NASA Contractor Report 172157

ICASE

NASA-CR-172157
19830021840

PRECONDITIONED MINIMAL RESIDUAL METHODS FOR CHEBYSHEV SPECTRAL CALCULATIONS

Claudio Canuto

and

Alfio Quarteroni

Contract No. NAS1-17070
June 1983

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia 23665

Operated by the Universities Space Research Association

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23665

PRECONDITIONED MINIMAL RESIDUAL METHODS FOR CHEBYSHEV SPECTRAL CALCULATIONS

Claudio Canuto
Institute for Computer Applications in Science and Engineering
and
Istituto di Analisi Numerica del CNR, Pavia (Italy)

Alfio Quarteroni
Institute for Computer Applications in Science and Engineering
and
Istituto di Analisi Numerica del CNR, Pavia (Italy)

ABSTRACT

The problem of preconditioning the pseudospectral Chebyshev approximation of an elliptic operator is considered. The numerical sensitiveness to variations of the coefficients of the operator are investigated for two classes of preconditioning matrices: one arising from finite differences, the other from finite elements. The preconditioned system is solved by a conjugate gradient type method, and by a DuFort-Frankel method with dynamical parameters. The methods are compared on some test problems with the Richardson method [12] and with the minimal residual Richardson method [17].

Research was supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-17070 while the authors were in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665 and by the Istituto di Analisi Numerica del CNR, Pavia (Italy).

Introduction

Spectral approximations of elliptic boundary value problems lead to full and very ill-conditioned matrices. In the special case of constant coefficient operators, efficient direct methods have been proposed, [8], [9]. For nonconstant coefficient problems, considerable attention has been devoted after Orszag's paper [12] to the simultaneous use of iterative methods and preconditioning techniques.

In the present paper, we present and discuss the results of a number of numerical tests on the iterative solution of preconditioned systems arising from Chebyshev approximations. The first part is devoted to the analysis of the preconditioning of spectral matrices. The sensitiveness to variations of the coefficients, to leading and lower order terms is investigated. Besides the standard finite difference matrix proposed in [12], we consider a finite element matrix, which essentially retains the same preconditioning properties, being moreover symmetric. In both cases, an incomplete factorization of Wong's type [16] is used.

Two iterative methods are considered next. A preconditioned conjugate gradient method (which has been recently used in fluid dynamics and transonic flow calculations via finite elements, see [5] and the references therein) resulted to be rather slow on the tested problems, although it may be very robust in more complicated situations. The DuFort-Frankel method (first applied by Gottlieb et. al. [6], [7] to spectral calculations and here considered as a two-parameter preconditioned iterative method) yields good results when the optimal parameters are used. In order to overcome the difficulty of finding such parameters, we propose a modified version of the DuFort-Frankel method, devised according to a "minimal residual" strategy. The new method, compared with other iterative techniques in the literature, was the fastest in terms of speed of convergence.

No attempt is made in this report to give theoretical justifications to the methods, nor to consider nonlinear problems. Both the aspects are, however, under investigation.

Part of this work has been made while the authors were visiting ICASE. The numerical results reported here were obtained on the Honeywell 6040 at the the University of Pavia. Programs were written in double precision. The eigenvalues of Section 2 were obtained by EISPACK routines.

2. The Preconditioning of Spectral Matrices

Let L be a smooth second-order elliptic partial differential operator over the interval $\Omega^1 = (-1,1)$ or the square $\Omega^2 = (-1,1)^2$. We consider homogeneous Dirichlet boundary conditions for L, i.e., functions on which L acts will be assumed to vanish identically on the boundary. L_{sp} will denote the Chebyshev pseudospectral approximation of L of order N. This means that the approximate solution is a polynomial of degree N and derivatives are computed after interpolating the function by a polynomial of degree N at the Chebyshev nodes ($\{x_j = \cos \frac{\pi j}{N}\}$, $j=0,\cdots,N$ if $\Omega = \Omega^1$; $\{x_i,x_j\}$, $0 \leq i,j \leq N$ if $\Omega = \Omega^2$). We identify L_{sp} with a matrix which maps the set of values of a polynomial u at the interior Chebyshev nodes into the set of values of the spectral approximation of Lu at the same nodes.

It is known that L_{sp} has a full structure. Moreover, its condition number is $O(N^4)$. These are considered negative aspects of spectral Chebyshev approximations versus finite difference and finite element methods. However, a tremendous improvement in the computational efficiency of spectral methods comes from the observation that L_{sp} can be easily approximated by a sparse

matrix A, such that the condition number $\kappa(A^{-1} L_{sp})$ is close to 1 (see Orszag [12]). Throughout the paper, we refer to the ratio $\kappa = \kappa(M) = |\lambda_{max}|/|\lambda_{min}|$ as to the "condition number" of the matrix M, even when M is not symmetric.

In the following, A will denote any matrix having these properties, and it will be called a preconditioning matrix. A is assumed to be related to some discretization of the operator L, usually by finite differences or finite elements. Sometimes we shall relate A to some other elliptic operator $\mathscr{L}$, with the same principal part as L.

In one space dimension, the simplest way of building a preconditioning matrix is to use non-equally spaced finite differences at the Chebyshev nodes. The resulting matrix is tridiagonal, and it can be factorized in O(N) operations. If $Lu = -u_{xx}$, the corresponding preconditioning matrix is given by $A = \{a_{ij}\}$, where

$$(2.1) \qquad \begin{cases} a_{jj} = \dfrac{2}{h_j h_{j-1}} ; & a_{j,j-1} = \dfrac{-2}{h_{j-1}(h_j + h_{j-1})} ; \\ a_{j,j+1} = \dfrac{-2}{h_j(h_j + h_{j-1})} ; & h_j = x_j - x_{j+1} . \end{cases}$$

In Table 2.1, the operator $Lu = -u_{xx}$ is considered. The smallest and the largest eigenvalue λ_{min} and λ_{max}, and the ratio $\kappa = \lambda_{max}/\lambda_{min}$ are reported for both the matrices L_{sp} and $A^{-1} L_{sp}$.

Table 2.1. $Lu = -u_{xx}$

Au = finite differences at Chebyshev points for Lu.

N	L_{sp}			$A^{-1} L_{sp}$		
	λ_{min}	λ_{max}	κ	λ_{min}	λ_{max}	κ
4	2.46	.20 E2	.80 E1	1.	1.75	1.75
8	2.47	.21 E3	.87 E2	1.	2.13	2.13
16	2.47	.32 E4	.13 E4	1.	2.30	2.30
32	2.47	.50 E5	.20 E5	1.	2.38	2.38
64	2.47	.80 E6	.32 E6	1.	2.43	2.43
128	2.47	.13 E8	.52 E7	1.	2.45	2.45

As expected, the largest eigenvalue of L_{sp} grows like N^4, while the eigenvalues of the preconditioned matrix $A^{-1} L_{sp}$ lie in the interval [1., 2.5]. The spectrum of $A^{-1} L_{sp}$ exhibits a similar behavior even if the elliptic operator L contains lower order terms (see [12]).

The preconditioning properties of the matrix A seem to be rather insensitive to the lower order terms of L. The following table shows that the condition number $\kappa(A^{-1} L_{sp})$ is kept small when A is just the finite difference approximation of the second-order term of L, and the lower order terms are not prevailing. This implies that the preconditioning matrix can be kept fixed in solving nonlinear problems in which the lower order terms only change during the iterations. In all the cases considered below, the smallest eigenvalue λ_{min} is close to 1, and it converges to 1 from above as N increases.

Table 2.2. Condition number $\kappa(A^{-1} L_{sp})$
$Lu = -u_{xx} + \delta u_x + \gamma u$
Au = finite differences for $\mathscr{L}u = -u_{xx}$.

N	$\delta = 0.$ $\gamma = 1.$	$\delta = 0.$ $\gamma = 10.$	$\delta = 1.$ $\gamma = 0.$	$\delta = 10.$ $\gamma = 0.$	$\delta = 10.$ $\gamma = 10.$
4	1.38	2.25	1.45	2.29	1.61
8	1.87	3.45	1.90	2.01	1.96
16	2.16	4.30	2.15	2.28	2.41
32	2.32	4.73	2.31	2.47	2.82
64	2.40	4.92	2.38	2.85	3.25
128	2.44	4.99	2.43	3.08	3.54

When the magnitude of the lower order terms is exceedingly large, the condition number of $A^{-1} L_{sp}$ deteriorates. However the spectrum is still uniformly bounded in N.

Table 2.3. Condition number $\kappa(A^{-1} L_{sp})$
$Lu = -u_{xx} + \delta u_x + \gamma u$
Au = finite differences for $\mathscr{L}u = -u_{xx}$.

N	$\delta = 0.$ $\gamma = 100.$	$\delta = 100.$ $\gamma = 0.$	$\delta = 1000.$ $\gamma = 0.$
4	3.95	21.76	217.43
8	17.70	19.55	195.39
16	28.90	18.08	181.05
32	35.89	19.18	177.83
64	39.17	21.55	176.24
128	40.57	24.32	204.05

A family of variable coefficient operators $Lu = -(\alpha u_x)_x$, with $0 < \alpha_0 \leq \alpha(x) \leq \alpha_1$, is considered in the next table. The eigenvalues of the matrix $A^{-1} L_{sp}$ are bounded independently of N, although the bound is larger than for the constant coefficient operator. The condition number κ is close to the one in Table 2.1 when a moderate perturbation is applied, otherwise it grows slowly and linearly with the total variation of α.

Table 2.4. $Lu = -\left((1 + 10^{\nu}x^2)u_x\right)_x$
Au = finite differences for Lu.

N	$\nu = 0$		$\nu = 1$		$\nu = 2$	
	λ_{min}	κ	λ_{min}	κ	λ_{min}	κ
4	1.04	2.49	1.11	5.03	1.13	7.09
8	1.01	3.04	1.10	7.11	1.01	14.77
16	1.00	3.27	1.00	7.70	1.00	21.65
32	1.00	3.38	1.00	7.94	1.00	24.03
64	1.00	3.43	1.00	8.06	1.00	24.44
128	1.00	3.46	1.00	8.12	1.00	24.61

When the coefficient α depends itself on the solution u (as in the full potential equation) one would not change the preconditioning matrix at each iteration, in order to save factorization time. This situation is simulated to a certain extent in the next table. The effects of preconditioning the spectral matrix of a variable coefficient operator by a constant coefficient operator matrix are reported.

Table 2.5. $Lu = -\left((1 + 10^{\nu}x^2)u_x\right)_x$
Au = finite differences for $\mathscr{L}u = -u_{xx}$.

N	$\nu = 0$		$\nu = 1$	
	λ_{min}	κ	λ_{min}	κ
4	1.74	2.23	8.51	3.70
8	1.48	3.07	4.70	7.61
16	1.27	3.79	2.71	12.79
32	1.16	4.27	1.83	18.46
64	1.10	4.60	1.43	23.29

The spectrum of $A^{-1} L_{sp}$ is bounded independently of N. κ is comparable with the one of Table 2.4 when the perturbation is moderate, but it becomes noticeably worse when the distance between the preconditioning and the spectral operators increases. In this case, if the factorization is carried out in $O(N)$ operations only (in one dimension or with an incomplete factorization), the worsening of the condition number may not be balanced by the saving in factorization time (unless the computation of the entries of A is particularly expensive).

The matrices A we considered so far arose from a finite difference approximation of the operator $\mathscr{L}$ at the Chebyshev points. Even if $\mathscr{L}$ is formally self-adjoint, A is not symmetric, as well as L_{sp}. Actually A splits up as $A = D \cdot \tilde{A}$, with D diagonal and $\tilde{A}$ symmetric. Some iterative techniques require the symmetry of the preconditioning matrix (see, e.g., the next section). This can be accomplished by discretizing a suitable variational formulation of the elliptic operator via finite elements as follows.

If $\mathscr{L}u = -(\alpha u_x)_x$, the bilinear form associated to $\mathscr{L}$ is

$$a(u,v) = \int_{\Omega^1} \alpha\, u_x (v\omega)_x \, dx \tag{2.2}$$

where $\omega(x) = (1 - x^2)^{-1/2}$. The form $a(\cdot,\cdot)$ is continuous and coercive on the weighted Sobolev space $H^1_{0,\omega}(\Omega^1)$ (cf. [2]), but it is not symmetric. However the "reduced" form

$$\tilde{a}(u,v) = \int_{\Omega^1} \alpha\, u_x v_x\, \omega \, dx \tag{2.3}$$

is still coercive and continuous on $H^1_{0,\omega}(\Omega^1)$, and trivially symmetric. Assuming u and v continuous and piecewise linear between contiguous Chebyshev knots, we associate a matrix $A = \{a_{ij}\}$ to (2.3) by setting

$$a_{ij} = \tilde{a}(\phi_i, \phi_j) \tag{2.4}$$

where ϕ_k is continuous piecewise linear and $\phi_k(x_\ell) = \delta_{k\ell}$. For instance, if $\mathscr{L}u = -u_{xx}$, we have after dropping the common factor $\frac{\pi}{N}$:

$$a_{jj} = \frac{1}{h_j^2} + \frac{1}{h_{j-1}^2}\,; \qquad a_{j,j-1} = \frac{-1}{h_{j-1}^2}\,; \qquad a_{j,j+1} = \frac{-1}{h_j^2} \tag{2.5}$$

(compare with (2.1)). The spectrum of A behaves like the spectrum of the corresponding finite difference matrix, and the preconditioning properties are only slightly worse, as shown in the next table.

Table 2.6. $Lu = -u_{xx}$
Au = finite elements at the Chebyshev points for Lu (see(2.5)).

N	$\lambda_{min}(A^{-1} L_{sp})$	$\lambda_{max}(A^{-1} L_{sp})$	$\kappa(A^{-1} L_{sp})$
4	1.25	3.29	2.63
8	1.16	4.10	3.55
16	1.13	4.53	3.99
32	1.13	4.74	4.19
64	1.13	4.84	4.29
128	1.13	4.89	4.33

Up to now we considered one-dimensional problems. In 2D one can still use a finite difference or finite element matrix, say B, in the preconditioning. The corresponding results are similar to those in 1D. However, the exact "inversion" of such a matrix is more expensive, since the factors in its LU decomposition have a bandwidth of order $O(N)$ instead of $O(1)$. In order to overcome this drawback, different techniques of incomplete factorization have been successfully proposed (cf., e.g., [10], [11], [16]). The idea is to replace the exact factors L and U by some approximations $\tilde{L}$ and $\tilde{U}$ of them, which retain a very sparse structure. $\tilde{L}$ and $\tilde{U}$ are computed by incomplete steps of Gaussian elimination, under the condition that certain quantities depending on the product $\tilde{L}\,\tilde{U}$ agree with the corresponding quantities for B. The matrix $A = \tilde{L}\,\tilde{U}$ is then used in the preconditioning.

In our computations, the incomplete factorization was done according to Wong's row-sums agreement condition ([16]). Namely, let $b^{(0)}$ and $b^{(k)}$ denote the diagonal and the off-diagonals of a m-th order matrix B, i.e.,

$b^{(k)} = \{b_{i,i+k} | 1 \leq i, i+k \leq m\}$. If B is the finite difference matrix for a second-order operator at the Chebyshev points in the square, then only $b^{(0)}$, $b^{(\pm 1)}$, and $b^{(\pm N)}$ are not identically zero. The incomplete factors $\tilde{L}$ and $\tilde{U}$ of B have $\tilde{\ell}^{(0)}$, $\tilde{\ell}^{(-1)}$, $\tilde{\ell}^{(-N)}$ and $\tilde{u}^{(0)}$, $\tilde{u}^{(1)}$, $\tilde{u}^{(N)}$ respectively as nonzero (off)-diagonals. $\tilde{u}^{(0)}$ is chosen to be $\equiv 1.$, while the off-diagonal elements are easily determined by the condition that $a^{(\pm 1)} \equiv b^{(\pm 1)}$ and $a^{(\pm N)} \equiv b^{(\pm N)}$, where $A = \tilde{L}\,\tilde{U}$. Finally, $\tilde{\ell}^{(0)}$ is such that the sum of each row in A equals the corresponding sum in B.

Henceforth, we list some results about the preconditioning by an incomplete factorized finite difference matrix (for other results see [17]). Table 2.7 refers to constant coefficients operators. The different ratios between the coefficients of u_{xx} and u_{yy} are supposed to mimic the effect of the stretching of coordinates in a mapping process. The spectral matrix of a variable coefficients operator was preconditioned by the finite differences representation of the same operator (Table 2.8), or by that of a constant coefficient operator (Table 2.9).

Table 2.7. $Lu = -\alpha u_{xx} - u_{yy}$
Au = incomplete factorized finite difference matrix for LU.

N	$\alpha = 1.$		$\alpha = 10.$		$\alpha = 100.$	
	λ_{min}	κ	λ_{min}	κ	λ_{min}	κ
4	1.08	1.72	1.01	1.75	1.00	1.76
8	1.06	2.72	1.03	2.43	1.01	2.13
16	1.04	4.06	1.03	5.34	1.01	3.27

Table 2.8.

$Lu = (\alpha u_x)_x - (\beta u_y)_y$

$\alpha = \beta = 1 + 10x^2y^2$

Au = incomplete factorized finite difference matrix for Lu.

N	λ_{min}	κ
4	1.09	3.29
8	1.08	4.92
16	1.04	9.33

Table 2.9.

Lu as in Table 2.8

Au = incomplete factorized finite differences for $\mathcal{L}u = -u_{xx} - u_{yy}$.

N	λ_{min}	κ
4	1.48	6.89
8	1.44	10.94
16	1.23	19.90

Unlike the case of complete factorization, the condition number grows linearly with the number of unknowns. However, it ranges within moderate bounds (except when a different operator is used in the preconditioning). This gives evidence to the convenience of using incompletely factorized preconditioning matrices in spectral calculations. Better results can be achieved, with slightly more computational effort, by a higher order incomplete factorization in which $\tilde{L}$ and $\tilde{U}$ have one more nonzero off-diagonal (see [16] for more details).

3. A Conjugate Gradient Method

Even if the differential operator L is self-adjoint, the matrix arising from a Chebyshev spectral approximation is not symmetric. Thus, one can apply the standard conjugate gradient method (CG) to the normal equations of the preconditioned system; or one can use CG-type methods for nonsymmetric systems, like those proposed by Vinsome [13], Young and Jea [14], Axelsson

[1], or those by Concus and Golub [3], Widlund [15]. The methods of the first class may require the storage of back steps of the solution (see however, Wong [17] for an application of a truncated version of [1] to spectral calculations), while the methods of the second class need that the symmetric part of the system be easily solvable.

In the previous section it has been pointed out that the spectral matrix can be preconditioned using a symmetric positive definite matrix, connected with some finite element approximation of the elliptic operator. This suggests the use of the following preconditioned version of the CG method (see, e.g., [5]): Minimize

$$J(u) = r^T A^{-1} r \qquad r = L_{sp} u - f \tag{3.1}$$

by CG iterations in $\mathbb{R}^n$ equipped with the inner product

$$((u,v)) = u^T A v. \tag{3.2}$$

The corresponding algorithm is as follows.

$$\left\{\begin{array}{l} \text{Given } u^0 \in \mathbb{R}^n, \text{ compute } z^0 = A^{-1}(f - L_{sp} u^0) \\ \qquad\qquad g^0 = A^{-1} L_{sp}^T z^0, \quad w^0 = g^0. \\ \text{Then set for } k = 0,1,\cdots: \\ u^{k+1} = u^k + \alpha^k w^k, \qquad \text{where } \alpha^k = \dfrac{(z^k, L_{sp} w^k)}{(L_{sp} w^k, A^{-1} L_{sp} w^k)} \\ z^{k+1} = z^k - \alpha^k A^{-1} L_{sp} w^k \\ g^{k+1} = A^{-1} L_{sp}^T z^{k+1} \\ w^{k+1} = g^{k+1} + \gamma^{k+1} w^k, \qquad \text{where } \gamma^{k+1} = \dfrac{((g^{k+1} - g^k, g^{k+1}))}{((g^k, g^k))}. \end{array}\right. \tag{3.3}$$

Here $(u,v) = u^T v$ denotes the Euclidean product in $\mathbb{R}^n$. The product $L_{sp}^T z^{k+1}$ can be executed through Fast Fourier Transforms, and the entries of the matrix L_{sp} need not to be computed. Actually, assume that $L_{sp}u = -[P_c(\alpha u_x)]_x$ be the Chebyshev pseudospectral approximation of $Lu = -(\alpha u_x)_x$, where $P_c w$ is the N-th degree polynomial interpolating w at the nodes $x_j, j=0,\cdots,N$. We identify a N-th degree polynomial vanishing at $x = \pm 1$ with the vector of its values at $x_j, j=1,\cdots,N-1$. Recall that

$$(3.4) \qquad \int_{-1}^{1} u(x)\, v(x)\, \omega(x)\, dx = \frac{\pi}{N} \sum_{j=1}^{N-1} u(x_j)\, v(x_j) + \frac{2\pi}{N} \{u(-1)\, v(-1) + u(1)\, v(1)\}$$

for any u,v such that $uv \in \mathbb{P}_{2N-1}$. Then

$$(u, L_{sp}^T v)_{\mathbb{R}^{N-1}} = (L_{sp}u, v)_{\mathbb{R}^{N-1}} = -\frac{N}{\pi} \int_{-1}^{1} [P_c(\alpha u_x)]_x\, v\, \omega\, dx.$$

Integration by parts and several applications of (3.4) yield

$$(3.5) \qquad \begin{cases} L_{sp}^T v = -(P_c z)_x + \beta z \\ \text{where} \quad z = \alpha(v_x + \beta v), \qquad \beta(x) = \dfrac{\omega_x}{\omega}(x) = \dfrac{x}{1 - x^2}. \end{cases}$$

Similar expressions hold in two dimensions.

Algorithm (3.3) was used to compute the spectral solution for the test problems:

$$(3.6) \qquad \begin{cases} Lu \equiv -(\alpha u_x)_x = f, & -1 < x < 1 \\ \alpha \equiv 1. \quad \text{or} \quad \alpha(x) = 1 + 10x^2 \\ u(x) = \sin \pi x, \end{cases}$$

and

$$(3.7)\qquad \begin{cases} Lu \equiv -\left(\alpha u_x\right)_x - \left(\alpha u_y\right)_y = f & -1 < x,y < 1 \\ \alpha \equiv 1. \quad \text{or} \quad \alpha(x,y) = 1 + 10x^2y^2 & \\ u(x,y) = \sin \pi x \sin \pi y. & \end{cases}$$

In the following tables, we report the minimum number NIT of iterations required to get RES $< 10^{-8}$, where the relative residual is defined by

$$(3.8)\qquad RES^2 = \frac{(r,r)}{(f,f)}\,, \qquad r = f - L_{sp}\,u.$$

The initial guess was $u^0 \equiv 0$. ERR is the corresponding relative error on the solution

$$(3.9)\qquad ERR = \frac{\|u_{sp} - u_{exact}\|}{\|u_{exact}\|}\,,$$

where $\|u\| = (u,u)^{1/2}$ is the discrete ℓ^2-norm on the grid.

Table 3.1. CG Method for Problem (3.6)
Au = finite element matrix for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	1	.31 E-17	.18 E0	1	.90 E-17	.11 E-1
8	3	.12 E-16	.31 E-3	3	.19 E-13	.40 E-3
16	7	.53 E-14	.27 E-11	8	.13 E-15	.61 E-11
32	14	.58 E-8	.12 E-9	16	.60 E-12	.88 E-14
64	20	.52 E-8	.84 E-10	24	.24 E-8	.18 E-10
128	26	.52 E-8	.47 E-11	29	.35 E-8	.13 E-10

Table 3.2. CG Method for Problem (3.7)
Au = incomplete factorized finite element matrix for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2y^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	21	.31 E-8	.18 E0	32	.88 E-8	.10 E-1
8	44	.75 E-8	.31 E-3	72	.99 E-8	.15 E-3
16	80	.99 E-8	.89 E-9	70	.13 E-2	.56 E-4

It is seen that the number of iterations NIT to match the stopping criterion RES $< 10^{-8}$ increases sublinearly in 1D and linearly in 2D with the degree N of polynomials. This seems qualitatively in accordance with the behavior of the condition number of the matrix $A^{-1} L_{sp}$. However, we are unable to find a satisfactory explanation to the slow convergence properties of the method in 2D.

4. The DuFort-Frankel (DF) Method

The DuFort-Frankel method can be applied to the numerical solution of steady-state equations

$$Bu = g, \tag{4.1}$$

(the eigenvalues of B having positive real parts) as an iterative procedure depending on two parameters δ and γ:

$$\frac{u^{k+1} - u^{k-1}}{2\delta} = g - Bu^k - \gamma(u^{k+1} - 2u^k + u^{k-1}). \tag{4.2}$$

This can be written as a one-step method in the form

$$\begin{bmatrix} u^{k+1} \\ u^k \end{bmatrix} = G(B;\delta,\gamma) \begin{bmatrix} u^k \\ u^{k-1} \end{bmatrix} + \frac{1 + 2\delta\gamma}{2\delta} \begin{bmatrix} g \\ 0 \end{bmatrix}, \tag{4.3}$$

with proper definition of the matrix G.

The DF scheme was studied in connection with spectral methods by Gottlieb and Gustafsson [6] and by Funaro [4]. If B has real strictly positive eigenvalues (the largest and the smallest eigenvalues being denoted by λ_{max} and λ_{min} respectively), then it is seen that the method is convergent if

$$\gamma > \gamma_{LIM} = \frac{\lambda_{max}}{4}. \tag{4.4}$$

Moreover, Funaro [4] proves that the spectral radius $\rho(G)$ as a function of δ and γ has a curve of local minima (with respect to increments in the δ or in the γ direction separately) given by the branches of hyperbola

$$\gamma = \frac{1 + \delta^2\lambda_{max}^2}{4\delta^2\lambda_{max}} \qquad \text{if} \quad \gamma < \frac{\lambda_{min} + \lambda_{max}}{4}, \tag{4.5}$$

$$\gamma = \frac{1 + \delta^2\lambda_{min}^2}{4\delta^2\lambda_{min}} \qquad \text{if} \quad \gamma > \frac{\lambda_{min} + \lambda_{max}}{4}. \tag{4.6}$$

$\rho(G)$ attains its absolute minimum at the intersection of the two branches, i.e., at the "optimal parameters"

$$\delta^* = \frac{1}{\sqrt{\lambda_{min}\lambda_{max}}}, \qquad \gamma^* = \frac{\lambda_{min} + \lambda_{max}}{4}, \tag{4.7}$$

where

$$\rho(G)_{opt} = \rho^* = \frac{\sqrt{\lambda_{max}/\lambda_{min}} - 1}{\sqrt{\lambda_{max}/\lambda_{min}} + 1} . \tag{4.8}$$

The DF method with the optimal parameters (4.7) was applied to the solution of the test problems (3.6) - (3.7) by a preconditioned spectral method. Hence, we set in (4.2) $Bu = A^{-1} L_{sp} u$ and $g = A^{-1} f$, where A is the finite difference matrix associated to L, incompletely factorized in 2D. One DF iteration requires one multiplication $z = L_{sp} u^k$ and one forward-backward substitution $Aw = z$. The optimal parameters were computed using the exact values of λ_{min} and λ_{max} obtained in the previous section. The initial guess was $u^0 \equiv 0$, while u^1 was computed by a step of the Modified Euler method for the preconditioned system. NIT, RES, ERR are defined as in Section 3, formulae (3.8) - (3.9).

Table 4.1. DF Method with optimal parameters for Problem (3.6)
Au = finite differences for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	9	.33 E-8	.18 E0	19	.41 E-8	.11 E-1
8	12	.19 E-8	.13 E-3	24	.97 E-8	.40 E-3
16	12	.98 E-8	.39 E-8	26	.52 E-8	.56 E-8
32	14	.25 E-8	.14 E-8	29	.37 E-8	.79 E-8
64	14	.29 E-8	.29 E-8	28	.83 E-8	.75 E-8
128	14	.57 E-8	.56 E-8	28	.88 E-8	.86 E-8

Table 4.2. DF Method with optimal parameters for Problem (3.7)
Au = incomplete factorized finite differences for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2y^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	9	.70 E-8	.18 E0	15	.20 E-8	.10 E-1
8	14	.97 E-8	.13 E-3	20	.84 E-8	.15 E-3
16	20	.56 E-8	.11 E-8	30	.98 E-8	.64 E-8
32	59	.86 E-8	.48 E-8	49	.82 E-8	.92 E-9

It is seen that the number of iterations needed to satisfy the stopping test is bounded as a function of N in the 1D tests, while it is linearly growing in 2D. This corresponds to the behavior of the condition number of the matrix $A^{-1} L_{sp}$, as reported in Tables 2.4 and 2.8.

Moreover, NIT is comparable with the one relative to the CG method in 1D, and definitely smaller in 2D. Since one DF iteration is faster than one CG iteration (by a factor of 1.7 both in 1D and in 2D) we conclude that the DF method with optimal parameters exhibits a globally better behavior than the CG method on the tested problems.

The speedup in the convergence due to the use of a preconditioning technique is particularly impressive for the DF method. This is suggested by formula (4.8), which shows the dependence of the optimal spectral radius on the condition number of B. Table 4.3 reports the performance of the DF method with optimal parameters without preconditioning (i.e., $Bu = L_{sp} u$) for problem (3.6) with $\alpha \equiv 1$. (compare with Table 4.1)

Table 4.3. DuFort-Frankel method without preconditioning

N	4	8	16	32	64
NIT	23	84	327	>400	>>400

The practical interest of formulae (4.7) relies on the explicit knowledge of λ_{min} and λ_{max}, which is rarely the case. Approximate values of δ^* and γ^* may be obtained in different ways, for instance by estimates on the eigenvalues of B or by extrapolation of correct values of δ^* and γ^* computed on coarser grids. It was found that linear extrapolation on the parameters as functions of N may lead to negative values of γ^*. Instead, linear extrapolation on the ratios of contiguous values of the parameters gives accurate answers. Actually, the case $N = 32$ in Table 4.2 was run with extrapolated "optimal" parameters.

Unfortunately, the method appears to be rather sensitive to the choice of parameters, especially around the curve of optimality. The qualitative behavior of $\rho(G)$ as a function of γ for fixed δ (or conversely) is similar to the one encountered in a SOR method. Table 4.4 shows the values of NIT for problem (3.6) with $\alpha = 1 + 10x^2y^2$, $N = 32$ and finite differences preconditioning, as a function of γ and δ.

Table 4.4. NIT as function of γ and δ.

$x\delta^*$	γ_{LIM}/γ^*	1	2	4	8
4	>>400	193	438	>>400	>>400
2	>400	98	162	416	>>400
1	>400	49	97	183	392
1/2	>400	87	61	134	196
1/4	>>400	147	145	86	151

$x\gamma^*$

The previous considerations suggest that the DF method, although in principle very powerful, may be poorly efficient in applications if the user attempts to fix the constants δ and γ once for all.

However, it is possible to transform the DF method into a completely parameter-free iterative scheme following a "minimal residual" strategy which has been proven successful in connection with other iterative schemes. The parameters γ and δ are computed at each iteration in order to minimize the ℓ^2-norm of the residual $r = f - L_{sp}\, u$. Given u^k, r^k and u^{k-1}, r^{k-1} then u^{k+1} and r^{k+1} are defined according to (4.2)

$$(4.9) \qquad \begin{cases} u^{k+1} = c_1\, A^{-1}\, r^k + c_2\, u^k + c_3\, u^{k-1} \\ r^{k+1} = -c_1\, L_{sp}\, A^{-1}\, r^k + c_2\, r^k + c_3\, r^{k-1} \end{cases},$$

where

$$(4.10) \qquad c_1 = \frac{2\delta}{1 + 2\delta\gamma}, \qquad c_2 = \frac{4\delta\gamma}{1 + 2\delta\gamma}, \qquad c_3 = \frac{1 - 2\delta\gamma}{1 + 2\delta\gamma}.$$

(r^{k+1},r^{k+1}) is minimized if one sets in (4.10)

$$(4.11)\qquad \begin{cases} \delta = \delta^k = \dfrac{1}{2}\,\dfrac{(q,r^{k-1} - \alpha s)}{(q,q - \beta s)} \\ \gamma = \gamma^k = \dfrac{1}{2\delta^k}\,\dfrac{(p,2\delta^k q - r^{k-1})}{(p,s)} \end{cases},$$

where $p = r^k - r^{k-1}$, $q = L_{sp}\, A^{-1}\, r^k$, $s = r^k + p$, $\alpha = (p,r^{k-1})/(p,s)$ and $\beta = (p,q)/(p,s)$.

This algorithm can be called "Minimal Residual DuFort-Frankel" (MRDF) Method. One MRDF iteration requires one forward-backward substitution $Az = r^k$ and one multiplication $w = L_{sp}\, z$; moreover, r^{k-1} needs to be stored with u^{k-1}. Note that if $u^1 = u^0$ but $r^0 \neq 0$ the algorithm cannot converge. Hence u^1 should be chosen in such a way that $u^1 - u^0$ and r^0 be roughly comparable. For instance u^1 can be computed from u^0 by one step of the Minimal Residual Richardson method (see Section 5 (b)).

Tables 4.5 and 4.6 are analogous of Tables 4.1 and 4.2, except that the MRDF method was used instead of the DF method with optimal parameters.
In 1D, the gain in the speed of convergence over the DF method with optimal parameters is spectacular, although this may depend on particular circumstances. In 2D, the improvement of the performances is less impressive.

However, one must not forget that the main improvement of the MRDF over the DF method relies on the complete automatization in the choice of parameters.

Table 4.5. MRDF Method for Problem (3.6)
Au = finite differences for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	1	.11 E-17	.18 E0	1	.70 E-17	.11 E-1
8	5	.48 E-9	.13 E-3	8	.55 E-8	.40 E-3
16	7	.71 E-8	.38 E-9	11	.46 E-8	.35 E-8
32	4	.56 E-9	.13 E-9	9	.48 E-8	.22 E-8
64	3	.10 E-9	.20 E-10	3	.45 E-8	.18 E-8
128	2	.33 E-9	.68 E-10	2	.10 E-8	.56 E-9

Table 4.6. MRDF Method for Problem (3.7)
Au = incomplete factorized finite difference matrix for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2y^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	7	.23 E-8	.18 E0	10	.84 E-8	.10 E-1
8	13	.47 E-8	.13 E-3	20	.77 E-8	.15 E-3
16	19	.63 E-8	.80 E-9	29	.90 E-8	.79 E-8
32	36	.50 E-8	.96 E-9	46	.78 E-8	.86 E-9

5. Comparisons with Other Methods

The preconditioned CG and DF methods were compared with two other iterative techniques recently suggested for spectral calculations: the Richardson iteration proposed by Orszag [12], and the Minimal Residual

Richardson method proposed by Wong [17]. We briefly review these techniques and we report for the sake of completeness their behavior on the test problems used throughout this paper.

(a) Richardson Method ([12]):

Given u^0, compute u^{k+1} from u^k by solving

$$Au^{k+1} = Au^k - \alpha(L_{sp}\, u^k - f), \tag{5.1}$$

where $0 < \alpha < 2/\lambda_{max}$, λ_{min} and λ_{max} being the smallest and the largest eigenvalue of $A^{-1} L_{sp}$. The optimal value of α,

$$\alpha_{opt} = \frac{2}{\lambda_{min} + \lambda_{max}}, \tag{5.2}$$

was computed exactly and used in the following tests. One iteration requires one multiplication $z = L_{sp}\, w$ and one forward-backward substitution $Ax = b$.

Table 5.1. Richardson Method for Problem (3.6)
Au = finite differences for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	8	.90 E-8	.18 E0	33	.63 E-8	.11 E-1
8	17	.81 E-8	.13 E-3	62	.95 E-8	.40 E-3
16	20	.77 E-8	.54 E-8	71	.82 E-8	.68 E-8
32	21	.70 E-8	.64 E-8	73	.96 E-8	.88 E-8
64	22	.70 E-8	.41 E-8	74	.97 E-8	.94 E-8
128	22	.42 E-8	.51 E-8	75	.87 E-8	.86 E-8

Table 5.2. Richardson Method for Problem (3.7)
Au = incomplete factorized finite difference matrix for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2y^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	12	.87 E-8	.18 E0	23	.84 E-8	.10 E-1
8	24	.71 E-8	.13 E-3	45	.69 E-8	.15 E-3
16	39	.92 E-8	.92 E-8	90	.99 E-8	.21 E-8

(b) Minimal Residual Richardson (MRR) Method ([17]):

In the previous scheme, compute $\alpha = \alpha^k$ at each iteration in order to minimize the residual (r^{k+1}, r^{k+1}). Hence one gets:

Given u^0, compute $r^0 = f - L_{sp}\, u^0$, $z^0 = A^{-1}\, r^0$,

then set

$$(5.3)\qquad \begin{cases} u^{k+1} = u^k + \alpha^k\, z^k \qquad \text{where} \quad \alpha^k = \dfrac{(r^k, L_{sp}\, z^k)}{(L_{sp}\, z^k, L_{sp}\, z^k)}\,, \\ r^{k+1} = r^k - \alpha^k\, L_{sp}\, z^k \\ z^{k+1} = A^{-1}\, r^{k+1}\,. \end{cases}$$

The computational effort per iteration is comparable with that of the Richardson method. Note that this method is obtained from the previous one by the same strategy used in deriving the MRDF from the pure DF method.

Table 5.3. MRR Method for Problem (3.6)
Au = finite differences for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	1	.12 E-17	.18 E0	1	.15 E-17	.11 E-1
8	10	.32 E-8	.13 E-3	13	.22 E-10	.40 E-3
16	8	.78 E-8	.29 E-9	13	.76 E-8	.28 E-8
32	5	.56 E-9	.19 E-11	10	.62 E-8	.37 E-8
64	4	.14 E-9	.12 E-10	4	.58 E-8	.15 E-8
128	3	.34 E-9	.48 E-10	3	.16 E-8	.12 E-8

Table 5.4. MRR Method for Problem (3.7)
Au = incomplete factorized finite difference matrix for Lu.

N	$\alpha \equiv 1.$			$\alpha = 1 + 10x^2y^2$		
	NIT	RES	ERR	NIT	RES	ERR
4	9	.98 E-8	.18 E0	18	.96 E-8	.10 E-1
8	18	.11 E-8	.13 E-3	22	.29 E-8	.15 E-3
16	23	.90 E-8	.53 E-8	32	.14 E-8	.60 E-9
32	58	.88 E-8	.19 E-8	58	.89 E-8	.43 E-9

(c) Comparisons

The speed of convergence of the methods previously discussed was compared on the basis of the number of iterations and the CPU time. Two significant cases were considered.

CASE 1: Problem (3.6) with $\alpha = 1 + 10x^2$
N = 128, i.e., 127 grid points in the interval (-1,1).

CASE 2: Problem (3.7) with $\alpha = 1 + 10x^2y^2$
N = 32, i.e., 31 × 31 grid points in the square $(-1,1)^2$.

Define for the sake of simplicity the following labels:

A Richardson method (5.1) with optimal parameter (5.2)
B Minimal residual method (5.3)
C Conjugate gradient method (3.3)
D DuFort-Frankel method (4.2) with optimal parameters (4.7)
E Minimal residual DuFort-Frankel method (4.9)

We used the standard finite difference (finite element for method C) preconditioning matrix on the spectral grid, incompletely factorized in Case 2 according to Wong's method described in Section 2. The optimal parameters were computed with the exact values of λ_{min} and λ_{max}. $u^0 \equiv 0$ was the initial guess.

The results in Figure 5.1 and in Figure 5.2 are in a sense machine- and programmer-independent. The relative performances of the methods can be analyzed according to the global CPU-time consumption, using the following table.

Table 5.5. CPU-time per iteration in hours

METHOD	A	B	C	D	E
CASE 1	.272 E-3	.280 E-3	.467 E-3	.273 E-3	.285 E-3
CASE 2	.128 E-2	.128 E-2	.217 E-2	.127 E-2	.130 E-2

Hence, Figure 5.1 and Figure 5.2 also summarize the relative performances of the methods in terms of cost, except for method C which is roughly 1.7 times slower than the others.

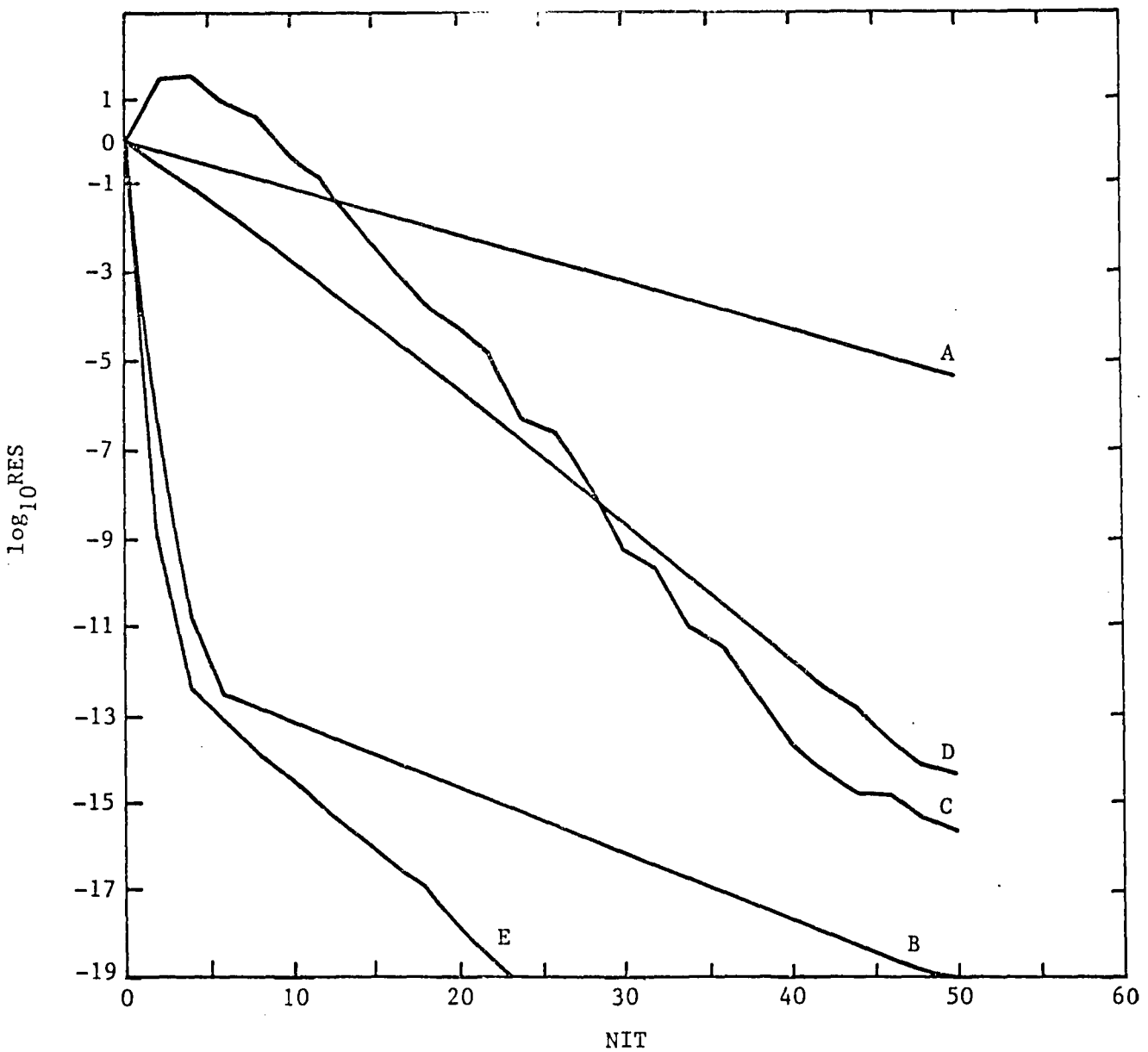

Figure 5.1. Case 1. Convergence histories versus number of iterations.

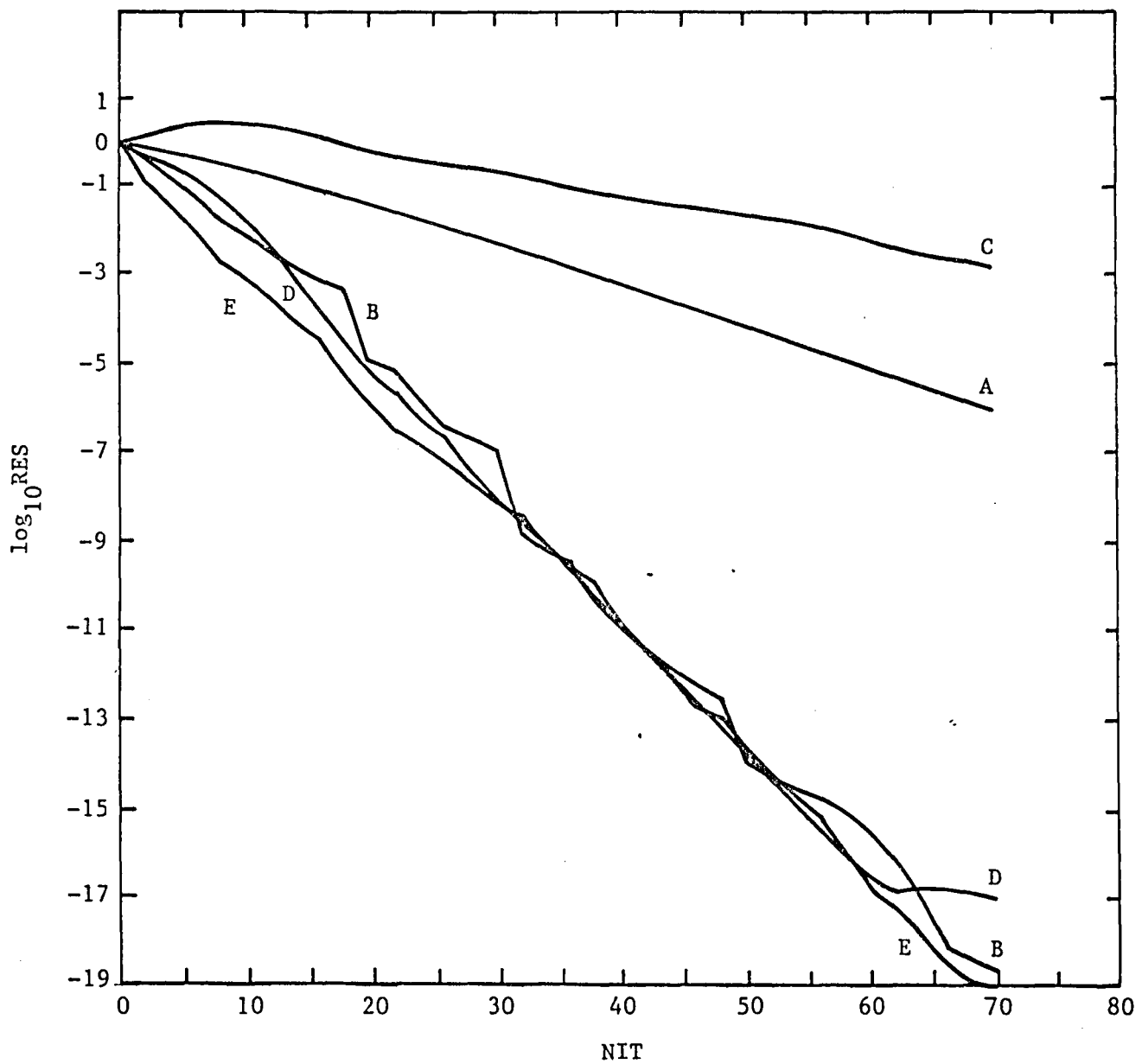

Figure 5.2. Case 2. Convergence histories versus number of iterations.

Comments

Globally, the results confirm the utility of preconditioning techniques in spectral calculations: few iterations are needed to reach the spectral accuracy, which corresponds in the test problems to a relative residual of 10^{-18}.

Methods A and D behave exactly like predicted by the theory: the error is reduced at each iteration by a factor $\frac{\kappa-1}{\kappa+1}$ for method A, and $\frac{\sqrt{\kappa-1}}{\sqrt{\kappa+1}}$ for method B (κ is the condition number of the preconditioned matrix).

The conjugate gradient method gives contradictory answers in terms of speed of convergence: in 1D the factor of reduction of the error is smaller than that for method D, while in 2D it is comparable with that of method A. In both cases, the method turns out to be not competitive in terms of computer time.

The "minimal residual" strategy is always winning over the "optimal parameter" strategy, also where the exact optimal parameters can be used. In particular, the MRR method is superior even to the Richardson method with Chebyshev acceleration, proposed in [12] (according to [12], p. 86, the Chebyshev acceleration increases the speed of Richardson method by a factor of 2, although it requires the extra-storage of the vector u^{k-1}).

The MRDF method requires the storage of u^{k-1} and r^{k-1}, being a two-step method. However, the extra memory required results in a better accuracy, and the MRDF method appears in all cases the fastest method among those tested in this report.

References

[1] O. AXELSSON, "On Preconditioning and Convergence Acceleration in Sparse Matrix Problems," Report CERN 74-10, Geneva, Switzerland, 1974.

[2] C. CANUTO and A. QUARTERONI, in Proc. Workshop on Spectral Methods for Partial Differential Equations, (R. Voigt, ed.), CBMS Regional Conference Series in Applied Mathematics, SIAM 1983 (to appear).

[3] P. CONCUS and G. H. GOLUB, in Lecture Notes in Economics and Mathematical Systems, Vol. 134 (R. Glowinski and J. L. Lions, eds.), Springer-Verlag, Berlin-Heidelberg-New York, 1976, 56-65.

[4] D. FUNARO, "Analysis of the DuFort-Frankel method for differential systems," to appear in SIAM J. Numer. Anal.

[5] R. GLOWINSKI, B. MANTEL, J. PERIAUX, O. PIRONNEAU, and G. POIRIER, in Computing Methods in Applied Sciences and Engineering, (R. Glowinski and J. Lions, eds.), 445-487, North Holland, Amsterdam-NewYork, 1980.

[6] D. GOTTLIEB and B. GUSTAFSSON, SIAM J. Numer. Anal. 13 (1976), 129-144.

[7] D. GOTTLIEB and L. LUSTMAN, "The DuFort-Frankel Chebyshev method for parabolic initial boundary value problems," ICASE Report No. 81-42, 1981.

(8) D. GOTTLIEB and S. A. ORSZAG, "Numerical Analysis of Spectral Methods: Theory and Applications," CBMS Regional Conference Series in Applied Mathematics, SIAM, 1977.

[9] D. B. HAIDVOGEL and T. A. ZANG, J. Comput. Phys., 30 (1979), 167-180.

[10] J. A. MEIJERINK and H. A. van der VORST, Math. Comp. 31 (1977), 148-162.

[11] J. A. MEIJERINK and H. A. van der VORST, J. Comput. Phys. 44 (1981), 134-155.

[12] S. A. ORSZAG, J. Comput. Phys. 37 (1980), 70-92.

[13] P. K. W. VINSOME, in Proc. of the Fourth Symposium on Reservoir Simulation, 1976, Soc. of Petroleum Engineers of AIME, 149-159.

[14] D. M. YOUNG and K. C. JEA, Linear Algebra Appl. 34 (1980), 159-194.

[15] O. WIDLUND, SIAM J. Numer. Anal. 15 (1978), 801-812.

[16] Y. S. WONG, in Numerical Methods in Thermal Problems (R. W.Lewis and K. Morgan, eds.), Pineridge Press, Swansea, 1979.

[17] T. ZANG, Y. S. WONG, and M. Y. HUSSAINI, J. Comput. Phys. 48 (1982), 485-501.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-172157		

4. Title and Subtitle	5. Report Date
Preconditioned Minimal Residual Methods for Chebyshev Equations	June 1983
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Claudio Canuto and Alfio Quarteroni	83-28
	10. Work Unit No.
9. Performing Organization Name and Address	
Institute for Computer Applications in Science and Engineering Mail Stop 132C, NASA Langley Research Center Hampton, VA 23665	11. Contract or Grant No. NAS1-17070
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Contractor Report
National Aeronautics and Space Administration Washington, DC 20546	14. Sponsoring Agency Code

15. Supplementary Notes

Langley Technical Monitor: Robert H. Tolson
Final Report

16. Abstract

The problem of preconditioning the pseudospectral Chebyshev approximation of an elliptic operator is considered. The numerical sensitiveness to variations of the coefficients of the operator are investigated for two classes of preconditioning matrices: one arising from finite differences, the other from finite elements. The preconditioned system is solved by a conjugate gradient type method, and by a DuFort-Frankel method with dynamical parameters. The methods are compared on some test problems with the Richardson method [12] and with the minimal residual Richardson method [17].

17. Key Words (Suggested by Author(s))	18. Distribution Statement
iterative methods Chebyshev methods	64 Numerical Analysis Unclassified-Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	33	A03

End of Document