Spectral Analysis of Defect-Correction Algorithms for Hyperbolic Equations

Marie-Claude Ciccoli, Jean-Antoine Desideri

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Spectral analysis of Defect-Correction algorithms for hyperbolic equations

M.C. Ciccoli & J.A. Désidéri

N 2831

Mars 1996

THÈME 4

Rapport de recherche

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

Spectral analysis of Defect-Correction algorithms for hyperbolic equations

M.C. Ciccoli[*] & J.A. Désidéri[**]

Thème 4 — Simulation
et optimisation
de systèmes complexes
Projet Sinus

Rapport de recherche n2831 — Mars 1996 — 40 pages

Abstract: In this report, variants are proposed to the Defect Correction algorithm analyzed for hyperbolic equations in [1] [2]. Potentially more efficient preconditioners are identified for a class of second-order partially-upwind approximation schemes including one that is third-order; these preconditioners are constructed by averaging the first-order upwind operator with the central-difference operator. Additionally, it is shown that the implicit equations can be solved by relaxation; for this purpose, a 3-stage iterative algorithm is proposed.

Key-words: Hyperbolic equations, upwind approximations, Defect Correction algorithm, eigenvalues, spectral radius, optimal preconditioners.

(Résumé : tsvp)

[*] Université de Rouen, URA CNRS 230/Coria, 76821 Mont Saint-Aignan Cedex (ciccoli@coria.fr),

[**] INRIA, 2004 Route des Lucioles, BP 93, F-06902 Sophia Antipolis Cedex – France (desideri@sophia.inria.fr).

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

Analyse spectrale d'algorithmes de Résidus Corrigés pour les équations hyperboliques

Résumé : Dans ce rapport, on propose des variantes de l'algorithme des Résidus Corrigés précédemment analysé pour les équations hyperboliques en [1] [2]. On identifie des préconditionneurs potentiellement plus efficaces d'une classe de schémas d'approximation partiellement décentrée du second ordre contenant un schéma du troisième ordre; ces préconditionneurs sont construits en moyennant l'opérateur décentré d'ordre un à l'opérateur centré. On montre également qu'on peut résoudre les équations implicites par relaxation; on propose pour cela un algorithme itératif à 3 sous-pas.

Mots-clé : Équations hyperboliques, approximations décentrées, algorithme des Résidus Corrigés, valeurs propres, rayon spectral, préconditionneurs optimaux.

Contents

1 Introduction – Model Problem

The two major steps in the numerical solution of a time-dependent problem governed by a (set of) partial-differential equation(s) (PDE) are the definition of an appropriate (consistent) spatial approximation scheme and the construction of an efficient and/or accurate algorithm to integrate forward in time the set of discrete equations. The two steps are evidently linked; in the linear case for example, when the solution algorithm is meant to be a pseudo-time-integration iteration, its election is usually strongly guided by the spectral properties of the approximation scheme, whose identification, for at least a simple model problem, is thus essential.

In this theoretical study, we have in mind future applications to inviscid gas dynamics. There, a standard approach to solve steady problems consists in integrating the Euler equations forward in time until the solution reaches the steady state. Thus, the time integration is used somewhat artificially to construct an iterative solution method and it is important to assess its performance with respect to cost-efficiency.

For the purpose of making a linear analysis of different potential schemes, we are here considering, after many others and following particularly the analysis of [1] [2], the quarter-plane pure-advection problem :

$$\begin{cases} u_t + c\,u_x = 0 \quad (c>0) \quad (\,x \in [0,1]\,;\; t>0\,) \\ u(x,0) = u^0(x) \\ u(0,t) = \text{ const.} \end{cases} \tag{1}$$

This simple PDE contains only one partial derivative with respect to x in the quantity

$$A(u) = c\,u_x \tag{2}$$

Because such operator A involves only one wavespeed c which is assumed positive, upwind differences are constructed as backward differences. Thus, identifying in our notation finite-difference operators with their matrix analogs, the operator A is here chosen to be approximated over a uniform mesh by a second-order finite-difference operator of the following form

$$B_\beta = \frac{1}{\Delta x}\delta_2^\beta = \frac{1}{\Delta x}\left[(1-\beta)\,\delta_2^C + \beta\,\delta_2^U\right] \tag{3}$$

where

$$\delta_2^C = \text{Trid}\ \left(-\frac{1}{2}, 0, \frac{1}{2}\right) \tag{4}$$

is the central-difference operator,

$$\delta_2^U = \text{Pentad}\ \left(\frac{1}{2}, -2, \frac{3}{2}, 0, 0\right) \tag{5}$$

is the second-order fully-upwind operator, and β is a parameter controlling the degree of upwinding in this approximation. In particular, for $\beta = 1/2$, one gets a Fromm-type "half-upwind" differencing scheme, and for $\beta = 1/3$, the 3rd-order accurate upwind-biased scheme.

For the solution of (1), one may use an implicit time integration scheme (known to be unconditionally stable) such as :

$$[I + \Delta t\, A'_\theta(u^n)] \left(u^{n+1} - u^n\right) = -\Delta t\, B_\beta(u^n) \tag{6}$$

in which $A'_\theta(u)$, to satisfy the consistency condition, should be the Jacobian of an approximation $A_\theta(u)$ of $B_\beta(u)$. More generally, it is some appropriate preconditioner controlling the stability when large timesteps are used. In the linear context, the operator A_θ and its Jacobian A'_θ admit the same matrix representation, and we consider in this article the following particular choice which generalizes that of [1] [2] :

$$A_\theta = \frac{1}{\Delta x}\left[(1-\theta)\,\delta_1 + \theta\,\delta_2^C\right] \tag{7}$$

where δ_1 is the first-order upwind-difference operator which is one-sided (since $c > 0$) :

$$\delta_1 = \text{Trid}\ (-1, 1, 0)$$

and θ is a parameter to be optimized.

We would like to first give some justification of the present choice of preconditioner A_θ. In (future) practical applications to nonlinear hyperbolic systems involving wavespeeds of different signs, upwind schemes rely on local linearization and diagonalization by means for example, of flux or flux-difference splitting. Even well mastered today for the case of the Euler equations (see e.g. [3] [4]), these operations remain delicate and costly. In fact, in usual implementations, the right-hand side (RHS) of (6) is computed by direct discretization of the quantity $A(u)$ and the matrix B_β is not calculated. Inversely, the matrix structure of the preconditioner appearing on the left-hand side is actually computed. This is one reason why the simple first-order accurate difference operator alone ($\theta = 0$) is very often preferred to the exact Jacobian of the RHS in the construction of the preconditioner, since this does not modify the converged (or steady-state) solution. Another advantage of using $A_\theta = \delta_1$ is that the matrix is (marginally) diagonally dominant. Thus, in that particular case, (6) can be solved by classical relaxation. The disadvantage of this simplification is that quadratic convergence of Newton's method is lost when infinite timesteps are employed; in addition the convergence is then pathological for $\beta = 0$ (central-differencing of RHS) or 1 (fully-upwind differencing of RHS) [1] [2]. The more general form of preconditioner adopted in (7) reflects our attempt to make it closer to the operator in the RHS by approaching more accurately a second-order operator without actually forming the matrix associated with the second-order upwind scheme. In addition, in the slightly more representative case where the sign of c would change, the first-order upwind-difference operator would not be uniformly one-sided. Thus in that case, the introduction of central differencing in (7) would not complicate the already-tridiagonal preconditioning matrix.

When infinite timesteps can be used stably, the implicit algorithm becomes

$$A_\theta \, u^{n+1} = A_\theta \, u^n - B_\beta \, u^n. \tag{8}$$

In this form, the iteration is identified to a particular version of the Defect-Correction algorithm [5]. For $\theta = 0$, this algorithm has been thoroughly investigated in [1] [2]. A major result of that analysis is that the spectral radius of the corresponding iterative algorithm is independent of the mesh size Δx and of the upwinding parameter β, and is equal to $1/2$. The aim of the present contribution is to present an analysis of the cases corresponding to $\theta \neq 0$ and thus investigate possible variants of the basic Defect Correction algorithm. For this, we introduce the amplification matrix G for which

$$u^{n+1} = G \, u^n + b \tag{9}$$

for some constant vector b that accounts for the known right-boundary value of u. One has :

$$\boxed{G = I - A_\theta^{-1} B_\beta = I - \left[(1-\theta)\,\delta_1 + \theta\,\delta_2^C\right]^{-1} \left[(1-\beta)\,\delta_2^C + \beta\,\delta_2^U\right]} \qquad (10)$$

Let $\{g_m\}$ (m : mode index) denote the eigenvalues of matrix G, and ρ the spectral radius so that

$$\rho = \max_m \mid g_m \mid \,, \qquad (11)$$

in which

$$g_m = 1 - \lambda_m \qquad (12)$$

where λ_m is the generic eigenvalue of the following generalized eigenproblem :

$$\boxed{\left\{\left[(1-\beta)\,\delta_2^C + \beta\,\delta_2^U\right] - \lambda\left[(1-\theta)\,\delta_1 + \theta\,\delta_2^C\right]\right\}\ \mathrm{u} = 0} \qquad (13)$$

The identification of the best preconditioner in our formulation relies on the solution of the above generalized eigenproblem. The formal solution to this problem is given in the next section and the sensitivity of the iteration convergence rate assessed. Next, we examine how the preconditioned set of discrete equations can itself be solved iteratively by using an appropriate "annihilation" strategy. Finally, a numerical study of the spectral radius corresponding to an analogous two-dimensional model problem is presented.

2 One-Dimensional Analysis

2.1 Eigenvalues

For the one-dimensional model problem, the spectrum of eigenvalues of matrix G has been calculated formally for $\theta \neq 0$ (see appendix A.1). If N denotes the number of points of the discretization, $\omega_m = m\pi/N$, and

$$\theta_m = \frac{3 - \sqrt{1 + 8\beta - 4\beta \sin^2 \omega_m (1-\beta)}}{2 + \beta \sin^2 \omega_m}$$

$(m = 1, .., N-1)$, these eigenvalues are given by :

$$\boxed{g_m = \frac{-B_m + 2i\cos\omega_m\sqrt{-\delta_m}}{(2-\theta)^2}}$$

if $0 \leq \theta < \theta_m$, and

$$\boxed{g_m = \frac{-B_m + 2\cos\omega_m\sqrt{\delta_m}}{(2-\theta)^2}}$$

if $\theta_m \leq \theta < 1$, where :

$$\begin{array}{l} B_m = -(2-\theta)(1-2\beta-\theta) + 2\beta\theta\cos^2\omega_m \\ \delta_m = -(\beta^2\sin^2\omega_m + 2\beta)\theta^2 + 6\beta\theta - 4\beta(1-\beta) \end{array}$$

Fig. 1 and 2 illustrate this result for several values of θ in the case where $N = 30$, $\beta = 1/3$. (A verification of these formulae was made by direct numerical computation of the eigenvalues using a library routine.)

2.2 Spectral radius

The spectral radius of the amplification matrix is the parameter that controls the algorithm convergence rate. The optimum preconditioner is associated with the minimum spectral radius.

Firstly, in order to optimize θ for a fixed value of $\beta = 1/3$, the values of the spectral radius ρ for $N = 9$ and for different values of θ have been collected in the following table :

θ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ρ	0.473	0.447	0.418	0.387	0.353	0.315	0.618	0.995	1.4	1.88	2.5

Table 1: Spectral radius depending on θ ($N = 9$, $\beta = 1/3$)

In this case where N is finite, the spectral radius achieves one minimum visibly attained for a certain θ between 0.5 and 0.7.

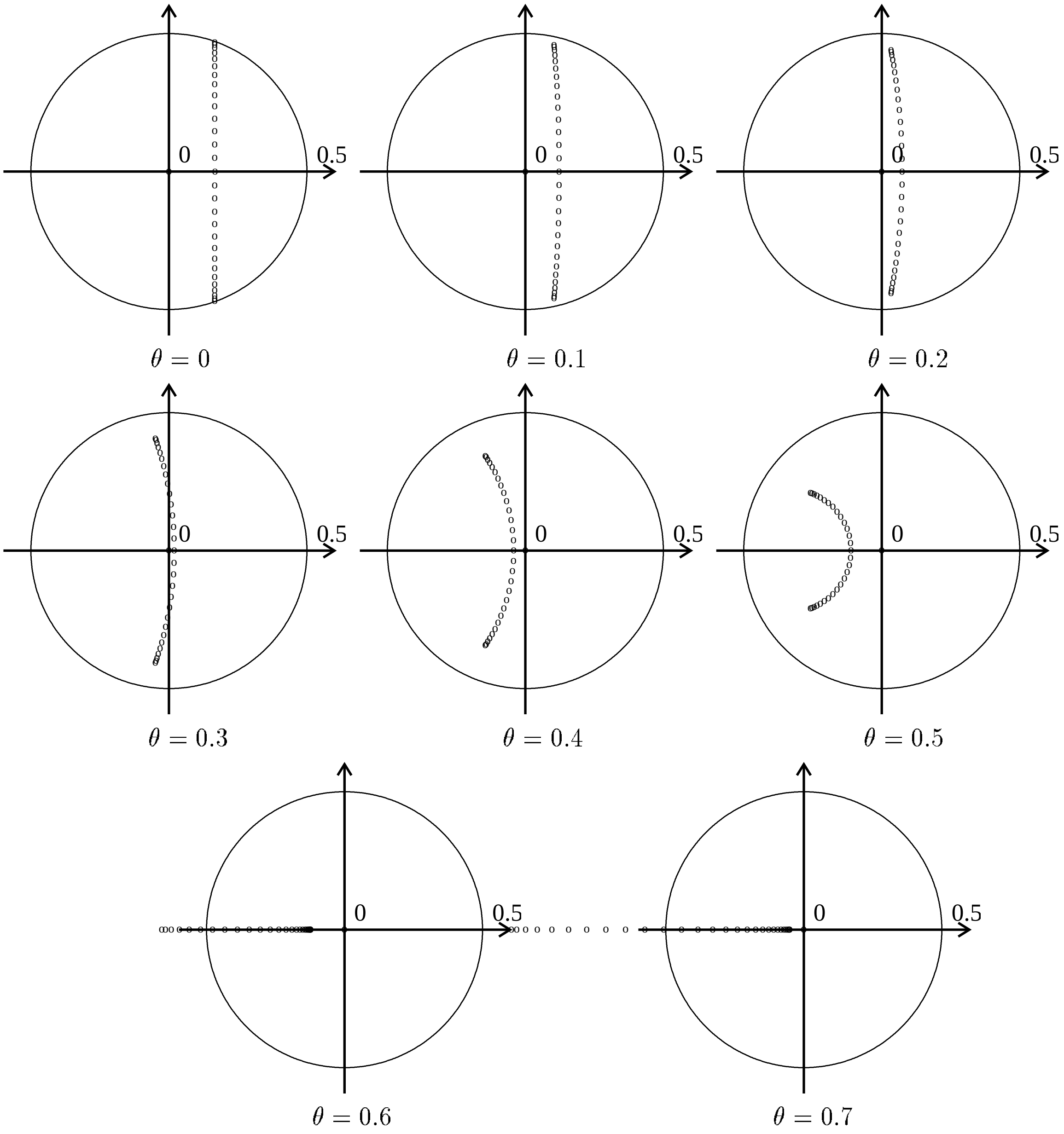
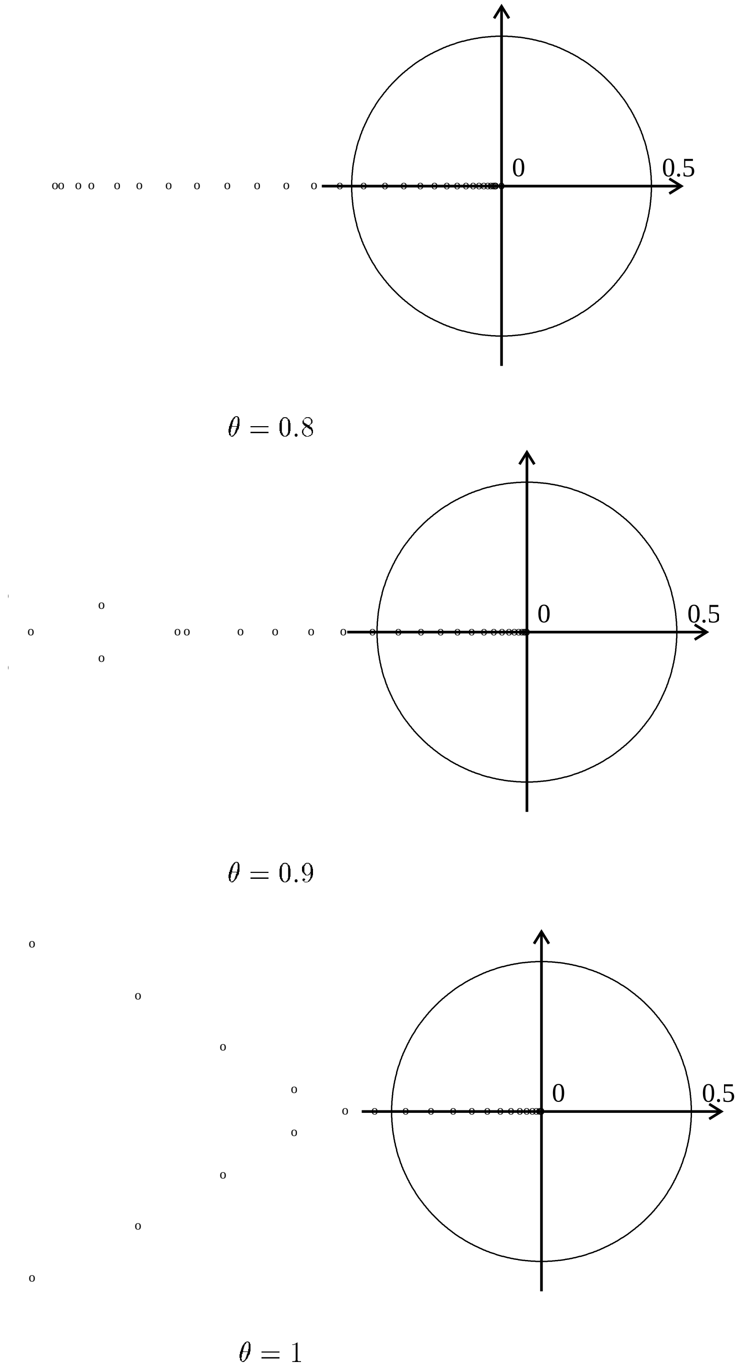


Figure 1: Spectrum of matrix G for different values of θ ($\beta = 1/3$)



$\theta = 0.8$

$\theta = 0.9$

$\theta = 1$

Figure 2: Spectrum of matrix G for different values of θ ($\beta = 1/3$)

Secondly, in the following table, the values of the spectral radius corresponding to different numbers of discretization points are given for a fixed value of $\beta = 1/3$ and $\theta = 1/2$:

N	4	9	19	29
ρ	0.2484	0.3155	0.3293	0.3316

Table 2: Spectral radius depending on the number of points
($\beta = 1/3$, $\theta = 1/2$)

We observe that the spectral radius rapidly reaches a limit value as N increases. Thus, in the remaining of this section, we only consider the "limit spectral radius" (as $N \to \infty$) which is calculated in Appendix A.2. Given β and θ, the maximum eigenvalue modulus in the subset of eigenvalues for which $\theta \geq \theta_m$ is given by :

$$\rho_1(\beta,\theta) = \frac{-(2-\theta)(1-\theta) + 4\beta + 2\sqrt{\beta\,(4\beta - 2(2-\theta)(1-\theta))}}{(2-\theta)^2}$$

and the maximum eigenvalue modulus in the subset of eigenvalues for which $\theta \leq \theta_m$ is instead :

$$\rho_2(\beta,\theta) = \begin{cases} \sqrt{(1-\theta)^2 + 4\beta(\beta+\theta-1)}/(2-\theta) & \text{if } \beta+\theta-1 \geq 0 \\ (1-\theta)/(2-\theta) & \text{if } \beta+\theta-1 < 0 \end{cases}$$

Of course, the (limit) spectral radius $\rho(\beta,\theta)$ is the following maximum :

$$\rho(\beta,\theta) = \max\{\rho_1(\beta,\theta), \rho_2(\beta,\theta)\} \tag{14}$$

In particular, for $\theta = 0$, as expected, one gets $\rho(\beta, 0) = 1/2$. To illustrate this result, the variation of the spectral radius ρ with θ is plotted on Fig. 3 for $\beta = 1/3$ and $\beta = 1/2$.

2.3 Optimal preconditioner

Our aim is to identify, for a given β, the best preconditioner A_θ. Refering to Appendix A.3, the optimal θ that minimizes the spectral radius of G is :

$$\boxed{\theta^*(\beta) = \frac{3-\sqrt{1+8\beta}}{2}}$$

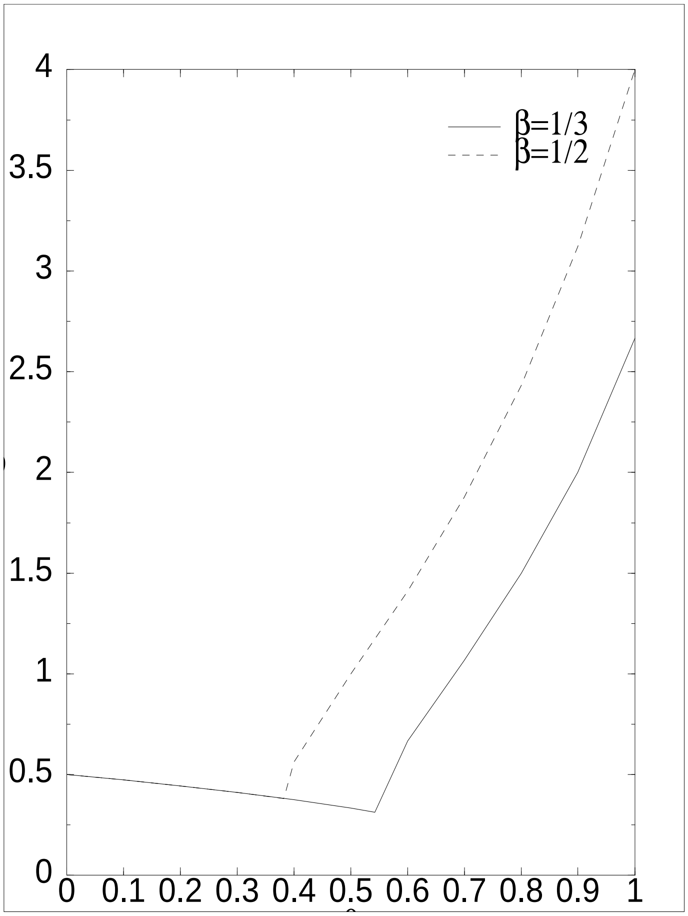


Figure 3: Spectral radius depending on θ ($\beta = 1/3$ and $\beta = 1/2$)

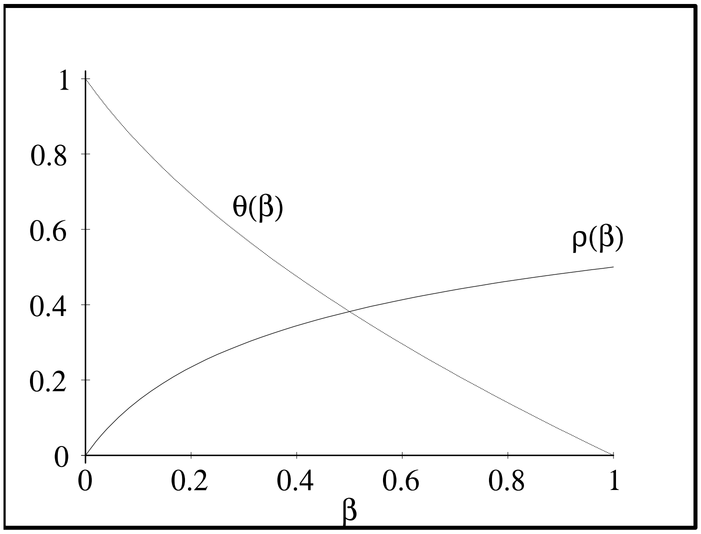


Figure 4: Optimal θ (θ^*) and optimal spectral radius (ρ^*) depending on β

and the corresponding spectral radius is :

$$\boxed{\rho^*(\beta) = \rho(\beta, \theta^*(\beta)) = \frac{\sqrt{1+8\beta}-1}{\sqrt{1+8\beta}+1}}$$

On Fig. 4, the optimal θ (θ^*) and spectral radius (ρ^*) is plotted versus β. We observe that for $\beta = 0$, the best preconditioner is obtained for $\theta = 1$. This is not surprising, since the present formulation permits the central difference operator to be preconditioned by itself and this is trivially optimal ($\rho = 0$). For $\beta = 1$, the optimal preconditioner is obtained for $\theta = 0$, which means that the second-order upwind difference operator is optimally preconditioned by the first-order upwind difference operator, which is less intuitive. For $\beta = 1/3,\ 1/2,\ 2/3$ the following optimal values have been found :

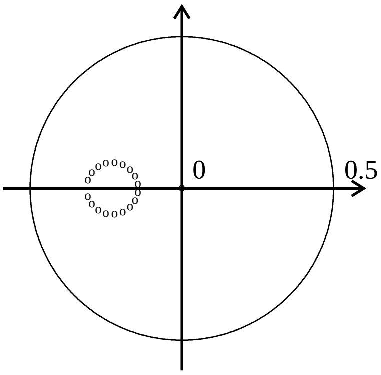


Figure 5: Spectrum of G for $\beta = 1/3$ and $\theta = \theta^*$

β	1/3	1/2	2/3
θ^*	0.5425	0.3819	0.2417
ρ^*	0.3138	0.3819	0.4313

Table 3: Optimal θ and optimal spectral radius depending on β

In the most interesting case $\beta = 1/3$, the spectral radius decreases from $1/2$ (obtained with $\theta = 0$) to 0.3138 when the optimal preconditioner is used. In the case $\beta = 1/2$, it decreases from $1/2$ to 0.3819. Therefore, in both cases, the convergence of the algorithm is somewhat accelerated by the use of the optimal preconditioner.

The spectrum of G corresponding to $\beta = 1/3$ and $\theta = \theta^*$ is plotted on Fig. 5. Note that in general (for any β), if $\theta = \theta^*(\beta)$, the nonzero eigenvalues of G are situated in the complex plane on the circle of center $C_G = -\beta + 4\beta^2/\left(1 + 4\beta + \sqrt{1+8\beta}\right)$ and radius $R_G = \beta\left(3 - \sqrt{1+8\beta}\right) / \left(1 + 4\beta + \sqrt{1+8\beta}\right)$.

Note that the parameters θ and ρ converge rapidly to their limits as the number of discretization points N increases, as shown on Figs. 6 and 7 for θ and ρ respectively. For example, for $N = 29$ the relative difference between the

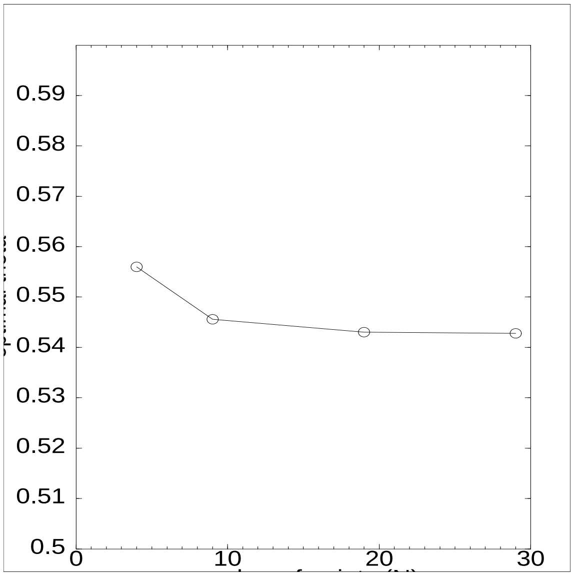


Figure 6: Optimal θ depending on the number of points; $\beta = 1/3$

optimal value for finite N and the limit θ^* is near 4 10^{-4} and correspondingly for the spectral radius near 5 10^{-3}.

2.4 Convergence acceleration

Examining again Fig. 5, it is remarkable that the subset of the spectrum made of the nonzero eigenvalues of the amplification matrix G is very localized, and in the neighborhood of a real number. In such a case, after application of the basic iteration to remove the error components along the eigendirections associated to zero eigenvalues, the iteration can easily be accelerated by over (or under) relaxation which can be viewed as a particular case of application of an eigenmode annihilation technique (see for example [6]). Without changing the fixed point, the iteration is then replaced by :

$$u^{n+1} = (I - \tau \mathsf{A})\ u^n + \tau b \tag{15}$$

in which the matrix A is defined as $\mathsf{A} = I - G$, and the number τ is the relaxation parameter (or pseudo-timestep). Since, the eigenvalues of A are

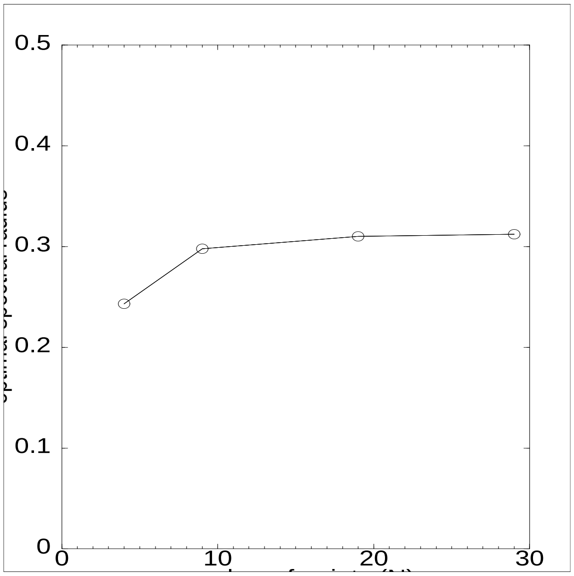


Figure 7: Optimal spectral radius depending on the number of points; $\beta = 1/3$

situated on a circle of center $C_A = 1 - C_G = 1 + \beta - 4\beta^2 / \left(1 + 4\beta + \sqrt{1+8\beta}\right)$ and radius $R_A = R_G$ (given above), i.e. a small number, it is natural to pick :

$$\tau = 1/C_A \tag{16}$$

which is a real number.

As a result, the spectral radius of the new iteration is given by :

$$f(\beta) = \max_{m=1,..,N-1} |1 - \lambda_m / C_A| \tag{17}$$

where $\{\lambda_m\}$ $(m = 1, .., N-1)$ are the known eigenvalues of A. One has :

$$\begin{aligned} |1 - \lambda_m / C_A| &= |1 - (C_A + R_A \cos\,\omega_m + \mathrm{i}\, R_A \sin\,\omega_m)\,/C_A| = \frac{R_A}{C_A} \\ &= \frac{\dfrac{\beta(3-\sqrt{1+8\beta})}{1+4\beta+\sqrt{1+8\beta}}}{1+\beta-\dfrac{4\beta^2}{1+4\beta+\sqrt{1+8\beta}}} = \frac{\beta(3-\sqrt{1+8\beta})}{1+5\beta+(1+\beta)\sqrt{1+8\beta}}. \end{aligned}$$

Thus

$$\boxed{f(\beta) = \frac{\beta(3-\sqrt{1+8\beta})}{1+5\beta+(1+\beta)\sqrt{1+8\beta}}.}$$

For example, for $\beta = 1/3$, the reduction factor is $f(1/3) \simeq 0.07$. This implies that only 5 iterations are sufficient to reduce the error of 6 orders of magnitude.

3 Solution of the Preconditioned Implicit Equations

After optimizing the global iteration (or pseudo-temporal integration) that yields the sequence $\{u^n\}$, we now turn to the question of how is the linear system solved at a given iteration. The system to be solved is :

$$A_\theta \; v = g$$

where : $v = u^{n+1}$ and $g = (A_\theta - B_\beta)\, u^n$.

One possibility is to solve by a multigrid technique as in [7] which is efficient but complex in applications other than model problems. Alternately, we propose here to solve the system by relaxation, using again but in a more general form, a strategy of eigenmode annihilation. Here, the eigenvalues of A_θ are situated on a segment in the complex plane :

$$\lambda_m = 1 - \theta + i\,\sqrt{\theta(2-\theta)}\,\cos\,\omega_m$$

(see Fig. 8). Following [6], two complex conjugate pseudo-timesteps $\tau, \overline{\tau}$ are selected such that their inverses are representative of the above spectrum, namely :

$$\tau = \left(1 - \theta \pm r\sqrt{\theta(2-\theta)}\,i\right)^{-1}$$

where $r \in [0,1]$. Then, without changing the fixed point, an iterative cycle made of two substeps is constructed :

$$\begin{cases} v^{(\alpha+1)} = (I - \tau A_\theta)\, v^{(\alpha)} + \tau g \\ \\ v^{(\alpha+2)} = (I - \overline{\tau} A_\theta)\, v^{(\alpha+1)} + \overline{\tau} g \end{cases}$$

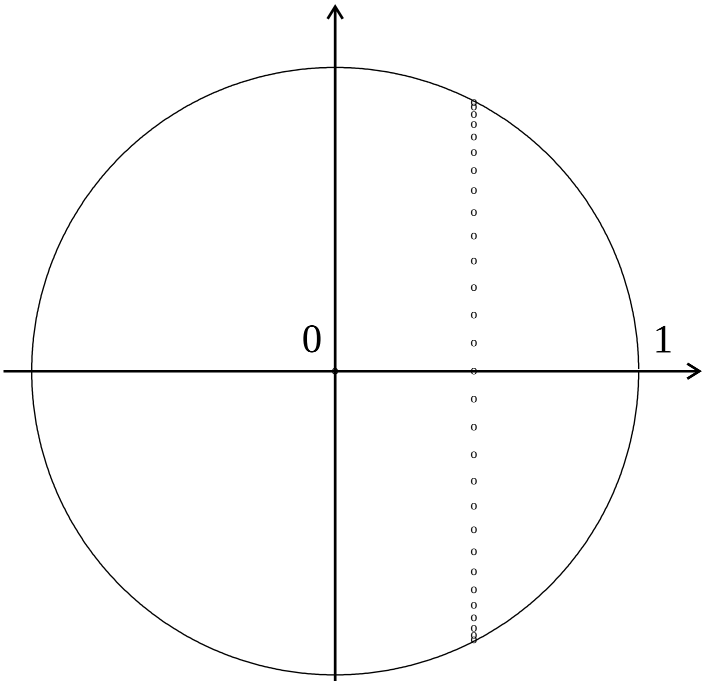


Figure 8: Spectrum of A_θ for $\beta = 1/3$ and $\theta = \theta^*$

The two substeps being conjugate to one another, it is easy to show that the resulting cycle can be implemented as a predictor-corrector sequence using real arithmetics only [6].

As a result of the above definitions, one application of the cycle has the effect of reducing the error component along the mth-eigenmode of the following factor :

$$g^{'}_m(r) = (1-\tau\lambda_m)(1-\overline{\tau}\lambda_m) = \frac{(r^2-\gamma_m^2)\theta(2-\theta)}{(1-\theta)^2+r^2\theta(2-\theta)}$$

where $\gamma_m = \cos\,\omega_m$. To simplify the analysis, γ_m is replaced by a parameter γ varying continuously in the interval [0,1] and $g^{'}_m(r)$ by :

$$g^{'}(r,\gamma) = \frac{(r^2-\gamma^2)\theta(2-\theta)}{(1-\theta)^2+r^2\theta(2-\theta)}.$$

Optimality is achieved by solving the following min-max problem :

$$\min_{r\in[0,1]}\max_{\gamma\in[0,1]}|g^{'}(r,\gamma)|$$

whose solution is given by :

$$r^* = \frac{1}{\sqrt{2}} \qquad \text{and} \qquad g^{'}_m(r^*) = \frac{1-2\gamma_m^2}{1+2\varpi^2}$$

where $\varpi = (1-\theta)\,/\sqrt{\theta(2-\theta)}$, yielding the following reduction factor :

$$\max_{m=1,..,N-1}|g^{'}_m| = |g^{'}_1| = \frac{1-2\gamma_1^2}{1+2\varpi^2} = \frac{\cos 2\omega_1}{1+2\varpi^2}.$$

Thus, in the limit $N \to \infty$, the spectral radius of the annihilation cycle tends to :

$$\boxed{\rho^{'}_2(\theta) = \frac{1}{1+2\varpi^2} = \frac{1-(1-\theta)^2}{1+(1-\theta)^2}}$$

and the cost-equivalent spectral radius per iteration is

$$\rho^{''}_2(\theta) = \sqrt{\rho^{'}_2(\theta)} = \sqrt{\frac{1-(1-\theta)^2}{1+(1-\theta)^2}}$$

(the unit of cost being one Jacobi-type iteration). We observe that it suffices that $0 < \theta < 1$ to ensure iterative convergence since then

$$0 < \rho_2'(\theta) < 1$$

In particular, one can let $\theta = \theta^*(\beta)$ so that :

$$\rho_2'(\theta^*) = \frac{1 - 4\beta + \sqrt{1+8\beta}}{3 + 4\beta - \sqrt{1+8\beta}}.$$

For $\beta = 1/3$, we get : $\rho_2'(\theta^*) \simeq 0.65$. In this example, 20 cycles (or 40 equivalent Jacobi-type iterations) are sufficient to reduce the iterative error in the solution of the linear system to 4 orders of magnitude. Thus, this algorithm provides us with a way to solve the linear system but its convergence rate is rather mediocre.

One way to improve it consists in using a larger number of annihilation parameters. For instance, to better annihilate the spectrum of A_θ which lies on a segment, one can form a cycle with 3 pseudo-timesteps. For this, the inverse of the midpoint of the segment can be added to τ and $\overline{\tau}$. This midpoint, $C = 1 - \theta$, is a real number and therefore, the additional substep is a simple over-relaxation. For the new cycle, the reduction factor associated with a given mode m is :

$$\begin{aligned} g_m'(r^*) &= (1 - \tau\lambda_m)(1 - \overline{\tau}\lambda_m)(1 - \lambda_m/C) \\ &= \frac{1 - 2\gamma_m^2}{1 + 2\varpi^2}\left(1 - \frac{1 - \theta + i\sqrt{\theta(2-\theta)}\gamma_m}{1-\theta}\right) \\ &= \frac{1 - 2\gamma_m^2}{1 + 2\varpi^2}\left(-i\frac{\sqrt{\theta(2-\theta)}}{1-\theta}\gamma_m\right) \end{aligned}$$

The maximum value of $|g_m'(r^*)|$ $(m = 1, N-1)$ is reached again for $m = 1$ and the spectral radius of the annihilation cycle becomes :

$$\boxed{\rho_3'(\theta) = \frac{1 - (1-\theta)^2}{1 + (1-\theta)^2}\frac{\sqrt{\theta(2-\theta)}}{1-\theta}}$$

while the cost-equivalent spectral radius per iteration is :

$$\rho_3''(\theta) = \left(\frac{1-(1-\theta)^2}{1+(1-\theta)^2} \frac{\sqrt{\theta(2-\theta)}}{1-\theta} \right)^{1/3} .$$

Note that $\rho_3'(\theta) < \rho_2'(\theta)$, $\forall \theta < 1$. For instance, for $\beta = 1/3$ and $\theta = \theta^*(1/3)$, we get $\rho_3'(\theta) \simeq 0.34$ while $\rho_2'(\theta) \simeq 0.65$ and $\rho_3''(\theta) \simeq 0.7$ while $\rho_2''(\theta) \simeq 0.8$. This reduction of the spectral radius allows us to compute only 8 cycles instead of 20 previously (or 24 equivalent Jacobi-type iterations instead of 40 previously).

Of course, even more complicated annihilation algorithms such as those proposed in [6] or [8] could be devised to improve the efficiency of the solution of the linear system.

4 Two-Dimensional Analysis

In order to investigate some of the effects of the spatial dimensionality, we consider in this section the following two-dimensional linear hyperbolic model problem :

$$\left\{ \begin{array}{l} u_t + c_x\, u_x + c_y\, u_y = 0 \quad (\, c_x > 0\, ,\; c_y > 0\,) \quad (\, (x,y) \in [0,1] \times [0,1]\, ;\; t > 0\,) \\ u(x,y,0) = u^0(x,y) \\ u(0,y,t) = u(x,0,t) = \text{const.} \end{array} \right. \tag{18}$$

The mesh is assumed to be uniform, the number of discretization points along the coordinates axes are denoted N_x and N_y, and one lets : $\nu_x = c_x/\Delta x$ and $\nu_y = c_y/\Delta y$. The stencils of the first-order, second-order central and upwind (here backward) operators write :

$$\delta_1 = \left(\begin{array}{ccc} & 0 & \\ -\nu_x & \nu_x + \nu_y & 0 \\ & -\nu_y & \end{array} \right) = \left(\begin{array}{ccc} & 0 & \\ -\nu_x & \nu_x + \nu_y & 0 \\ & -\nu_y & \end{array} \right)$$

$$\delta_2^C = \left(\begin{array}{ccc} & \nu_y & \\ -\nu_x & 0 & \nu_x \\ & -\nu_y & \end{array} \right)$$

$$\delta_2^U = \frac{1}{2}\left(\begin{array}{ccccc} & & 0 & & \\ & & 0 & & \\ \nu_x & -4\nu_x & 3(\nu_x+\nu_y) & 0 & 0 \\ & & -4\nu_y & & \\ & & \nu_y & & \end{array}\right)$$

Thus, the stencil of the preconditioner is :

$$\delta_1^\theta = (1-\theta)\delta_1 + \theta\delta_2^C = \left(\begin{array}{ccc} & \frac{\theta}{2}\nu_y & \\ \left(\frac{\theta}{2}-1\right)\nu_x & (\theta-1)(\nu_x+\nu_y) & \frac{\theta}{2}\nu_x \\ & \left(\frac{\theta}{2}-1\right)\nu_y & \end{array}\right)$$

whereas for the (partially-upwind) second-order difference operator :

$$\delta_2^\beta = (1-\beta)\delta_2^C + \beta\delta_2^U = \frac{1}{2}\left(\begin{array}{ccccc} & & 0 & & \\ & & (1-\beta)\nu_y & & \\ \beta\nu_x & -(1+3\beta)\nu_x & 3\beta(\nu_x+\nu_y) & (1-\beta)\nu_x & 0 \\ & & -(1+3\beta)\nu_y & & \\ & & \beta\nu_y & & \end{array}\right)$$

We have not been able to express the formal solution to the generalized eigenvalue problem in this two-dimensional case. Instead, a numerical study of the eigenvalues has been conducted whose results are presented here.

Several observations can be made on the 2D case compared with the 1D case. For all β and θ, we first observe that

$$\rho_{2D}(\beta,\theta) \geq \rho_{1D}(\beta,\theta)$$

This is because the 2D spectrum always contains the 1D spectrum.

The 2D spectrum of G corresponding to a 9×9 points discretization is plotted on Figs. 9, 10 and 11 for $\beta = 1/3$ and for several values of θ.

It is when $\nu_x = \nu_y$ that dimensionality has the greatest effect. In this case, for $N_x = N_y$, $\rho_{2D}(\beta,\theta) = \rho_{1D}(\beta,\theta)$ for certain values of β and θ, as illustrated on Fig. 12 for $N_x = N_y = 9$ and $\beta = 1/3, 1/2$ and $2/3$. For $\beta = 1/3$, $\rho_{2D} = \rho_{1D}$ for $0.3 \leq \theta \leq 0.8$ and for $\beta = 1/2$ or $\beta = 2/3$, $\rho_{2D} = \rho_{1D}$ for $\theta \leq 0.8$. But the more important consequence is that, in this case, the optimal θ and the spectral radius are the same as in 1D.

If now $\nu_x = \nu_y$ but $N_x \neq N_y$, the 1D and 2D optimal discrete θ and spectral radii differ, as shown in the following table for $\beta = 1/3$, $N_x = 9$ and different values of N_y :

N_y	9	19	29
θ^*	0.5456	0.5444	0.5442
ρ^*	0.298	0.304	0.305

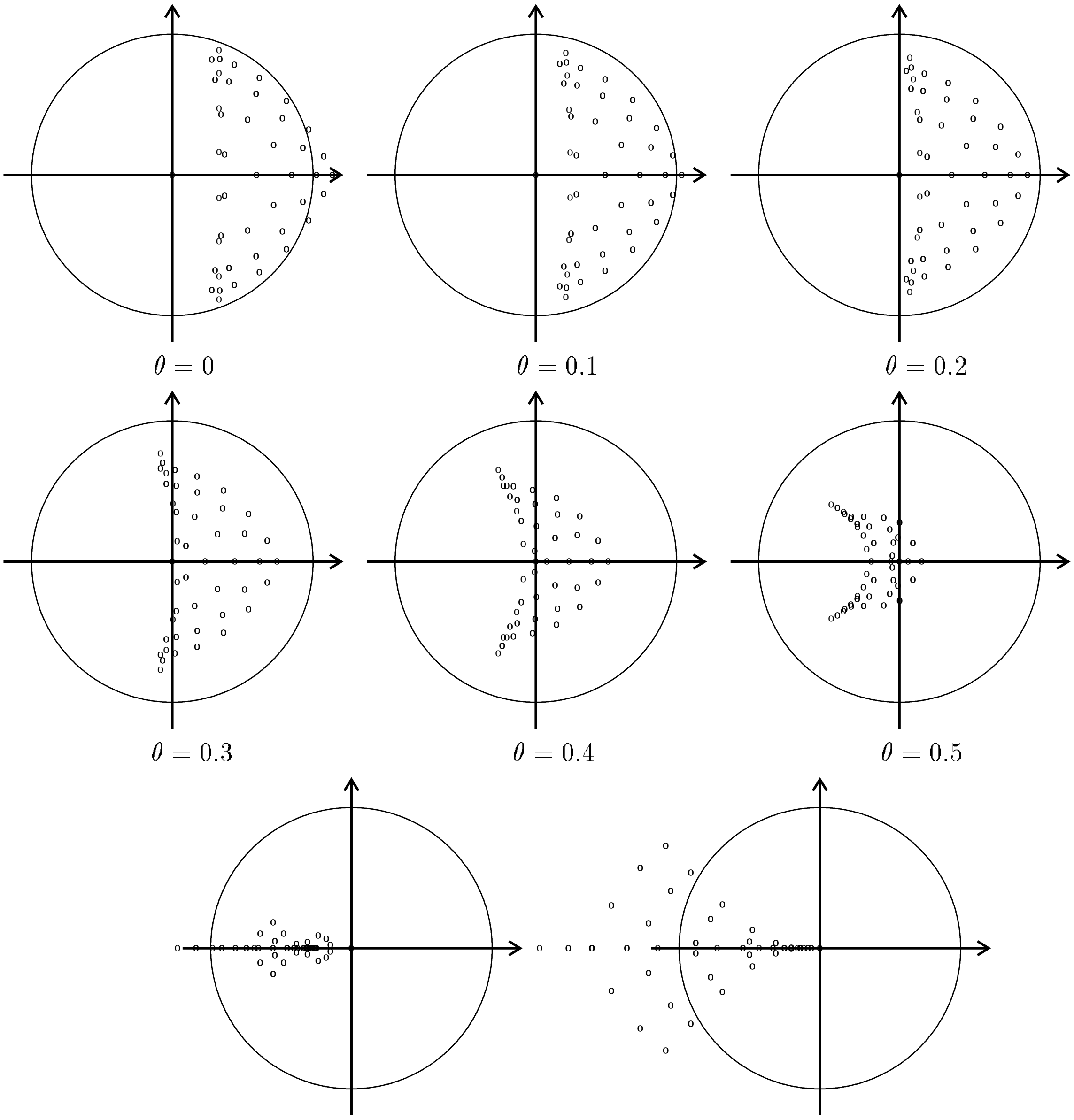
Table 4: 2D optimal θ and spectral radius versus N_y $(N_x = 9\,;\ \nu_x = \nu_y)$

The trend of variation of θ^* and ρ^* as N_y increases indicates that these parameters admit limits, but no calculations with larger values of N_y were made to confirm this. One can observe in the following table that for $N_x \neq N_y$, the optimal values of the parameters are always between those associated with the $N_x \times N_x$ and $N_y \times N_y$ discretizations :

$N_x \times N_y$	9x9	9x29	29x29
θ^*	0.5456	0.5442	0.5428
ρ^*	0.298	0.305	0.3123

Table 5: 2D optimal θ and spectral radius
for different domain discretizations $(\nu_x = \nu_y)$

Therefore the limit optimal values of the parameters (as $N_y \to \infty$), are also the 1D theoretical limits.

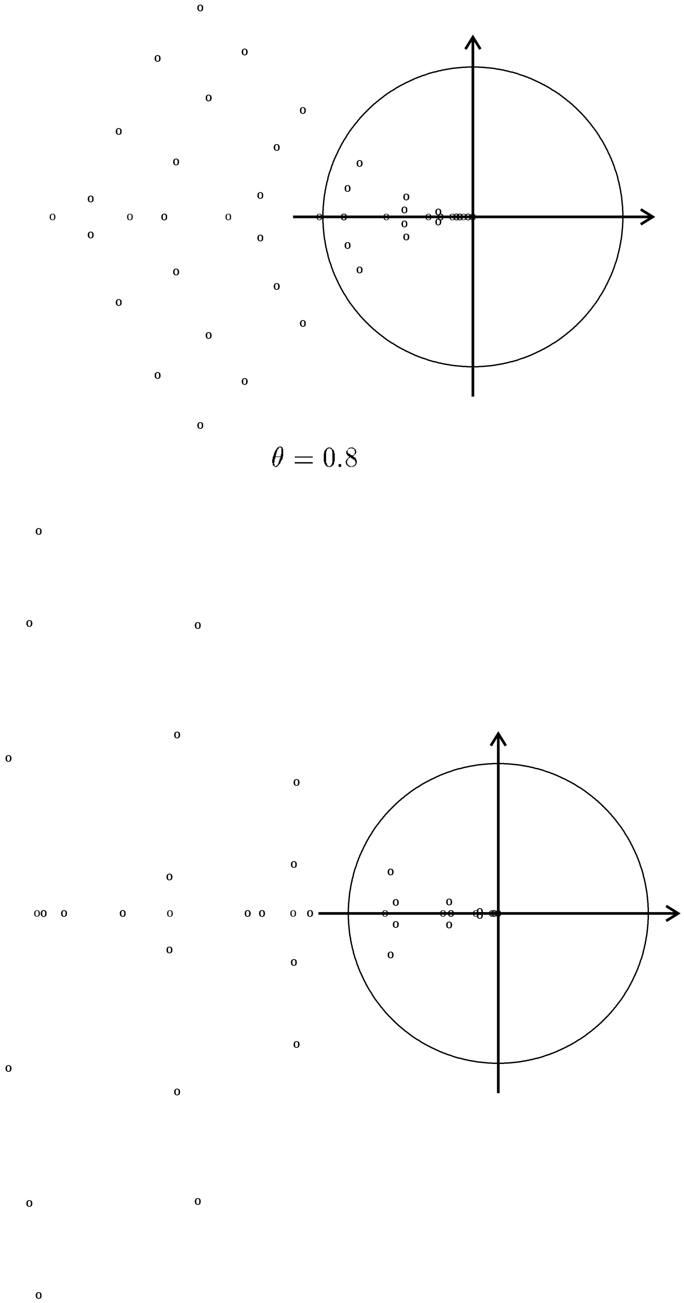


$\theta = 0$ $\theta = 0.1$ $\theta = 0.2$

$\theta = 0.3$ $\theta = 0.4$ $\theta = 0.5$

$\theta = 0.6$ $\theta = 0.7$

Figure 9: Spectrum of matrix G for $\beta = 1/3$ and $0 \leq \theta \leq \frac{7}{10}$



$\theta = 0.8$

$\theta = 0.9$

Figure 10: Spectrum of matrix G for $\beta = 1/3$ and $\theta = 8/10 \;\; 9/10$

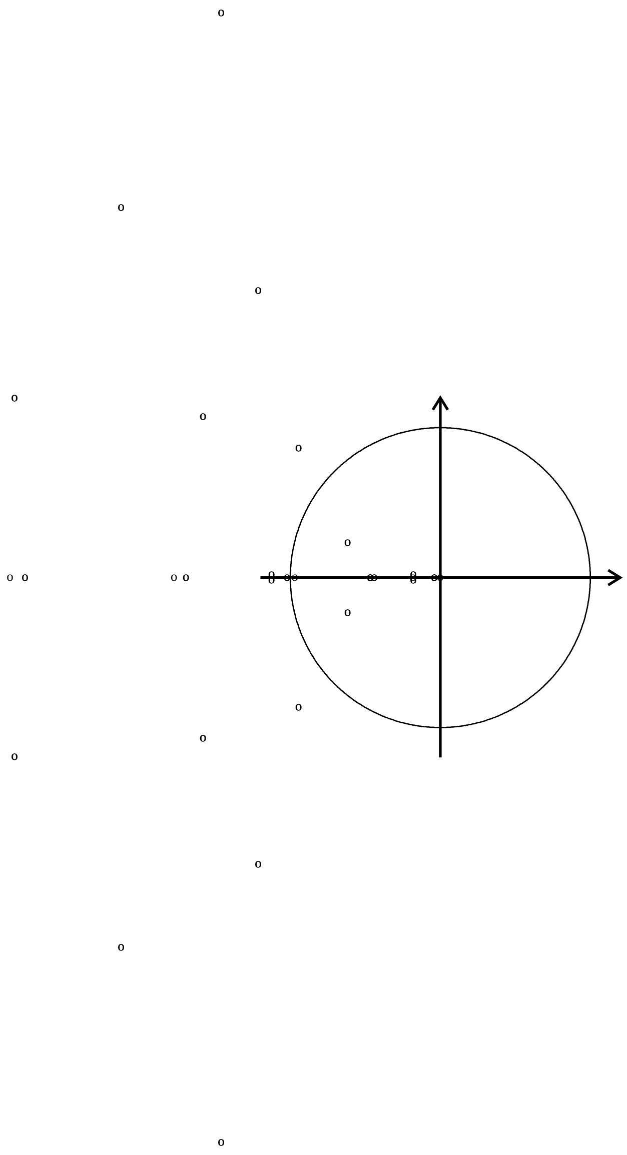


Figure 11: Spectrum of matrix G for $\beta = 1/3$ and $\theta = 1$

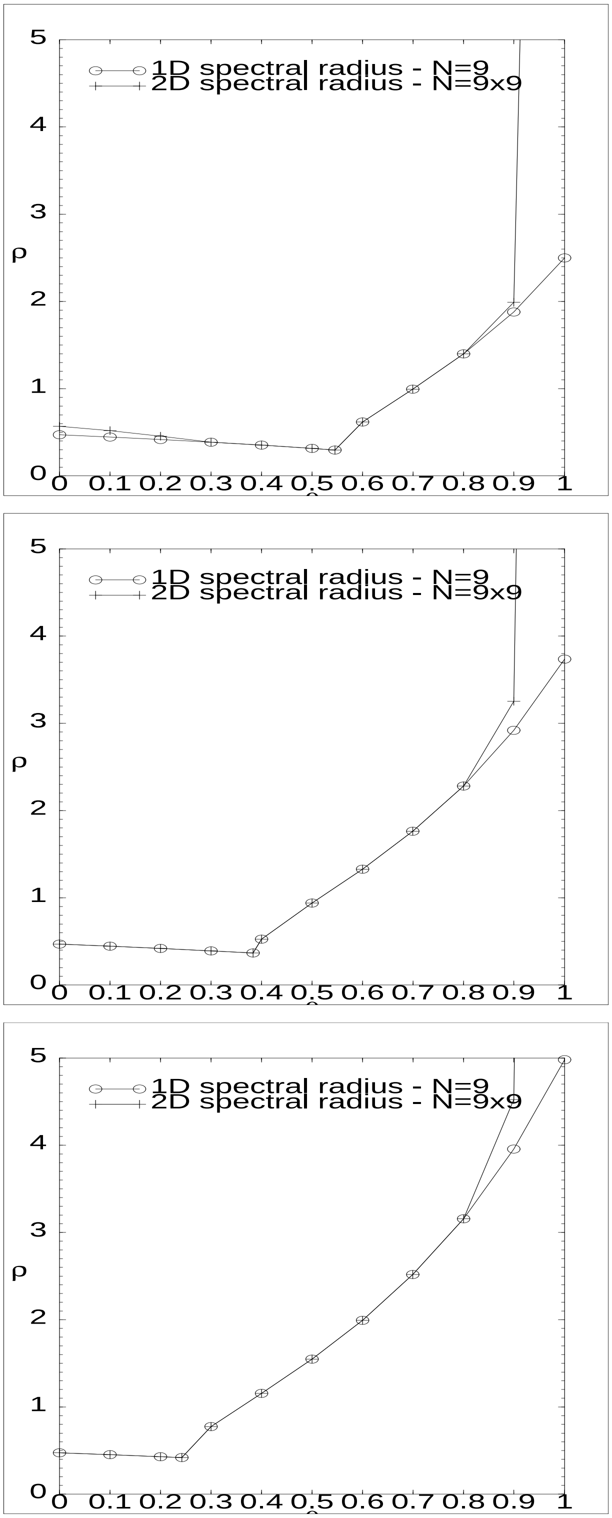


Figure 12: Comparison of the 1D and 2D spectral radii depending on θ; $\beta = 1/3, 1/2, 2/3$

In the case where $\nu_x \neq \nu_y$, θ^* and ρ^* are not the same as in 1D, but their limits (as $N_x \times N_y \to \infty$) are observed to be the same, as indicated by the following table made for $\nu_x = 100\nu_y$:

$N_x \times N_y$	9x9	29x9	9x29	29x29
θ^*	0.5444	0.5413	0.5442	0.5426
ρ^*	0.2983	0.3131	0.2985	0.3142

Table 6: 2D optimal θ and spectral radius
for different domain discretizations ($\nu_x = 100\nu_y$)

We observe from the following table that the variation of the discrete optimal values as the ratio ν_x/ν_y varies (assuming $N_x = N_y$) is not regular; however, the trend indicates that the quantities reach limits that are the same as those obtained with $\nu_x = \nu_y$ and also the same as in the 1D case for $N_x = N_y$:

ν_x/ν_y	1	100	1000	10000	100000
θ^*	0.5456	0.5444	0.5452	0.5455	0.5456
ρ^*	0.2978	0.2983	0.2981	0.2979	0.2978

Table 7: 2D optimal θ and spectral radius
for different mesh aspect ratios ν_x/ν_y ($N_x = N_y = 9$)

In all the cases we have described, the 2D limit optimal discrete values as the number of unknowns increases, are the same as the 1D discrete values.

The overall conclusion of this numerical identification of the spectrum of the amplification matrix G is that the optimal preconditioner identified in the one-dimensional case is optimal or close to optimal in the two-dimensional case also, so long as the number of degrees of freedom is large, or the mesh aspect ratios ν_x and ν_y very different in magnitude ($\nu_y/\nu_x \gg 1$ or $\ll 1$).

5 Conclusions

Preconditioners for a family of second-order (and one third-order) partially upwind approximations of hyperbolic equations, similar to those commonly employed in the discretization of the Euler equations have been proposed and their efficiency evaluated formally in the cases of pure-advection linear model problems. In addition, iterative algorithms have been proposed for the solution of the implicit system to be solved at a given (pseudo-timestep) iteration. Techniques of eigenmode annihilation have been shown to be potentially efficient to accelerate these algorithms also.

The results have been somewhat extended to the two-dimensional case by solving numerically the generalized eigenvalue problem. Basically, the preconditioners identified in the one-dimensional case remain potentially efficient in the case of two dimensions.

Several possible sequels of this study can be considered. The next step in our investigations will be to conduct experiments on nontrivial hyperbolic problems such as the solution of the Euler equations. Possibly more efficient algorithms could be constructed to solve the linear system. The one-dimensional analysis could also be extended to another type of preconditioners or to higher-order discretizations or to parabolic equations.

A Eigenvalues of the amplification matrix in the 1D case

The eigenvalues are calculated in the way introduced in [2] and [1].

A.1 Eigenvalues

The amplification operator G of the defect correction iteration is given by

$$G = I - A_\theta^{-1} B_\beta = I - \left[(1-\theta)\,\delta_1 + \theta\,\delta_2^C\right]^{-1} \left[(1-\beta)\,\delta_2^C + \beta\,\delta_2^U\right].$$

The eigenvalues λ and corresponding eigenfunctions u_λ of G satisfy the relation :

$$B_\beta\, u_\lambda = (1-\lambda) A_\theta\, u_\lambda$$

that is :

$$\left[(1-\beta)\,\delta_2^C + \beta\,\delta_2^U\right] u_\lambda = (1-\lambda)\left[(1-\theta)\,\delta_1 + \theta\,\delta_2^C\right] u_\lambda$$

Substituting

$$\left(u_\lambda\right)_j = \mu^j$$

in the above equation gives :

$$\begin{array}{l} \beta\mu^{j-2} + (-3\beta-1)\mu^{j-1} + 3\beta\mu^j + (1-\beta)\mu^{j+1} = \\ 2(1-\lambda)\left(\left(-1+\dfrac{\theta}{2}\right)\mu^{j-1} + (1-\theta)\mu^j + \dfrac{\theta}{2}\mu^{j+1}\right) \end{array}$$

or :

$$\begin{array}{c} (1-\beta-(1-\lambda)\theta)\,\mu^3 + (3\beta - 2(1-\lambda)(1-\theta))\,\mu^2 + \\ (-(3\beta+1) + (1-\lambda)(2-\theta))\,\mu + \beta = 0 \end{array}$$

The solutions of this equation are $\mu = 1$ (corresponding to $\lambda = 0$) and μ_1, μ_2 roots of the following 2nd-order equation for μ :

$$(1-\beta-(1-\lambda)\theta)\,\mu^2 + (2\beta + 2\lambda - 1 + (1-\lambda)\theta)\,\mu - \beta = 0 \qquad (19)$$

$\mu_1, \mu_2 \in \mathbb{C}$, $\mu_1 \neq \mu_2$ and $|\mu_1| = |\mu_2|$ give $\mu_2 = \mu_1 e^{2i\omega}, \omega \neq 0\ (\pi)$.
From (19), we can write :

$$\mu_1\mu_2 = \frac{-\beta}{1-\beta-(1-\lambda)\theta} = \mu_1^2 e^{2i\omega}$$

Denote $z \in \mathbb{C}$ such that :

$$\mu_1^2 = \frac{-\beta}{1-\beta-(1-\lambda)\theta} e^{-2i\omega} = z^2 e^{-2i\omega}$$

μ_1 and μ_2 write :

$$\mu_1 = z e^{-i\omega}, \quad \mu_2 = z e^{i\omega}$$

From (19), we can also write :

$$\mu_1 + \mu_2 = 2\ z\ \cos\omega = \frac{-2\beta - 2\lambda + 1 - (1-\lambda)\theta}{1-\beta-(1-\lambda)\theta}$$

which yields :

$$z = \frac{-2\beta - 2\lambda + 1 - (1-\lambda)\theta}{2(1-\beta-(1-\lambda)\theta)\cos\omega}$$

Thus, z verifies :

$$z^2 = \left[\frac{-2\beta - 2\lambda + 1 - (1-\lambda)\theta}{2(1-\beta-(1-\lambda)\theta)\cos\omega}\right]^2 = \frac{-\beta}{1-\beta-(1-\lambda)\theta}$$

This equation gives us the following condition on λ :

$$(-2\beta - 2\lambda + 1 - (1-\lambda)\theta)^2 = 4\beta\cos^2\omega(-1+\beta+(1-\lambda)\theta)$$

Let us note

$$\nu = 1 - \lambda$$

ν has to verify :

$$(-1 - 2\beta + \nu(2-\theta))^2 = -4\beta\cos^2\omega(1-\beta-\nu\theta)$$

in other terms :

$$(2-\theta)^2\nu^2 - \left(2(2-\theta)(1+2\beta) + 4\beta\theta\ \cos^2\omega\right)\nu + (1+2\beta)^2 + 4\beta(1-\beta)\cos^2\omega = 0$$

The discriminant of this equation is :

$$\begin{aligned} \Delta &= 16\cos^2\omega\left(-4\beta + 4\beta^2 + 6\beta\theta - 2\beta\theta^2 - \beta^2\theta^2\sin^2\omega\right) \\ &= 16\cos^2\omega \times \delta \end{aligned}$$

where

$$\boxed{\delta = -(\beta^2 \sin^2 \omega + 2\beta)\theta^2 + 6\beta\theta - 4\beta(1-\beta)}$$

β being given, the sign of δ has to be identified depending on θ. Define :

$$\varphi(\beta,\theta) = -(\beta^2 \sin^2 \omega + 2\beta)\theta^2 + 6\beta\theta - 4\beta(1-\beta)$$

We notice that

$$\begin{array}{l} \varphi(\beta,0) = -4\beta(1-\beta) < 0 \\ \varphi(\beta,1) = \beta^2(4-\sin^2\omega) > 0 \end{array}$$

Let us solve :

$$\varphi(\beta,\theta) = 0 \Longleftrightarrow (\beta^2 \sin^2 \omega + 2\beta)\theta^2 - 6\beta\theta + 4\beta(1-\beta) = 0$$

The discriminant is :

$$D = 4\beta^2 + (32 - 16\sin^2\omega)\beta^3 + 16\beta^4 \sin^2\omega > 0$$

The roots θ_1 and θ_2 are :

$$\theta_1 = \frac{3 - \sqrt{1 + 8\beta - 4\beta \ \sin^2\!\omega(1-\beta)}}{2 + \beta\sin^2\omega}, \quad \theta_2 = \frac{3 + \sqrt{1 + 8\beta - 4\beta \ \sin^2\!\omega(1-\beta)}}{2 + \beta\sin^2\omega}$$

Remark that :

$$0 \le \theta_1 \le 1, \quad \theta_2 > 1$$

Thus, if $0 \le \theta < \theta_1$, then $\delta < 0$ and

$$\nu = \frac{2(2-\theta)(1+2\beta) + 4\beta\theta\cos^2\omega \pm 4i\cos\omega\sqrt{-\delta}}{2(2-\theta)^2}$$

whereas, if $\theta_1 \le \theta < 1$, then $\delta \ge 0$ and

$$\nu = \frac{2(2-\theta)(1+2\beta) + 4\beta\theta \ \cos^2\!\omega \pm 4\cos\omega\sqrt{\delta}}{2(2-\theta)^2}$$

As a result, the eigenvalues $\{g_m\}$ of G are defined by :

$$\boxed{g_m = \frac{(2-\theta)(1-2\beta-\theta) - 2\beta\theta\cos^2\omega_m + 2i\cos\omega_m\sqrt{-\delta_m}}{(2-\theta)^2}}$$

if $0 \le \theta < \theta_m$, and

$$\boxed{g_m = \frac{(2-\theta)(1-2\beta-\theta) - 2\beta\theta\cos^2\omega_m + 2\cos\omega_m\sqrt{\delta_m}}{(2-\theta)^2}}$$

if $\theta_m \le \theta < 1$, where

$$\begin{aligned} \omega_m &= m\pi/N \\ \theta_m &= \frac{3 - \sqrt{1 + 8\beta - 4\beta\sin^2\omega_m(1-\beta)}}{2 + \beta\sin^2\omega_m} \\ \delta_m &= -(\beta^2\sin^2\omega_m + 2\beta)\theta^2 + 6\beta\theta - 4\beta(1-\beta) \end{aligned}$$

and $m = 1, .., N-1$.

A.2 Spectral radius

For a given θ, let K be the index such that :

$$\theta_1 = \theta_{N-1} < \theta_2 = \theta_{N-2} ... < \theta_K = \theta_{N-K} < \theta < \theta_{K+1} = \theta_{N-K-1} < ... < \theta_{N/2}$$

Then, for $m = 1, .., K$ and $m = N-K, ..., N-1$,

$$g_m = \frac{(2-\theta)(1-2\beta-\theta) - 2\beta\theta\cos^2\omega_m + 2\cos\omega_m\sqrt{\delta_m}}{(2-\theta)^2}$$

and

$$|g_m| = \frac{\left|(2-\theta)(1-2\beta-\theta) - 2\beta\theta\cos^2\omega_m + 2\cos\omega_m\sqrt{\delta_m}\right|}{(2-\theta)^2}$$

whereas, for $m = K+1, .., N-K-1$,

$$g_m = \frac{(2-\theta)(1-2\beta-\theta) - 2\beta\theta\cos^2\omega_m + 2i\cos\omega_m\sqrt{-\delta_m}}{(2-\theta)^2}$$

and instead

$$\begin{aligned}|g_m| &= \frac{\sqrt{((2-\theta)(1-2\beta-\theta)-2\beta\theta\cos^2\omega_m)^2+4\cos^2\omega_m(-\delta_m)}}{(2-\theta)^2}\\ &= \frac{\sqrt{(1-\theta)^2+4\beta\sin^2\omega_m(\beta+\theta-1)}}{2-\theta}\end{aligned}$$

Let us put :

$$\rho_1(\beta,\theta)=\max\{|g_m|\ ;\ m\in\{1,2,...,K\}\cup\{N-K,N-K+1,...,N-1\}\}$$

and

$$\rho_2(\beta,\theta)=\max\{|g_m|\ ;\ m\in\{K+1,K+2,...,N-K-1\}\}$$

To determine $\rho_1(\beta,\theta)$, let us denote ϕ the function :

$$\begin{aligned}\phi:\quad (-1,1) &\to I\!R\\ t &\mapsto (2-\theta)(1-2\beta-\theta)-2\beta\theta t^2\\ &\quad +2t\sqrt{-(\beta^2(1-t^2)+2\beta)\theta^2+6\beta\theta-4\beta(1-\beta)}\end{aligned}$$

$$\begin{aligned}\phi(-1) &= (2-\theta)(1-2\beta-\theta)-2\beta\theta-2\sqrt{-2\beta\theta^2+6\beta\theta-4\beta(1-\beta)}\\ &= (2-\theta)(1-\theta)-4\beta-2\sqrt{\beta\,(4\beta-2(2-\theta)(1-\theta))},\\ \phi(1) &= (2-\theta)(1-\theta)-4\beta+2\sqrt{\beta\,(4\beta-2(2-\theta)(1-\theta))}\end{aligned}$$

Remark that $4\beta-2(2-\theta)(1-\theta)>0$. Thus $(2-\theta)(1-\theta)-4\beta<0$ and $\phi(-1)<0$.

Now observe that :

$$\begin{aligned}&0<(2-\theta)^2(1-\theta)^2,\\ &16\beta^2-8\beta(2-\theta)(1-\theta)<16\beta^2-8\beta(2-\theta)(1-\theta)+(2-\theta)^2(1-\theta)^2,\\ &4\beta\,(4\beta-2(2-\theta)(1-\theta))<16\beta^2-8\beta(2-\theta)(1-\theta)+(2-\theta)^2(1-\theta)^2,\\ &2\sqrt{\beta\,(4\beta-2(2-\theta)(1-\theta))}<4\beta-(2-\theta)(1-\theta).\end{aligned}$$

Therefore, $\phi(1)$ is negative.

Denote $D(t) = -(\beta^2(1-t^2)+2\beta)\theta^2 + 6\beta\theta - 4\beta(1-\beta)$. The derivative of ϕ is expressed by

$$\phi^{'}(t) = -4\beta\theta t + 2\sqrt{D(t)} + \frac{2\beta^2\theta^2 t^2}{\sqrt{D(t)}}.$$

$\phi'(t)$ has the same sign as $-2\beta\theta t\sqrt{D(t)} + D(t) + \beta^2\theta^2 t^2 = (\sqrt{D(t)} - \beta\theta t)^2$. $\phi'(t)$ is therefore positive. Thus, the maximum of $|\phi(t)|$ is attained at $t=-1$. This implies that

$$\begin{aligned}
\rho_1(\beta,\theta) &= |g_{N-1}| \\
&= \frac{\left|(2-\theta)(1-2\beta-\theta) - 2\beta\theta\cos^2\omega_{N-1} + 2\cos\omega_{N-1}\sqrt{\delta_{N-1}}\right|}{(2-\theta)^2} \\
&= \frac{-(2-\theta)(1-2\beta-\theta) + 2\beta\theta\cos^2\omega_{N-1} - 2\cos\omega_{N-1}\sqrt{\delta_{N-1}}}{(2-\theta)^2}
\end{aligned}$$

On the other hand, the maximum of $\{|g_{K+1}|, ..., |g_{N-K-1}|\}$ is attained at $N/2$ if $\beta+\theta-1 \geq 0$ and at $K+1$ if $\beta+\theta-1 \leq 0$. Consequently :

$$\rho_2(\beta,\theta) = \begin{cases} \dfrac{\sqrt{(1-\theta)^2 + 4\beta\sin^2\omega_{N/2}(\beta+\theta-1)}}{2-\theta} & \text{if } \beta+\theta-1 \geq 0 \\ \dfrac{\sqrt{(1-\theta)^2 + 4\beta\sin^2\omega_{K+1}(\beta+\theta-1)}}{2-\theta} & \text{if } \beta+\theta-1 \leq 0 \end{cases}$$

A.3 Optimal θ

Observe that :

$$\rho_1(\beta,\theta) \simeq \frac{-(2-\theta)(1-\theta) + 4\beta + 2\sqrt{\beta\,(4\beta - 2(2-\theta)(1-\theta))}}{(2-\theta)^2}$$

and :

$$\rho_2(\beta,\theta) \leq \frac{1-\theta}{2-\theta}$$

if $\beta+\theta-1 \leq 0$, and

$$\rho_2(\beta,\theta) \simeq \frac{\sqrt{(1-\theta)^2 + 4\beta(\beta+\theta-1)}}{2-\theta}$$

if $\beta+\theta-1 \geq 0$.

Note that if $\beta+\theta-1 \leq 0$, $\rho_1(\beta,\theta)$ is defined if

$$\theta \geq \theta_1 = \frac{3-\sqrt{1+8\beta-4\beta\sin^2\omega_1(1-\beta)}}{2+\beta\sin^2\omega_1}$$

and, in this case,

$$\rho_2(\beta,\theta) \leq \rho_1(\beta,\theta)$$

Therefore, the minimum spectral radius is achieved by letting

$$\beta+\theta-1 \leq 0$$

and

$$\theta \leq \theta_1 = \frac{3-\sqrt{1+8\beta-4\beta\sin^2\omega_1(1-\beta)}}{2+\beta\sin^2\omega_1} \simeq \frac{3-\sqrt{1+8\beta}}{2}$$

We verify that $\theta_1 \leq 1-\beta$ which implies that :

$$\begin{array}{l} \beta \leq 1 \\ 4\beta^2 \leq 4\beta \\ 1+4\beta+4\beta^2 \leq 1+8\beta \\ 1+2\beta \leq \sqrt{1+8\beta} \\ \dfrac{3-\sqrt{1+8\beta}}{2} \leq 1-\beta \end{array}$$

Consequently,

$$\rho(\beta,\theta) = \rho_2(\beta,\theta) = \frac{\sqrt{(1-\theta)^2+4\beta\sin^2\omega_1(\beta+\theta-1)}}{2-\theta} \simeq \frac{1-\theta}{2-\theta}.$$

From now on, we assume $\rho(\beta,\theta) = \dfrac{1-\theta}{2-\theta}$ and $\theta_1 = \dfrac{3-\sqrt{1+8\beta}}{2}$. Then :

$$\frac{\partial\rho(\beta,\theta)}{\partial\theta} = \frac{-1}{(2-\theta)^2} < 0$$

Thus, the optimal spectral radius is obtained for

$$\boxed{\theta^*(\beta) = \frac{3-\sqrt{1+8\beta}}{2}}$$

and is equal to

$$\boxed{\rho^*(\beta) = \rho(\beta, \theta^*) = \frac{1-\theta^*}{2-\theta^*} = \frac{\sqrt{1+8\beta}-1}{\sqrt{1+8\beta}+1}.}$$

B Eigenvalues of the preconditioner in the 1D case

We first calculate the eigenvalues of A_θ :

$$A_\theta \, u_\lambda = \lambda \, u_\lambda \quad \Longleftrightarrow \quad \left[(1-\theta)\,\delta_1 + \theta\,\delta_2^C\right] u_\lambda = \lambda\, u_\lambda$$

Substituting again $\left(u_\lambda\right)_j = \mu^j$ in the above equation gives :

$$\left(-1+\frac{\theta}{2}\right)\mu^{j-1} + (1-\theta)\mu^j + \frac{\theta}{2}\mu^{j+1} = \lambda\,\mu^j$$

which simplifies to the following 2nd-order equation for μ :

$$\frac{\theta}{2}\mu^2 + (1-\theta-\lambda)\mu - 1 + \frac{\theta}{2} = 0 \tag{20}$$

$(\lambda,\mu) = (0,1)$ is one solution of this equation. For $\lambda \neq 0$, the simultaneous conditions

$$\mu_1, \mu_2 \in \mathbb{C}\,, \;\; \mu_1 \neq \mu_2\,, \;\; |\mu_1| = |\mu_2|$$

hold when

$$\mu_2 = \mu_1 e^{2i\omega} \quad (\omega \neq 0\;[\pi])$$

From (20), we can write :

$$\mu_1\mu_2 = \frac{\theta-2}{\theta} = \mu_1^2 e^{2i\omega}$$

$$\mu_1^2 = -\frac{2-\theta}{\theta}e^{-2i\omega}$$

$$\mu_1 = i\sqrt{\frac{2-\theta}{\theta}}e^{-i\omega}, \qquad \mu_2 = i\sqrt{\frac{2-\theta}{\theta}}e^{i\omega}$$

$$\mu_1+\mu_2 = 2\,i\sqrt{\frac{2-\theta}{\theta}}\cos\omega = \frac{2(-1+\theta+\lambda)}{\theta}$$

Finally, we obtain the following eigenvalues of the matrix A_θ :

$$\lambda = 1-\theta+i\,\sqrt{\theta(2-\theta)}\cos\omega$$

so that

$$|\lambda|^2 = 1 - \theta(2-\theta)\sin^2\omega$$

Consequently, the eigenvalues of the preconditioner A_θ^{-1} are :

$$\lambda^{-1} = \frac{1-\theta-i\sqrt{\theta(2-\theta)}\cos\omega}{1-\theta(2-\theta)\ \sin^2\!\omega}$$

These eigenvalues are situated on the circle of center

$$C = \frac{1}{2(1-\theta)}$$

and radius

$$R = \frac{1}{2(1-\theta)}$$

References

[1] J.A. DÉSIDÉRI, P.W. HEMKER, "Convergence analysis of the defect-correction iteration for hyperbolic problems", SIAM J. SCI. COMPUT, Vol. 16, No. 1, pp. 89-118, 1995.

[2] P.W. HEMKER, J.A. DÉSIDÉRI, "Convergence behaviour of defect correction for hyperbolic equations", J. OF COMP. AND APP. MATH. **45** (1993), 357-365.

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

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

[5] K. BOHMER, P. HEMKER, AND H. STETTER, "The defect correction approach", COMPUT. SUPPL., **5** (1984), pp. 1-32.

[6] J.A. DÉSIDÉRI, "La technique d'annihilation de modes propres et applications", *Rapport INRIA No. 1875, Mars 1993.*

[7] B. KOREN, "Multigrid and defect correction for the steady Navier-Stokes equations", CWI Tracts 74, Centrum voor Wiskunde en Informatica, Amsterdam, 1990.

[8] M.C. CICCOLI, A. DERVIEUX, J.A. DÉSIDÉRI, AND E. MORANO, "Efficient Solution Methods for Compressible Flow Computations", in *Lecture Notes in Pure and Applied Mathematics, Vol. 164, Finite Elements Methods, Fifty Years of the Courant Element, M. Krizek, P. Neittaanmäki, R. Stenberg, eds., Marcel Dekker, Inc., New York, Basel, Hong Kong (1994).*

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399