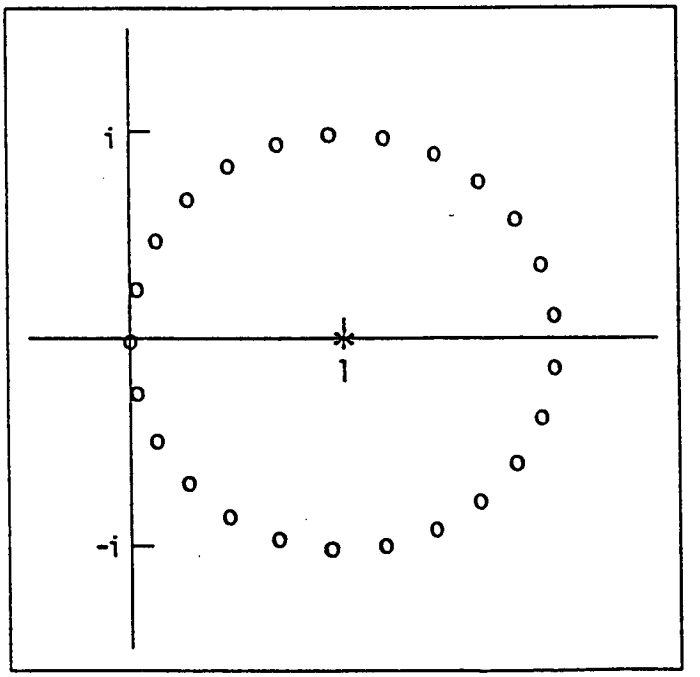
$$\delta^u_{x,1} = \begin{pmatrix} 1 & & & & & -1 \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{pmatrix}$$

(diagonalizable matrix)
$\lambda_m = 1 - e^{-i\theta_m}$

(b) non periodic case (*)
LB: Dirichlet (u_0 given)

$$\delta^u_{x,1} = \begin{pmatrix} 1 & & & & & \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{pmatrix}$$

(defective) triangular matrix
$\lambda_m = 1. \; \forall m$



(3) 2nd-order fully upwind (backward-difference operator)

(a) periodic case (o)

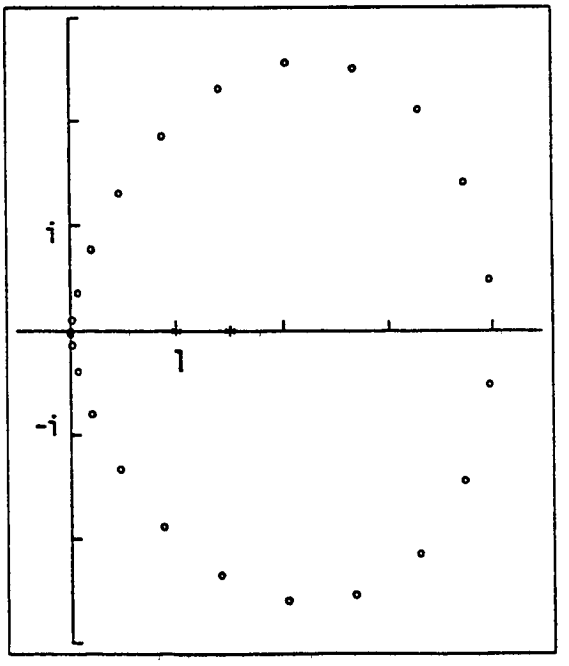
$$\delta^u_{x,2} = \frac{1}{2}\begin{pmatrix} 3 & & & & 1 & -4 \\ -4 & 3 & & & & 1 \\ 1 & -4 & 3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -4 & 3 & \\ & & & 1 & -4 & 3 \end{pmatrix}$$

(diagonalizable matrix)
$\lambda_m = \frac{1}{2}(3 - 4e^{-i\theta_m} + e^{-2i\theta_m})$

(b) non periodic case (*)
LB: Dirichlet (u_0 given), and 1st-order upwind scheme applied at gridpoint 1

$$\delta^u_{x,2} = \frac{1}{2}\begin{pmatrix} 2 & & & & & \\ -4 & 3 & & & & \\ 1 & -4 & 3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -4 & 3 & \\ & & & 1 & -4 & 3 \end{pmatrix}$$

(defective) triangular matrix
$\lambda_0 = 1,\ \lambda_m = \frac{3}{2}\ (m = 1, 2, \cdots, N-1)$



2. Diagonalization of the Amplification Matrix G_∞

In the particular case where the upwinding parameter $\beta = 0$ or 1, the amplification matrix G_∞ cannot be diagonalized and has been studied in details in [JAD1, JAD2]. In this appendix, the eigensystem is evaluated in the case where instead

$$0 < \beta < 1, \tag{107}$$

Let $N \in \mathbb{N}$ be the dimension of the matrix G_∞. The vector $u \in \mathbb{C}^N$ is an eigenvector of the amplification operator G_∞ of (22) or (27) and λ is an associated eigenvalue, iff

$$(I - \delta_1^{-1}\delta_2)\, u = \lambda u \tag{108}$$

or, equivalently

$$\delta_2\, u = (1 - \lambda)\, \delta_1\, u \tag{109}$$

The symbol δ_1 refers to the first-order backward difference operator or to the following matrix representation:

$$\delta_1 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \tag{110}$$

The second-order partially-upwind difference operator δ_2 is given in terms of the parameter β by:

$$\begin{aligned} \delta_2 &= (1-\beta)\,\delta_2^c + \beta\,\delta_2^u \\ &= (1-\beta)\,\mathrm{Trid}\left(-\frac{1}{2}, 0, \frac{1}{2}\right) + \beta\,\mathrm{Pentad}\left(\frac{1}{2}, -2, \frac{3}{2}, 0, 0\right) \end{aligned} \tag{111}$$

in which the symbols Trid and Pentad denote tridiagonal and pentadiagonal matrices with constant subdiagonals (except for first and last rows which reflect the applied boundary conditions). These matrices are given explicitly in Appendix, section 1. This gives

$$2\delta_2 = \mathrm{Pentad}(\beta, -3\beta - 1, 3\beta, 1 - \beta, 0)$$

$$= \begin{pmatrix} 2\beta & 1-\beta & 0 & \cdots & \cdots & 0 \\ -(3\beta+1) & 3\beta & 1-\beta & \ddots & & \vdots \\ \beta & -(3\beta+1) & 3\beta & 1-\beta & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \beta & -(3\beta+1) & 3\beta & 1-\beta \\ 0 & \cdots & 0 & \beta & -(2\beta+2) & \beta+2 \end{pmatrix} \tag{112}$$

Note that the bandwidth of the matrix is in general equal to 4 except in two cases that have been excluded by assumption, $\beta = 0$ and $\beta = 1$, for which it reduces to 3. Thus (2) is equivalent to the following system:

$$\begin{cases} \text{(a)} & 2\beta u_1 + (1-\beta)u_2 = 2(1-\lambda)u_1 \\ \text{(b)} & \beta u_{j-2} - (3\beta+1)u_{j-1} + 3\beta u_j + (1-\beta)u_{j+1} = 2(1-\lambda)(u_j - u_{j-1}) \\ & \qquad\qquad (j = 2, 3, \cdots, N-1) \\ \text{(c)} & \beta u_{N-2} - (2\beta+2)u_{N-1} + (\beta+2)u_N = 2(1-\lambda)(u_N - u_{N-1}) \end{cases} \tag{113}$$

For $j = 2$, (114b) holds provided a value u_0 is conventionally defined to be zero:

$$u_0 = 0 \tag{114}$$

Similarly, (114c) can be viewed as the particular form taken by (114b) for $j = N$, if one defines an additional value u_{N+1} such that:

$$\beta u_{N-2} - (2\beta+2)u_{N-1} + (\beta+2)u_N = \beta u_{N-2} - (3\beta+1)u_{N-1} + 3\beta u_N + (1-\beta)u_{N+1} \tag{115}$$

that is,

$$u_{N-1} - 2u_N + u_{N+1} = 0 \tag{116}$$

since $\beta \neq 1$. Consequently, the system to be solved is equivalent to:

$$\begin{cases} \text{(a)} & u_0 = 0 \\ \text{(b)} & (1-\beta)u_2 = \big(2(1-\lambda) - 2\beta\big)u_1 \\ \text{(c)} & \beta u_{j-2} + (1-2\lambda-3\beta)u_{j-1} + (3\beta-2+2\lambda)u_j + (1-\beta)u_{j+1} = 0 \\ & \qquad\qquad (j = 2, 3, \cdots, N) \\ \text{(d)} & u_{N-1} - 2u_N + u_{N+1} = 0 \end{cases} \tag{117}$$

The above system is a homogeneous 2-point boundary-value problem in which the unknown is a sequence u_j ($j = 0, 1, \cdots, N+1$). The first two equations, (118a,b), are left boundary conditions, and the last equation, (118d), a right boundary condition. The remaining equation, 118c), is a 4-level linear-homogeneous recurrence formula. This equation is solved by first calculating the roots r_0, r_1 and r_2 of the equation

$$(1-\beta)r^3 + (3\beta-2+2\lambda)r^2 + (1-2\lambda-3\beta)r + \beta = 0 \tag{118}$$

When these roots are distinct, the general solution is given by:

$$u_j = A_0 r_0^j + A_1 r_1^j + A_2 r_2^j \tag{119}$$

in which the constants A_0, A_1 and A_2 are determined from the boundary conditions. The remaining is a search for solutions associated with triplets of distinct roots, each triplet corresponding to a particular eigenvalue. Since this search does yield

the complete set of the eigenvalues (as verified *a posteriori*), it is unnecessary to consider the case of multiple roots.

One of the roots is obvious. It is,

$$r_0 = 1 \tag{120}$$

This is because the constant solution which is discretely represented by a vector all of whose components are equal, is an eigenmode (eigenvector) of all consistent difference operators, and thus in particular of both operators δ_1 and δ_2, and consequently the matrix G_∞. Dividing (119) by the polynomial $(r-1)$ yields the following equation for the remaining two roots:

$$(1-\beta)r^2 + (2\beta + 2\lambda - 1)r - \beta = 0 \tag{121}$$

This gives:

$$r_1, r_2 = \frac{1 - (2\beta + 2\lambda) \mp \sqrt{\left(1 - (2\beta + 2\lambda)\right)^2 + 4\beta(1-\beta)}}{2(1-\beta)} \tag{122}$$

Note that the following equations also hold:

$$\begin{aligned} r_1 + r_2 &= \frac{1 - (2\beta + 2\lambda)}{1-\beta} \\ r_1 r_2 &= \frac{-\beta}{1-\beta} \end{aligned} \tag{123}$$

Recall that the system to be solved is homogeneous and thus the sequence u_j can be normalized arbitrarily. To realize this, and simplify the solution, we put

$$u_1 = 1 - \beta \tag{124}$$

so that

$$u_2 = 2 - 2\lambda - 2\beta \tag{125}$$

Knowing u_0, u_1 and u_2, the constants A_0, A_1 and A_2, and the sequence u_j can be determined uniquely in terms of λ. Substituting the result in (118d) will produce the condition under which the system is compatible, that is, the characteristic equation yielding the values of λ.

Consequently, the system to be solved for the constants, A_0, A_1 and A_2 is:

$$\begin{cases} A_0 + A_1 + A_2 = u_0 = 0 \\ A_0 + A_1 r_1 + A_2 r_2 = u_1 = 1 - \beta \\ A_0 + A_1 r_1^2 + A_2 r_2^2 = u_2 = 2 - 2\lambda - 2\beta \end{cases} \tag{126}$$

One finds

$$\begin{cases} A_0 = \frac{1-\beta}{2\lambda} \\ A_1 = \frac{A_0}{r_2 - r_1}(r_2 - 1)\big((1-\beta)r_2 + 2\lambda + \beta - 1\big) \\ A_2 = -\frac{A_0}{r_2 - r_1}(r_1 - 1)\big((1-\beta)r_1 + 2\lambda + \beta - 1\big) \end{cases} \tag{127}$$

where some simplifications were made using (123-126). As a result,

$$u_j = A_0\left(1 + \frac{v_j}{r_2 - r_1}\right) \tag{128}$$

where

$$v_j = (r_2 - 1)\big((1-\beta)r_2 + 2\lambda + \beta - 1\big)\, r_1^j - (r_1 - 1)\big((1-\beta)r_1 + 2\lambda + \beta - 1\big)\, r_2^j \tag{129}$$

Making the corresponding substitutions in (118d) yields the condition on λ:

$$\begin{aligned} &(r_2 - 1)\big((1-\beta)r_2 + 2\lambda + \beta - 1\big)\big(r_1^{N+1} - 2r_1^N + r_1^{N-1}\big) = \\ &\qquad (r_1 - 1)\big((1-\beta)r_1 + 2\lambda + \beta - 1\big)\big(r_2^{N+1} - 2r_2^N + r_2^{N-1}\big) \end{aligned} \tag{130}$$

or

$$\begin{aligned} &(r_2 - 1)(r_1 - 1)^2\big((1-\beta)r_2 + 2\lambda + \beta - 1\big)r_1^{N-1} = \\ &\qquad (r_1 - 1)(r_2 - 1)^2\big((1-\beta)r_1 + 2\lambda + \beta - 1\big)r_2^{N-1} \end{aligned} \tag{131}$$

The above equation holds when either one of the following three possibilities applies:

$$\begin{cases} \text{(A)} & r_1 = 1 \text{ , or} \\ \text{(B)} & r_2 = 1 \text{ , or} \\ \text{(C)} & (r_1 - 1)\big((1-\beta)r_2 + 2\lambda + \beta - 1\big)r_1^{N-1} = \\ & \qquad (r_2 - 1)\big((1-\beta)r_1 + 2\lambda + \beta - 1\big)r_2^{N-1} \end{cases} \tag{132}$$

Note that

$$\begin{aligned} &(r_1 - 1)\big((1-\beta)r_2 + 2\lambda + \beta - 1\big) \\ &\qquad = (1-\beta)\big(r_1 r_2 - (r_1 + r_2) + 1\big) + 2\lambda(r_1 - 1) \\ &\qquad = 2\lambda r_1 \end{aligned} \tag{133}$$

where (124) has been used. Thus, in (133), (C) reduces to

$$\lambda r_1^N = \lambda r_2^N \tag{134}$$

or equivalently:

$$\begin{cases} \text{(C1)} & \lambda = 0 \text{ , or} \\ \text{(C2)} & r_1^N = r_2^N \end{cases} \tag{135}$$

In summary, λ is an eigenvalue iff either one of the following four equations holds (inclusively): (A) or (B) in (133), or (C1) or (C2) in (136).

It is easy to show that:

$$(\mathrm{A}) \text{ or } (\mathrm{B}) \Longrightarrow (\mathrm{C1}) \tag{136}$$

Conversely, $\lambda = 0$ implies that

$$\begin{cases} r_1 = \frac{-\beta}{1-\beta} \\ r_2 = 1 \end{cases} \tag{137}$$

and the satisfaction of (B). Therefore, one eigenvalue is indeed zero:

$$\lambda_0 = 0 \tag{138}$$

The corresponding eigenvector can directly be found to be:

$$u_j = j \tag{139}$$

or a multiple of it. This is no surprise, since it means that both first-order and second-order difference operators act exactly in the same way on a linear distribution of data, that is, with no truncation error. The remaining eigenvalues are found by solving the equation:

$$r_1^N = r_2^N \tag{140}$$

This gives:

$$\frac{r_2}{r_1} = e^{i\theta_m} \tag{141}$$

where

$$\theta_m = \frac{2\pi m}{N} \tag{142}$$

and $m = 1, 2, \cdots, N-1$. Note that the case $m = 0$ (or N) is not retained since it corresponds to a case where $r_1 = r_2$ which has been implicitly excluded. Substituting the expressions of the roots r_1 and r_2 into the above equation yields after some straightforward calculation, the remaining $N-1$ eigenvalues:

$$\lambda_m = \frac{1}{2} - \beta + i\sqrt{\beta(1-\beta)}\cos\frac{\pi m}{N} \tag{143}$$

($m = 1, 2, \cdots, N-1$). Since all the eigenvalues have been found, the search is complete.

We now examine heuristically how is the amplification matrix G_∞ conditioned. For this, and since the eigenvalues are not of disparate scales, one can simply examine the l_2-condition number of the eigenvector matrix,

$$\kappa_2(U) = \|U\|_2 \, \|U^{-1}\|_2 = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \tag{144}$$

in which $\lambda_{\min}$ and $\lambda_{\max}$ are respectively the minimum and maximum of the real positive eigenvalues of the matrix $U^* U$ (where U^* is the conjugate of the matrix U). The substitution of the formula for the eigenvalues into the expressions of the roots r_1 and r_2 gives:

$$r_1, r_2 = -i \sqrt{\frac{\beta}{1-\beta}} \, e^{\mp \frac{i\theta_m}{2}} \tag{145}$$

Consequently, and in view of (146), when the parameter $\frac{\beta}{1-\beta}$ is either very large or very small, large discrepancies in the magnitudes of the components of an eigenvector u will be observed. More precisely, when $\beta \to 0$ or 1, the magnitude of the ratio of the first component to the last, tends to infinity, and this for all frequency modes. in these limits, the condition number κ_2 tends to infinity, also because eigendirections coalesce. Conversely, it may be conjectured that the least condition number of the matrix G_∞ is achieved by the half-upwind scheme ($\beta = 1/2$), since it realizes the best separation of the eigenvalues, and minimizes $\left|\log \frac{\beta}{1-\beta}\right|$.

In conclusion, when the upwinding parameter is strictly between 0 and 1, the amplification matrix, G_∞, can be diagonalized except in the case where $\beta = 1/2$ and N is even for which the eigenvalue 0 is double, whereas only one eigenvector is associated with it. In addition, the spectral radius of G_∞ is uniformly less than 1/2, but the matrix is believed to be best conditioned when $\beta = 1/2$.

3. Defective Linear Iterations

In this appendix, we examine the case of a general linear non-homogeneous iteration,

$$w^{n+1} = Gw^n + b \tag{146}$$

in which the unknown w is an N-vector, G is a given $N \times N$ amplification matrix and b is a given constant N-vector. in the particular case where **the matrix G is defective**, that is, **cannot be diagonalized.**

By a suitable similarity transform, G can still be reduced to the so-called "**Jordan canonical form**" [STRA]:

$$J = X^{-1}GX \tag{147}$$

where X is the generalized eigenvector matrix, and J is block-diagonal:

$$J = BDiag(J_i) = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{pmatrix} \tag{148}$$

where s is the number of linearly independent (true) eigenvectors ($s < N$), and each block J_i has the following bidiagonal structure:

$$J_i = Bidiag(1, \lambda_i) = \begin{pmatrix} \lambda_i & & & & \\ 1 & \lambda_i & & & \\ & 1 & \lambda_i & & \\ & & \ddots & \ddots & \\ & & & 1 & \lambda_i \end{pmatrix} \tag{149}$$

The matrix J being triangular, it contains its eigenvalues in its main diagonal; these are the numbers λ_i ($i = 1, 2, .., s$) that also are the eigenvalues of the matrix G to which J is similar. To assure convergence of the iteration, it is assumed that the amplification matrix G satisfies the spectral radius condition [VARG]:

$$\rho(G) = \max_{i=1,2,\cdots,s} (|\lambda_i|) \tag{150}$$

In such case, none of the eigenvalues of the matrix G is equal to 1, the matrix $I - G$ is invertible, and the iteration admits a fixed-point solution $w^\infty = (I - G)^{-1}b$ that satisfies

$$w^\infty = Gw^\infty + b \tag{151}$$

Then defining the "**error vector**" e^n by

$$e^n = w^n - w^\infty \tag{152}$$

and subtracting (152) from (147) it follows that e^n satisfies the following homogeneous linear iteration

$$e^{n+1} = Ge^n \tag{153}$$

which implies that

$$e^n = G^n e^0 \tag{154}$$

But,

$$G^n = XJ^nX^{-1} \tag{155}$$

and in the basis of the generalized eigenvectors, the error vector becomes,

$$\epsilon^n = X^{-1}e^n \tag{156}$$

so that combining the last three equations yields the expression:

$$\epsilon^n = J^n\epsilon^0 \tag{157}$$

In addition, as a consequence of the block-diagonal structure of the matrix J given by (149),

$$J^n = BDiag(J_i^n) \tag{158}$$

It is therefore apparent that the attenuation with increasing n of the error-vector components is governed by the powers of the individual blocks J_i taken separately. For this reason, in what follows, and without great loss of generality, we consider the case of only one block ($s = 1$; $J = J_1$; $\lambda_1 = \lambda$). Using (150) several times successively, we obtain:

$$J^2 = \begin{pmatrix} \lambda^2 & & & & \\ 2\lambda & \lambda^2 & & & \\ 1 & 2\lambda & \lambda^2 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2\lambda & \lambda^2 \end{pmatrix} \tag{159}$$

$$J^3 = \begin{pmatrix} \lambda^3 & & & & & \\ 3\lambda^2 & \lambda^3 & & & & \\ 3\lambda & 3\lambda^2 & \lambda^3 & & & \\ 1 & 3\lambda & 3\lambda^2 & \lambda^3 & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & 3\lambda & 3\lambda^2 & \lambda^3 \end{pmatrix} \tag{160}$$

More generally, and for $n < N$, the block J^n is a banded lower-triangular matrix having n nonzero subdiagonals below the main diagonal:

$$J^n = \begin{pmatrix}
\lambda^n & & & & & & & & & \\
n\lambda^{n-1} & \lambda^n & & & & & & & & \\
C_n^2\lambda^{n-2} & n\lambda^{n-1} & \lambda^n & & & & & & & \\
\vdots & \ddots & \ddots & \ddots & & & & & & \\
C_n^j\lambda^{n-j} & \cdots & C_n^2\lambda^{n-2} & n\lambda^{n-1} & \lambda^n & & & & & \\
\vdots & \ddots & & \ddots & \ddots & \ddots & & & & \\
1 & \cdots & C_n^j\lambda^{n-j} & \cdots & C_n^2\lambda^{n-2} & n\lambda^{n-1} & \lambda^n & & & \\
0 & \ddots & & \ddots & & \ddots & \ddots & \ddots & & \\
\vdots & \ddots & \ddots & & \ddots & & \ddots & \ddots & \ddots & \\
0 & \cdots & 0 & 1 & \cdots & C_n^j\lambda^{n-j} & \cdots & C_n^2\lambda^{n-2} & n\lambda^{n-1} & \lambda^n
\end{pmatrix} \tag{161}$$

where $C_n^j = \frac{n!}{j!(n-j)!}$. Evidently, each nonzero entry of the matrix J^n is one of the monomials appearing in the expansion of $(1+\lambda)^n$.

The first column vector of the matrix J^n is precisely the value achieved by ϵ^n if ϵ^0 is set to $\epsilon^1 = (1, 0, 0, \cdots, 0)^T$, that is the first vector of the canonical basis. Since one of its components (the $n+1$st) is equal to 1, its sup-norm (or maximum component in absolute value) is greater or equal to 1, and this, over at least $N-1$ iterations. Therefore, in the first $N-1$ iterations the process cannot be expected to be dissipative, even in the particular case where the spectral radius is equal to 0 ($\lambda = 0$).

For $n \geq N$, the above vector is truncated to the first N components. In particular, its last component becomes $C_n^{N-1}\lambda^{n-N+1}$. This term, for fixed N, is equal to the value at n of a monomial of degree $N-1$ times λ^n. The other components involve monomials of lesser degrees. Therefore, in the most general case, the vector ϵ^n is a general linear combination of the column-vectors of the matrix J^n and any of its components, say ϵ_j^n, is of the form $P_j^{N-1}(n)\lambda^n$ where $P_j^{N-1}$is a polynomial of degree $N-1$ depending on j. Consequently, as $n \to \infty$, $\|\epsilon^n\|_\infty \to 0$ only at the rate of $n^{N-1}\rho^n$. From this we conclude that relatively to the regular case for which the amplification matrix G can be diagonalized and the error tends to 0 at the same rate as ρ^n, the asymptotic convergence is slightly degraded by the additional factor n^{N-1}. Unfortunately, another form of degradation of the convergence, believed to be more severe than the first, is now going to be demonstrated. For this, let:

$$\tau_n = \|\epsilon^n\|_\infty = \max_{j=0,1,\cdots,N-1} \left(\xi_j^n\right) \tag{162}$$

in which

$$\xi_j^n = C_n^j \rho^{n-j} \tag{163}$$

in which again, $\rho = \rho(G) = |\lambda| < 1$ is the spectral radius. It turns out that the asymptotic convergence rate is significant only after a relatively large number of iterations.

Let ν be the smallest integer such that the sequence $\{\tau_n\}$ $(n \geq \nu)$ is monotone decreasing. Then

$$\lim_{N \to \infty} \frac{\nu}{N} = \frac{1}{1-\rho} \tag{164}$$

In this analysis, N is large and since $n > N$, n is also large, a fortiori. First examine the variation with j of ξ_j^n for fixed n. We have:

$$\frac{\xi_j^n}{\xi_{j-1}^n} = \frac{n-j+1}{j\rho} > 1, \quad \text{iff} \quad j < \frac{n+1}{\rho+1} \tag{164}$$

Therefore, for fixed n, the sequence $\{\xi_j^n\}$ $(j = 0, 1, \cdots)$ is initially monotone increasing, and only eventually decreases if values of j greater than $\frac{n+1}{\rho+1}$ exist (knowing that $j \leq N-1$). This leads us to analyze two cases separately:

1st case: $N < n < (\rho+1)N - 1$ (N large)
In view of (165), it appears that since the ratio $(n+1)/(\rho+1)$ is less than N, it is for a value j_0 of j close to that ratio that ξ_j^n achieves its maximum over j, which implies that:

$$\tau_n = \xi_{j_0}^n$$

The statement in (165) also indicates that if n increases of 1, the maximum will be achieved at $j_1 = j_0$, or possibly $j_0 + 1$. Therefore:

$$\tau_{n+1} = \xi_{j_0}^{n+1}, \quad \text{or possibly} \quad \xi_{j_0+1}^{n+1}$$

In both cases, using $j_0/n \approx 1/(\rho+1)$, one obtains that

$$\frac{\tau_{n+1}}{\tau_n} \approx \rho + 1 > 1$$

and therefore n is found insufficiently large for the sequence $\{\tau_k\}$ $(k \geq n)$ to be monotone-decreasing.

2nd case: $n \geq (\rho+1)N - 1$ (N large)
Then (165) indicates that the sequence ξ_j^n monotonically increases with j. Therefore, for all $k \geq n$,

$$\tau_k = \xi_{N-1}^k = C_k^{N-1} \rho^{k-N+1}$$

and

$$\frac{\tau_{k+1}}{\tau_k} = \frac{(k+1)\rho}{(k+1-N+1)} = \frac{\rho}{1-\frac{N-1}{k+1}}$$

and this ratio is less than 1 for all $k \geq n$ iff

$$n > \frac{N-2+\rho}{1-\rho} \sim \frac{N}{1-\rho}$$

□

In summary:
If a linear iteration has a defective amplification matrix G, still satisfying the spectral radius condition, $\rho < 1$ that insures convergence, then for a general initial guess, the asymptotic convergence will only be like $n^{N-1}\rho^n$, where N is the dimension of the larger Jordan block appearing in the reduction of the amplification matrix. Moreover, if the number N is large, indicating that a large number of eigenvectors are missing, the asymptotic convergence rate is meaningful only after a number of iterations of the order of $N/(1-\rho)$, a particularly severe degradation if N is very large or if ρ is close to unity, or both.

4. Kronecker Products and Sums

To facilitate the understanding of the reader unfamiliar with the Kronecker product notation, its definition is recalled here along with a few related notions.

One considers a two-dimensional uniform rectangular mesh of $N_x \times N_y$ grid-points. An arbitrary function $v(x,y)$ can be represented discretely by either one of the following two matrices whose entries are the nodal values of the function, $v_{jk} = v(x_j, y_k)$:

$$V' = \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1,N_y} \\ v_{21} & v_{22} & \dots & v_{2,N_y} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N_x,1} & v_{N_x,2} & \dots & v_{N_x,N_y} \end{pmatrix} \tag{165}$$

(dimension: $N_x \times N_y$), and

$$V'' = \begin{pmatrix} v_{11} & v_{21} & \dots & v_{N_x,1} \\ v_{12} & v_{22} & \dots & v_{N_x,2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1,N_y} & v_{2,N_x} & \dots & v_{N_x,N_y} \end{pmatrix} = V'^T \tag{166}$$

(dimension: $N_y \times N_x$), or equivalently by the following column-vectors (identified in what follows with $N_x N_y \times 1$ matrices):

$$[v]' = \left(v_{11}\ v_{12}\ \dots\ v_{1,N_y}\ v_{21}\ v_{22}\ \dots\ v_{2,N_y}\ \dots\ v_{N_x,1}\ v_{N_x,2}\ \dots\ v_{N_x,N_y}\right)^T \tag{167}$$

or

$$[v]'' = \left(v_{11}\ v_{21}\ \dots\ v_{N_x,1}\ v_{12}\ v_{22}\ \dots\ v_{N_x,2}\ \dots\ v_{1,N_y}\ v_{2,N_y}\ \dots\ v_{N_x,N_y}\ \right)^T \tag{168}$$

These two vectors are related by a permutation matrix P so that

$$[v]' = P\,[v]'' \tag{169}$$

The matrix P contains one element equal to unity in each column and each row; all other entries are equal to zero; furthermore,

$$P\,P^T = P^T\,P = I \tag{170}$$

We now recall the following two definitions (see e.g. [KRON]):

Definitions:

(i) Kronecker product: The Kronecker product of two arbitrary given matrices $M = \{m_{\alpha\beta}\}$ $(1 \leq \alpha \leq p, 1 \leq \beta \leq q)$ *and* $N = \{n_{\gamma\delta}\}$ $(1 \leq \gamma \leq r, 1 \leq \delta \leq s)$*, is the matrix of dimension* $pr \times qs$ *denoted by* $M \otimes N$ *and defined by:*

$$M \otimes N = \begin{pmatrix} m_{11}N & m_{12}N & \dots & m_{1q}N \\ m_{21}N & m_{22}N & \dots & m_{2q}N \\ \vdots & \vdots & \ddots & \vdots \\ m_{p1}N & m_{p2}N & \dots & m_{pq}N \end{pmatrix} \tag{171}$$

in which for any (α, β).

$$m_{\alpha\beta}N = \begin{pmatrix} m_{\alpha\beta}n_{11} & m_{\alpha\beta}n_{12} & \dots & m_{\alpha\beta}n_{1s} \\ m_{\alpha\beta}n_{21} & m_{\alpha\beta}n_{22} & \dots & m_{\alpha\beta}n_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ m_{\alpha\beta}n_{r1} & m_{\alpha\beta}n_{r2} & \dots & m_{\alpha\beta}n_{rs} \end{pmatrix} \tag{172}$$

(ii) Kronecker sum: The Kronecker sum of two arbitrary given square matrices $M = \{m_{\alpha\beta}\}$ $(1 \leq \alpha, \beta \leq p)$ *and* $N = \{n_{\gamma\delta}\}$ $(1 \leq \gamma, \delta \leq r)$ *is the square matrix of dimension* $pr \times pr$ *denoted by* $M \oplus N$ *and defined by*

$$M \oplus N = M \otimes I_r + I_p \otimes N \tag{173}$$

in which I_p *and* I_r *are the identity matrices of dimension* $p \times p$ *and* $r \times r$ *respectively.*

Note that if M' and N' are two other square matrices of analogous dimensions,

$$(M \otimes N)\,(M' \otimes N') = (MM' \otimes NN') \tag{174}$$

so that, if the matrices M and N are invertible,

$$(M \otimes N)^{-1} = M^{-1} \otimes N^{-1} \tag{175}$$

We now return to the vector representations, (168) and (169), of a given function $v(x, y)$. If this function has separable variables, that is, if it is of the form

$$v(x, y) = f(x)\,g(y) \tag{176}$$

the corresponding vectors $[v]'$ and $[v]''$ have the structure of Kronecker products, that is,

$$\begin{aligned} [v]' &= [f] \otimes [g] \\ [v]'' &= [g] \otimes [f] \end{aligned} \tag{177}$$

where

$$[f] = (f(x_1)\, f(x_2) \ldots f(x_{N_x}))^T$$
$$[g] = (g(y_1)\, g(y_2) \ldots g(y_{N_y}))^T \tag{178}$$

In this case, we refer to the vectors $[v]'$ and $[v]''$ as "separable vectors". This is in particular the case of the usual discrete Fourier modes whose components are the nodal values of functions of the type $\exp(i(\alpha x + \beta y))$ (periodic case) or $\sin(\alpha x)\sin(\beta y)$ (Dirichlet, or Dirichlet-Neumann case).

Now consider a polynomial $\mathcal{P}(X, Y)$ of the symbols X and Y, the differential operator

$$\mathcal{D} = \mathcal{P}\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right) \tag{179}$$

and the image function $w(x, y)$ of the function $v(x, y)$,

$$w = \mathcal{D}v \tag{180}$$

For example, if $\mathcal{P}(X, Y)$ is the first-degree polynomial $aX + bY$, the operator $\mathcal{D}$ becomes

$$\mathcal{D} = a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y} \tag{181}$$

which is studied in Section III.2.

If we choose to discretely represent the functions v and w by "maps", that is matrices V' and W' as in (166), the representation of the finite-difference analog of the operator $\mathcal{D}$ will be a tensor of order 4. The corresponding notation would be somewhat inconvenient, and we prefer to introduce the vector representations $[v]'$ and analogously $[w]'$. Then let the matrices $(\Delta x^{-1})\delta_x$ (of dimension $N_x \times N_x$) and $(\Delta y^{-1})\delta_y$ (of dimension $N_y \times N_y$) denote matrix representations of finite-difference analogs of the operators $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ respectively (boundary conditions included, see APPENDIX, Section 1. In the two-dimensional problem, these operators are respectively represented by the $N_xN_y \times N_xN_y$-matrices $(\Delta x^{-1})\delta_x \otimes I_y$ and $I_x \otimes (\Delta y^{-1})\delta_y$ if I_x and I_y are the identity matrices of dimension $N_x \times N_x$ and $N_y \times N_y$ respectively. Then, a possible finite-difference approximation of the operator $\mathcal{D}$ is represented by the matrix:

$$\mathcal{D}_h = \mathcal{P}\big((\Delta x^{-1})\delta_x \otimes I_y\,,\, I_x \otimes (\Delta y^{-1})\delta_y\big) \tag{182}$$

In particular, if $\mathcal{D}$ is given by (77), letting

$$\mathcal{D}_x = \nu_x\, \delta_x$$
$$\mathcal{D}_y = \nu_y\, \delta_y \tag{183}$$

(where $\nu_x = a/\Delta x$, $\nu_y = b/\Delta y$) the matrix $\mathcal{D}_h$ has the following Kronecker sum structure:

$$\mathcal{D}_h = \mathcal{D}_x \otimes I_y + I_x \otimes \mathcal{D}_y = \mathcal{D}_x \oplus \mathcal{D}_y \tag{184}$$

In the remaining of this appendix, we study the eigensystem of a matrix A having a similar Kronecker sum structure

$$A = A_x \oplus A_y \tag{185}$$

assuming the matrices A_x and A_y (of dimension $N_x \times N_x$ and $N_y \times N_y$ respectively) known. The eigenvalues are first identified, and we then examine whether the (true or generalized) eigenvectors are "separable".

For this, we first introduce the reduction of the matrices A_x and A_y to Jordan forms which writes:

$$\begin{aligned} A_x &= U_x \, J_x \, U_x^{-1} \\ A_y &= U_y \, J_y \, U_y^{-1} \end{aligned} \tag{186}$$

This reduction is always possible and it identifies with the diagonalization when the matrices A_x and A_y are not defective; in this case, the matrix J_x (resp. J_y) is diagonal and contains the eigenvalues of the matrix A_x (resp. A_y), and the column vectors of the matrix U_x (resp. U_y) are the associated eigenvectors which form a complete set.

Otherwise, if the matrix A_x (resp. A_y) is defective, the matrix J_x (resp. J_y) is block-diagonal, and at least one diagonal block is a lower bi-diagonal Jordan sub-block of the type

$$\begin{pmatrix} \lambda_i & & & & \\ 1 & \lambda_i & & & \\ 0 & 1 & \lambda_i & & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & 1 & \lambda_i \end{pmatrix} \tag{187}$$

(see APPENDIX, Section 3). If unity appears below the (m, m) element in J_x (resp. J_y), the m-th column vector of the matrix U_x (resp. U_y) is not a true but a so-called generalized eigenvector of the matrix A_x (resp. A_y).

Now observe that

$$\begin{aligned} A_x \oplus A_y &= \left(U_x \, J_x \, U_x^{-1}\right) \otimes \left(U_y \, I_y \, U_y^{-1}\right) + \left(U_x \, I_x \, U_x^{-1}\right) \otimes \left(U_y \, J_y \, U_y^{-1}\right) \\ &= U_{x \otimes y} \, J_{x \oplus y} \, U_{x \otimes y}^{-1} \end{aligned} \tag{188}$$

where

$$U_{x\otimes y} = U_x \otimes U_y, \quad U_{x\otimes y}^{-1} = U_x^{-1} \otimes U_y^{-1}$$
$$J_{x\oplus y} = J_x \oplus J_y = J_x \otimes I_y + I_x \otimes J_y \tag{189}$$

In all cases, the matrix $J_{x\oplus y}$ is lower triangular, and thus contains in its main diagonal the eigenvalues of the matrix A which are the numbers:

$$(\alpha_{x\oplus y})_{j,k} = (\alpha_x)_j + (\alpha_y)_k \qquad (j = 1,2,\ldots,N_x\,;\ k = 1,2,\ldots,N_y) \tag{190}$$

where $\{(\alpha_x)_j\}$ $(j = 1,2,\ldots,N_x)$ and $\{(\alpha_y)_j\}$ $(k = 1,2,\ldots,N_y)$ are the eigenvalues of the matrices A_x and A_y respectively.

We now turn to the description of the structure of the eigenvectors, and for this, four cases are distinguished.

Firstly, if both matrices A_x and A_y can be diagonalized, both matrices J_x and J_y are diagonal, and evidently the matrix $J_{x\oplus y}$ is also diagonal. In this "standard" case, (189) realizes the diagonalization of the matrix A whose eigenvectors form a complete set and are immediately identified with the column vectors of the matrix $U_{x\otimes y}$; since each one of them is the Kronecker product of a column vector of the matrix U_x by the column vector of the matrix U_y, all the eigenvectors are found separable.

Secondly, if the matrix A_x is diagonalizable, but the matrix A_y defective, the matrix J_x is diagonal, but the matrix J_y contains at least one defective Jordan sub-block. It is then immediate that the matrix $J_{x\oplus y}$ although not diagonal is already in Jordan form. Consequently, (189) realizes the reduction to Jordan form of the matrix A, which is defective, and whose true and generalized eigenvectors are the column vectors of the matrix $U_{x\otimes y}$; again, all of them are separable.

Thirdly, we examine the symmetrical case where instead the matrix A_x is defective, whereas the matrix A_y can be diagonalized. For this case, we consider again the permutation matrix P of (170). The similarity transform associated with it has the effect of interchanging the roles played by the coordinates x and y, in particular in the vector ordering of the nodal values of functions of the two variables. Consequently, given any two matrices B_x and B_y of dimensions $N_x \times N_x$ and $N_y \times N_y$ respectively, we have

$$P^T\,(B_x \otimes B_y)\,P = B_y \otimes B_x \tag{191}$$

This implies that

$$P^T\,(J_x \oplus J_y)\,P = J_y \oplus J_x = J_{y\oplus x} \tag{192}$$

As in the previous case, the matrix $J_{y\oplus x}$ is a Jordan form, and since $P\,P^T = I$,

$$\begin{aligned} A &= U_{x\otimes y}\,P\,P^T\,J_{x\oplus y}\,P\,P^T\,U_{x\otimes y}^{-1} \\ &= U'\,J_{y\oplus x}\,U'^{-1} \end{aligned} \tag{193}$$

where the (generalized) eigenvectors are found to be the column vectors of the matrix

$$U' = U_{x\otimes y}\,P = (U_x \otimes U_y)\,P \tag{194}$$

Since in the above equation, the post-multiplication by the permutation matrix P has the mere effect of permuting the columns, it does not destroy their Kronecker product structure: the (true and generalized) eigenvectors are separable in this case also, as expected by symmetry.

Lastly, if both matrices A_x and A_y are defective, nonzero elements appear in more than one subdiagonals of the matrix $J_{x\oplus y}$; one expects that a permutation of the rows and columns is not sufficient to reduce this matrix to Jordan form, and that certain true as well as generalized eigenvectors are not separable. This can be verified with a counter example, e.g. let

$$A_x = J_x = A_y = J_y = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{195}$$

so that

$$A = A_x \oplus A_y \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 1 & 2 \end{pmatrix} \tag{196}$$

As expected all the diagonal entries of the matrix A are equal to 2, the sole eigenvalue. The matrix is found defective with only two true eigenvectors,

$$U_1 = (0\ 0\ 0\ 1)^T\ ,\ U_2 = (0\ 1\ -1\ 0)^T \tag{197}$$

The vector U_1 is indeed separable since equal to $\begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. However the vector U_2 is not separable.[1]

In this last case, even some of the true eigenvectors have a distinctive structure from the usual Fourier-type modes that are separable; in this sense, the algebraic problem is truly two-dimensional.

[1] The system $u_1v_1 = 0,\ u_1v_2 = 1,\ u_2v_1 = -1,\ u_2v_2 = 0$ has no solution.

In conclusion, if at least one of the matrices A_x or A_y is diagonalizable, the (true and generalized) eigenvectors of the Kronecker sum $A_x \oplus A_y$ are separable, that is, each one is the Kronecker product of an $N_x \times 1$ vector by an $N_y \times 1$ vector; otherwise, only some of them have this structure.[2]

[2] If the matrices A_x and A_y admit p and q true eigenvectors respectively, at least pq true eigenvectors of the matrix A are separable.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique