SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1079–1101
© 2000 Society for Industrial and Applied Mathematics

RESIDUAL-MINIMIZING KRYLOV SUBSPACE METHODS FOR STABILIZED DISCRETIZATIONS OF CONVECTION-DIFFUSION EQUATIONS*

OLIVER G. ERNST†

Abstract. We discuss the behavior of the minimal residual method applied to stabilized discretizations of one- and two-dimensional model problems for the stationary convection-diffusion equation. In the one-dimensional case, it is shown that eigenvalue information for estimating the convergence rate of the minimal residual method is highly misleading due to the strong nonnormality of these operators for large grid Péclet numbers. It is also shown that the field of values is a more reliable tool for assessing the convergence rate. In the two-dimensional model problems considered, we observe two distinct phases in the convergence of the iterative method: the first determined by the field of values and the second by the spectrum. We conjecture that the first phase lasts as long as the longest streamline takes to traverse the grid with the flow.

Key words. convection-diffusion equation, Krylov subspace methods, streamline diffusion, field of values

AMS subject classifications. 65F10, 65N30, 65N22

PII. S0895479897325761

1. Introduction.

It is well known [12] that when convection-diffusion problems are discretized using centered schemes such as central differences or the Galerkin finite element method, nonphysical oscillations can occur in the discrete solution whenever convection is the dominating term. In the context of finite element discretizations, an approach to remedy this convective instability has been developed by Brooks and Hughes [2] and Johnson, Nävert, and Pitkäranta [16]. These techniques are called the *streamline upwind Petrov Galerkin* (SUPG) and the *Galerkin least-squares* (GLS) methods and are based on adding a term to the variational formulation of the problem which is proportional to the residual of the discrete solution on each element. Extensive discussions of these stabilization techniques can be found in the recent monographs of Morton [18] and Roos, Stynes, and Tobiska [22].

Little attention, however, has been devoted to the effect of this stabilization on the resulting discrete linear system of equations and its solution by iterative methods. In this paper, we consider one- and two-dimensional model problems, compare the properties of the resulting discretization matrices, and analyze the convergence of residual minimizing Krylov subspace methods applied to these linear systems. We discover two things: First, for the one-dimensional Dirichlet problem, the stabilized discrete operators are highly nonnormal. As a result, using spectral information to estimate the convergence rate of Krylov subspace methods is potentially misleading (cf. [19, 24]) and it is shown that this is indeed so. While most of the examples of highly nonnormal matrices in, e.g., [24] are contrived mathematical pathologies, the one-dimensional stabilized convection-diffusion discretization arises very naturally in applications. We also show that, for this class of problems, the field of values (cf. [7]) is a more reliable tool for assessing the convergence rate. Our second observation con-

*Received by the editors August 8, 1997; accepted for publication (in revised form) by M. Eiermann December 18, 1998; published electronically March 21, 2000.
http://www.siam.org/journals/simax/21-4/32576.html
†Institut für Angewandte Mathematik II, TU Bergakademie Freiberg, Germany (ernst@math.tu-freiberg.de).

cerns the two-dimensional model problem. It is observed that the convergence of the minimal residual method consists of two distinct phases of linear convergence whose rates seem to be determined by the field of values and the spectrum, respectively. Thus, in this case the nonnormality only affects convergence in the first phase, whose duration we conjecture to be the number of iteration steps it takes for information to traverse the underlying grid along the longest streamline.

The paper is organized as follows. Section 2 introduces the continuous problem, its standard Galerkin discretization, and two variants of finite element stabilization techniques known as the SUPG and GLS methods. Section 3 introduces residual minimizing Krylov subspace techniques and reviews classical and more recent convergence results. In section 4 we consider a one-dimensional model problem and show that the field of values gives a much better estimate of the rate of convergence than the spectrum. Section 5 performs a computational study of two model problems in two dimensions.

2. The convection-diffusion equation and its stabilized discretization. This section briefly reviews the boundary value problem under consideration and its discretization by stabilized finite element methods.

2.1. The continuous problem. We consider a bounded domain $\Omega \subset \mathbb{R}^d$ on which a given solenoidal velocity field $\boldsymbol{a} : \Omega \to \mathbb{R}^d$ and a diffusivity tensor $\boldsymbol{\kappa} : \Omega \to \mathbb{R}^{d\times d}$ are defined. We seek a function $u : \Omega \to \mathbb{R}$ which satisfies the differential equation

$$-\nabla\cdot(\boldsymbol{\kappa}\nabla u - \boldsymbol{a}u) = f \tag{2.1}$$

in Ω with a given source term $f : \Omega \to \mathbb{R}$. Appropriate boundary conditions for this problem are the specification of the total flux $-\boldsymbol{n}\cdot\nabla u + \boldsymbol{a}u$ along the inflow boundary (where $\boldsymbol{a}\cdot\boldsymbol{n} < 0$, $\boldsymbol{n}$ being the exterior unit normal), the convective flux $\boldsymbol{a}u$ along the outflow boundary, or Dirichlet conditions on either part of the boundary. We restrict ourselves to Dirichlet conditions

$$u = g \quad \text{ on } \Gamma = \partial\Omega. \tag{2.2}$$

2.2. Stabilized finite element methods. The variational formulation of problem (2.1), (2.2) is to find a function $u \in V$ which satisfies

$$a(u, v) = \ell(v) \qquad \forall v \in V$$

with a bilinear form $a : V \times V \to \mathbb{R}$ and a linear functional $\ell : V \to \mathbb{R}$ given by

$$a(u, v) = \int_\Omega \nabla v \cdot (\boldsymbol{\kappa}\nabla u - \boldsymbol{a}u)\, dx \quad \text{ and } \quad \ell(v) = \int_\Omega vf\, dx,$$

along with a suitable trial/test space V depending on where essential boundary conditions are imposed. By choosing trial and weighting space to coincide, we make the assumption that all essential boundary conditions have been made homogeneous. For a pure Dirichlet problem we may choose the usual Sobolev space $V = H_0^1(\Omega)$.

Given a finite-dimensional subspace V_h of V, the Galerkin finite element method computes an approximate solution $u_h \in V_h$ determined by

$$a(u_h, v) = \ell(v) \qquad \forall v \in V_h. \tag{2.3}$$

The SUPG and GLS stabilizations of the Galerkin discretization lead to modified bilinear forms and right-hand side functionals, which we shall denote by $a_j^h : V_h \times V_h \to \mathbb{R}, j = 1, 2$, and $\ell_j^h : V^h \to \mathbb{R}, j = 1, 2$, respectively. When referring to both methods we shall omit the index j. To define these quantities we introduce the diffusive part $L^D u = -\nabla\cdot(\boldsymbol{\kappa}\nabla u)$ and the advective part $L^A u = \boldsymbol{a} \cdot \nabla u$ of the advection-diffusion operator $L = L^D + L^A$. The SUPG method is defined by

$$a_1^h(u, v) = a(u, v) + \sum_{K\in\mathcal{T}_h} (Lu, \tau L^A v)_K, \quad u, v \in V_h, \tag{2.4}$$

$$\ell_1^h(v) = \ell(v) + \sum_{K\in\mathcal{T}_h} (f, \tau L^A v)_K, \quad v \in V_h, \tag{2.5}$$

where K denotes an arbitrary element in the finite element mesh $\mathcal{T}_h$, $(\cdot, \cdot)_K$ denotes the L^2 inner product on K, and τ denotes an appropriately chosen stability parameter. The corresponding terms for GLS are given by

$$a_2^h(u, v) = a(u, v) + \sum_{K\in\mathcal{T}_h} (Lu, \tau Lv)_K, \quad u, v \in V_h, \tag{2.6}$$

$$\ell_2^h(v) = \ell(v) + \sum_{K\in\mathcal{T}_h} (f, \tau Lv)_K, \quad v \in V_h, \tag{2.7}$$

i.e., the stabilization term weights with the full operator instead of just the advective part. This makes it applicable to more general problems.

Following [3], we choose the stabilization parameter τ as

$$\tau = \frac{h}{2|\boldsymbol{a}|}\xi(\alpha) \qquad \text{with} \qquad \xi(\alpha) = \coth(\alpha) - \frac{1}{\alpha}, \tag{2.8}$$

where $|\cdot|$ denotes the Euclidean length of a vector and the parameter $\alpha = ah/2\kappa$ is the *grid Péclet number*, which measures the strength of convection versus diffusion relative to the mesh size. This choice leads to nodally exact solutions for the one-dimensional constant-coefficient problem and has been shown to converge with order $\mathcal{O}(h^{p+1/2})$ in the L^2-norm for higher dimensions, where p denotes the maximal degree of complete polynomials used in the finite element approximation (cf. [26]).

For the following, we make the assumption that the diffusivity tensor $\boldsymbol{\kappa}$ is isotropic, i.e., diagonal and elementwise constant. Moreover, we assume that the finite element space consists of either piecewise linear or piecewise bilinear functions. In this case we have

$$(Lu, \tau Lv)_K = (Lu, \tau L^A v)_K = (L^A u, \tau L^A v)_K = (\tau \boldsymbol{a}\boldsymbol{a}^T \nabla u, \nabla v)_K,$$

i.e., the stabilization term has the form of an additional diffusivity tensor given by $\tau\boldsymbol{a}\boldsymbol{a}^T$, which acts only in the direction of the flow. For this reason this type of stabilization scheme is also known as the *streamline diffusion* method.

3. Iterative solution of the discrete system. In this section we briefly review some well-known facts about residual-minimizing Krylov subspace methods—as implemented, e.g., by the popular GMRES algorithm of Saad and Schultz [23]—for solving the discrete linear system. The setting is the complex vector space $\mathbb{C}^n$ endowed with an arbitrary (not necessarily the Euclidean) inner product $(\cdot, \cdot)$. We denote the associated vector norm and induced matrix norm by $\|\cdot\|$.

3.1. Minimal residual methods. Minimal residual methods belong to the family of Krylov subspace algorithms, in which approximations to the solution $\boldsymbol{u}$ of the matrix equation $A\boldsymbol{u} = \boldsymbol{f}$ are sought in the sequence of *shifted Krylov spaces* $V_m := \boldsymbol{u}_0 + K_m(A, \boldsymbol{r}_0)$, where

$$K_m(A, \boldsymbol{r}_0) := \operatorname{span}\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^{m-1}\boldsymbol{r}_0\}, \quad m = 1, 2, \ldots,$$

is the mth Krylov space with respect to A and $\boldsymbol{r}_0 = \boldsymbol{f} - A\boldsymbol{u}_0$, the residual vector of the initial guess $\boldsymbol{u}_0$. The vectors $\boldsymbol{v}_m \in V_m$ are of the form $\boldsymbol{v}_m = \boldsymbol{u}_0 + q_{m-1}(A)\boldsymbol{r}_0$ with a polynomial $q_{m-1} \in \Pi_{m-1}$, the space of polynomials of degree not exceeding $m-1$. The residual of a vector in this space can then be written in the form $\boldsymbol{r}_m = \boldsymbol{f} - A\boldsymbol{v}_m = p_m(A)\boldsymbol{r}_0$ with a polynomial p_m in the set $\Pi_m^* := \{p \in \Pi_m : p(0) = 1\}$. The difference $\boldsymbol{e}_m = \boldsymbol{u} - \boldsymbol{v}_m$ between $\boldsymbol{v}_m \in V_m$ and the solution $\boldsymbol{u}$ is then given by $\boldsymbol{e}_m = p_m(A)\boldsymbol{e}_0$, where $\boldsymbol{e}_0 = \boldsymbol{u} - \boldsymbol{u}_0$ is the error in the initial guess. In particular, for the given vector norm and its induced operator norm,

$$\|\boldsymbol{e}_m\| = \|p_m(A)\boldsymbol{e}_0\|, \quad \text{implying} \quad \frac{\|\boldsymbol{e}_m\|}{\|\boldsymbol{e}_0\|} \le \|p_m(A)\|$$

with the analogous inequalities holding for $\boldsymbol{r}_m$. Thus, for a given matrix A, a Krylov subspace method will succeed in reducing error and residual for all right-hand sides and initial vectors to the extent it is able to find residual polynomials $p_m \in \Pi_m^*$ which make $\|p_m(A)\|$ small.

Different Krylov subspace algorithms result from different strategies for selecting a particular approximation $\boldsymbol{u}_m \in V_m$ or, equivalently, selecting a polynomial $p_m \in \Pi_m^*$. One possible and well-defined approach is to choose $\boldsymbol{u}_m$ to minimize the norm of the residual, i.e., that

$$\|\boldsymbol{r}_m\| = \min_{\boldsymbol{v} \in V_m} \|\boldsymbol{f} - A\boldsymbol{v}\|.$$

We will refer to algorithms which implement this strategy as *minimal residual methods*. The GMRES algorithm of Saad and Schultz [23] is an implementation of this method for general nonsingular linear systems of equations which employs an orthonormal basis of $K_m(A, \boldsymbol{r}_0)$ which is augmented by one basis vector in each iteration step using the Arnoldi process. Two other well-known algorithms implementing the minimal residual approach are MINRES by Paige and Saunders [21] for Hermitian indefinite systems and the conjugate residual (CR) method for Hermitian positive definite systems, which was first described in [14] by Hestenes and Stiefel on the conjugate gradient algorithm.

3.2. Error bounds. The most well-known convergence result, which follows immediately from the relations discussed above, is based on spectral properties of the matrix A.

THEOREM 3.1. *Let $A = V\Lambda V^{-1}$ be diagonalizable and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ contain the eigenvalues on its diagonal. Then the residual $\boldsymbol{r}_m$ after m steps of the GMRES algorithm satisfies*

$$\frac{\|\boldsymbol{r}_m\|}{\|\boldsymbol{r}_0\|} \le \|Vp_m(\Lambda)V^{-1}\| \le \operatorname{cond}(V) \min_{p_m \in \Pi_m^*} \|p_m\|_{\Lambda(A)}, \tag{3.1}$$

where $\|\phi\|_D$ denotes the maximum value of the function ϕ on the compact set D and $\operatorname{cond}(V) = \|V\|\,\|V^{-1}\|$ is the condition number of the eigenvector matrix with respect to the norm $\|\cdot\|$.

If A is a normal matrix, then $\mathrm{cond}(V) = 1$, and hence in this case the bound (3.1) for the convergence rate of the minimal residual method depends entirely on how rapidly the quantity $\max_{p\in\Pi_m^*} |p|_{\Lambda(A)}$ decreases with m, i.e., the issue is reduced to a polynomial approximation problem on the discrete set $\Lambda(A)$ of eigenvalues of the matrix A. The class of normal matrices includes the important cases of Hermitian, skew-Hermitian, and circulant matrices. In the general nonnormal case, however, $\mathrm{cond}(V)$ may become sufficiently large to have a profound influence on the convergence rate. This is particularly important when addressing the question of the convergence rate for a whole sequence of parameter-dependent matrices such as those belonging to the discretization of a convection-diffusion problem on successively refined meshes or of problems with increasing Péclet numbers on one given mesh. Although often ignored in applications, this effect may be strong enough to make any available eigenvalue information completely useless (cf. [24]).

One approach for addressing this difficulty is due to Trefethen [25], who has derived residual bounds based on *pseudospectra* of the matrix A. For a positive number ϵ, the associated ϵ-pseudospectrum of A is the set in the complex plane defined by $\Lambda_\epsilon(A) = \{z : \|(zI - A)^{-1}\| \geq 1/\epsilon\}$. This set contains the spectrum of A, hence its boundary Γ_ϵ may be used in the resolvent integral representation of any analytic function of A, in particular of any polynomial of A. This results in the bound

$$\|p_m(A)\| \leq \frac{\ell(\Gamma_\epsilon)}{2\pi\epsilon}\|p_m\|_{\Gamma_\epsilon}$$

for the norm of a residual polynomial p_m applied to A, where $\ell(\Gamma_\epsilon)$ denotes the length of the curve Γ_ϵ, which in turn implies the bound

$$\frac{\|\boldsymbol{r}_m\|}{\|\boldsymbol{r}_0\|} \leq \frac{\ell(\Gamma_\epsilon)}{2\pi\epsilon} \min_{p_m\in\Pi_m^*} \|p_m\|_{\Gamma_\epsilon}$$

for the residual reduction.

Pseudospectra can sometimes result in much more realistic bounds than (3.1) but are expensive to compute. Moreover, it is not always clear which value of ϵ leads to the most useful information. In this paper we turn to another set associated with a matrix A for predicting the convergence rate of minimum residual methods, namely, its *field of values*

$$W(A) = \left\{ \frac{(A\boldsymbol{z}, \boldsymbol{z})}{(\boldsymbol{z}, \boldsymbol{z})} : \quad 0 \neq \boldsymbol{z} \in \mathbb{C}^n \right\} = \{(A\boldsymbol{z}, \boldsymbol{z}) : \|\boldsymbol{z}\| = 1\},$$

sometimes also called its *numerical range*. The field of values of a matrix is a convex and compact set in the complex plane which contains the eigenvalues. For normal matrices $W(A)$ coincides with the convex hull of $\Lambda(A)$, whereas for nonnormal matrices it may be considerably larger. A measure for the size of $W(A)$ is the *numerical radius* $\mu(A) := \max\{|\zeta| : \zeta \in W(A)\}$. The numerical radius is related to the norm associated with the underlying inner product by $\frac{1}{2}\|A\| \leq \mu(A) \leq \|A\|$. Further properties of $W(A)$ can be found in [15, Chapter 1]. The bound (3.1) relies on the fact that the eigenvalues of $p(A)$ are related to those of A via the spectral mapping theorem. An analogous mapping theorem for the field of values does not hold in general. However, a useful result recently obtained by Eiermann [8] does hold for convex sets and a special sequence of polynomials, the Faber polynomials associated with this set.

THEOREM 3.2. *If $\{F_m\}_{m=0}^\infty$ denotes the sequence of Faber polynomials of the field of values $W(A)$ of a matrix $A \in \mathbb{C}^{n\times n}$, then the numerical radius $\mu(F_m(A))$ satisfies*

$$\mu(F_m(A)) \leq \|F_m\|_{W(A)}. \tag{3.2}$$

Proof. See [8] for the proof. ∎

The crucial point of this result is that it relates the field of values of $F_m(A)$ to the size of the polynomial F_m on the set $W(A)$. To use Faber polynomials for estimating the convergence rate of Krylov subspace methods, we turn to the normalized Faber polynomials $\hat{F}_m(z) := F_m(z)/F_m(0)$, which are admissible as residual polynomials. Hence we must require $F_m(0) \neq 0$, which is assured if we assume $0 \notin W(A)$. The asymptotic behavior of these polynomials as $m \to \infty$ is well understood: The normalized Faber polynomials $\{\hat{F}_m\}$ of a convex bounded set $0 \notin D \subset \mathbb{C}$ satisfy $\|\hat{F}_m\|_D \leq c_m \gamma^m$ with $0 < \gamma < 1$ and $c_m < 2/(1-\gamma^m)$. The number γ is known as the asymptotic convergence factor of the set D. In fact, if $p_m^* \in \Pi_m^*$ minimizes $\|p\|_D$ over all $p \in \Pi_m^*$, then

$$\gamma = \lim_{m\to\infty} \|p_m^*\|_D^{1/m}.$$

If p_m denotes the mth residual polynomial selected in the mth step of the minimal residual method, then the minimization property implies

$$\|\boldsymbol{r}_m\| = \|p_m(A)\boldsymbol{r}_0\| \leq \|\hat{F}_m(A)\boldsymbol{r}_0\| \leq \|\hat{F}_m(A)\|\,\|\boldsymbol{r}_0\|$$

for the residual after m steps of the algorithm. Theorem 3.2, together with the asymptotic properties of normalized Faber polynomials, thus yields the bound

$$\frac{\|\boldsymbol{r}_m\|_2}{\|\boldsymbol{r}_0\|_2} \leq \|\hat{F}_m(A)\|_2 \leq 2\mu(\hat{F}_m(A)) \leq 2\|\hat{F}_m\|_{W(A)} \leq 2c_m\gamma^m.$$

3.3. A special case. The one-dimensional model problem with constant coefficients and Dirichlet boundary conditions introduced in the next section is particularly easy to analyze since in this case the field of values is an ellipse, for which both the asymptotic convergence factor γ as well as the Faber polynomials are known explicitly. In particular, the Faber polynomials are just suitably scaled and shifted Chebyshev polynomials of the first kind.

The asymptotic convergence factor γ for domains D not containing the origin whose complement with respect to the extended complex plane $\hat{\mathbb{C}}$ is simply connected may be determined using conformal mapping: In this case there exists a conformal map Φ of the complement $\hat{\mathbb{C}}\backslash D$ of D to the exterior of the unit disk. The asymptotic convergence factor $\gamma = \gamma(D)$ is then given by (cf. [9])

$$\gamma = \frac{1}{|\Phi(0)|}. \tag{3.3}$$

For an ellipse $\mathcal{E}_\rho(\sigma,\tau)$ with foci at $\sigma \pm \tau$ defined by

$$\mathcal{E}_\rho(\sigma,\tau) = \left\{ z \in \mathbb{C} : |z-\sigma+\tau| + |z-\sigma-\tau| \leq |\tau|\left(\rho+\rho^{-1}\right)\right\}, \tag{3.4}$$

we obtain

$$\gamma(\mathcal{E}_\rho(\sigma,\tau)) = \rho \cdot \left| -\frac{\sigma}{\tau} + \sqrt{\left(\frac{\sigma}{\tau}\right)^2 - 1} \right|^{-1},$$

in which that branch of the square root is selected which results in $\gamma < 1$. Provided $0 \notin \mathcal{E}_\rho(\sigma,\tau)$, the Faber polynomials are given in terms of the first-kind Chebyshev polynomials T_m by

$$F_m(z) = \frac{T_m(\zeta(z))}{T_m(\zeta(0))}, \quad \zeta(z) = \frac{z-\sigma}{\tau}, \qquad m = 0, 1, \ldots.$$

4. The one-dimensional case. It is instructive to look at the one-dimensional case, since, at least for constant coefficients, all the important quantities associated with the discrete problem can be computed analytically. If we focus on the Dirichlet problem,

$$-\kappa u'' + au' = f \text{ on } (0, L), \qquad u(0) = u(L) = 0,$$

then a Galerkin discretization using piecewise linear elements on a grid with uniform spacing h and N interior mesh points leads to the discrete linear system of equations $A\boldsymbol{u} = \boldsymbol{f}$, in which, after scaling by h/κ, the coefficient matrix A is given by

$$A = \text{ tridiag}(-1-\alpha, 2, -1+\alpha) \in \mathbb{R}^{N\times N}.$$

The stabilized scheme leads to a linear system $\tilde{A}\boldsymbol{u} = \tilde{\boldsymbol{f}}$ with modified right-hand side $\tilde{\boldsymbol{f}}$ and modified coefficient matrix $\tilde{A}$, which, after scaling by $h/(\kappa+\tilde{\kappa})$, is given by

$$\tilde{A} = \text{ tridiag}(-1-\tilde{\alpha}, 2, 1+\tilde{\alpha}) \in \mathbb{R}^{N\times N}.$$

The parameter $\tilde{\alpha} = ah/(2(\kappa+\tilde{\kappa})), \kappa+\tilde{\kappa} = \kappa(1+\alpha\xi)$ may be interpreted as the effective Péclet number of the stabilized scheme.

4.1. Eigensystems. Both discretization matrices A and $\tilde{A}$ are of the form $T = \text{tridiag}(-1-t, 2, -1+t)$ and, using the results on the eigenvalues of tridiagonal Toeplitz matrices collected in the appendix, the eigenvalues of these matrices are given by

$$\tau_j = 2\left[1 + \sqrt{|(1+t)(1-t)|}e^{\frac{i}{2}[\pi+\arg(t-1)]}\cos\left(\frac{j\pi}{N+1}\right)\right], \qquad j = 1, \ldots, N.$$

For the Galerkin discretization matrix A this results in eigenvalues $\{\lambda_j\}_{j=1}^N$ given by

$$\lambda_j = 2\begin{cases} 1 - \sqrt{1-\alpha^2}\cos(j\pi h), & 0 \le \alpha \le 1, \\ 1 + i\sqrt{\alpha^2-1}\cos(j\pi h), & \alpha \ge 1, \end{cases} \qquad j = 1 \ldots, N.$$

The eigenvalues for the stabilized discretization are obtained by replacing the Péclet number α with $\tilde{\alpha} = \tanh(\alpha)$. The effective Péclet number $\tilde{\alpha}$ is a strictly increasing function of α: It equals α in the diffusion-dominated limit $\alpha = 0$ and approaches unity in the convection-dominated limit $\alpha \to \infty$. Hence $\tilde{\alpha} \in [0, 1)$ so that the eigenvalues $\{\tilde{\lambda}_j\}_{j=1}^N$ of $\tilde{A}$ are always real and given by

$$\tilde{\lambda}_j = 2\left(1 - \sqrt{1-\tilde{\alpha}^2}\cos(j\pi h)\right), \qquad j = 1, \ldots, N.$$

This results in the following qualitative behavior of the eigenvalues for varying Péclet number: In the diffusion-dominated limit both spectra lie in the real interval $[0, 4]$ with a slight clustering at the endpoints. As α approaches the critical value of unity, the eigenvalues of A coalesce at the value 2 and, as α grows beyond 1, the eigenvalues of A lie on the complex interval $[2 - i\alpha, 2 + i\alpha]$ parallel to the imaginary axis. In contrast, the eigenvalues of $\tilde{A}$ always lie on a real interval symmetric with respect to two which shrinks monotonically as α increases. The diameter of this interval is

$$\tilde{\lambda}_{\max} - \tilde{\lambda}_{\min} = \frac{4\cos\pi h}{\cosh\alpha},$$

which decreases at an exponential rate as the Péclet number α increases.

The GMRES convergence bound (3.1) is the product of the condition number of the eigenvector matrix and the maximum norm of the GMRES polynomial on the spectrum of the matrix. By the minimization property of GMRES, we obtain an upper bound by replacing the mth GMRES polynomial with the shifted and scaled Chebyshev polynomial

$$p_m(z) = \frac{T_m(\zeta(z))}{T_m(\zeta(0))}, \qquad \zeta(z) = \frac{z-\sigma}{\tau},$$

where T_m is the first-kind Chebyshev polynomial of degree m and $\sigma = (\lambda_1 + \lambda_N)/2$ is the center of the spectrum, which in our case is a line segment centered at $z = 2$ parallel to either the real or imaginary axis. By employing the bound (A.8) on the polynomials p_m from the appendix, we obtain the bounds

$$\mathrm{cond}_2(V)\,\|p_m\|_{\Lambda(A)} \le \left|\frac{\alpha+1}{\alpha-1}\right|^{(N-1)/2} \frac{2\gamma^m}{1-\gamma^{2m}}$$

for the matrix A and

$$\mathrm{cond}_2(\tilde{V})\,\|p_m\|_{\Lambda(\tilde{A})} \le e^{(N-1)\alpha}\frac{2\tilde{\gamma}^m}{1-\tilde{\gamma}^{2m}}$$

for the matrix $\tilde{A}$, respectively, where

$$\gamma = |\Phi(\zeta(0))|^{-1}, \qquad \Phi(\zeta) = \zeta + \sqrt{\zeta^2-1}, \;\; \zeta \notin [-1,1],$$

is the asymptotic convergence factor of $\Lambda(A)$, and $\tilde{\gamma}$ is the corresponding quantity for $\Lambda(\tilde{A})$.

If only the spectral distributions of A and $\tilde{A}$ are considered, the residual bound (3.1) would indicate that GMRES should converge much more rapidly for $\tilde{A}$ than for A in the convection-dominated case, since then the spectrum of the former rapidly shrinks to a point, hence low order polynomials are sufficient to yield small values on $\Lambda(\tilde{A})$. Numerical experience, however, results in almost identical behavior of GMRES for both systems.

An indication that something is going wrong is obtained from looking at the second term in (3.1), the condition number of the eigenvector matrix. Drawing again from the results in the appendix, the eigenvector matrix V of the tridiagonal Toeplitz matrix T has the form $V = DU$ with an orthogonal matrix U and a diagonal matrix

$$D = \mathrm{diag}(\delta, \ldots, \delta^N), \quad \text{where} \quad \delta = \delta(t) = \begin{cases} \sqrt{\frac{1+t}{1-t}}, & 0 < t < 1, \\ i\sqrt{\frac{t+1}{t-1}}, & t > 1. \end{cases}$$

Hence, the spectral condition number of the eigenvector matrix V is given by

$$\mathrm{cond}_2(V) = \|V\|_2\|V^{-1}\|_2 = \|D\|_2\|D^{-1}\|_2 = |\delta|^{N-1},$$

which, written in terms of the grid Péclet number α, yields

$$\mathrm{cond}_2(V) = \left|\frac{\alpha+1}{\alpha-1}\right|^{(N-1)/2} \qquad \text{and} \qquad \mathrm{cond}_2(\tilde{V}) = e^{(N-1)\alpha}$$

for the matrices A ($t = \alpha$) and $\tilde{A}$ ($t = \tilde{\alpha} = \tanh\alpha$), respectively. We see that $\mathrm{cond}_2(V)$ is bounded in the limit $\alpha \to \infty$ while $\mathrm{cond}(\tilde{V})$ grows exponentially.

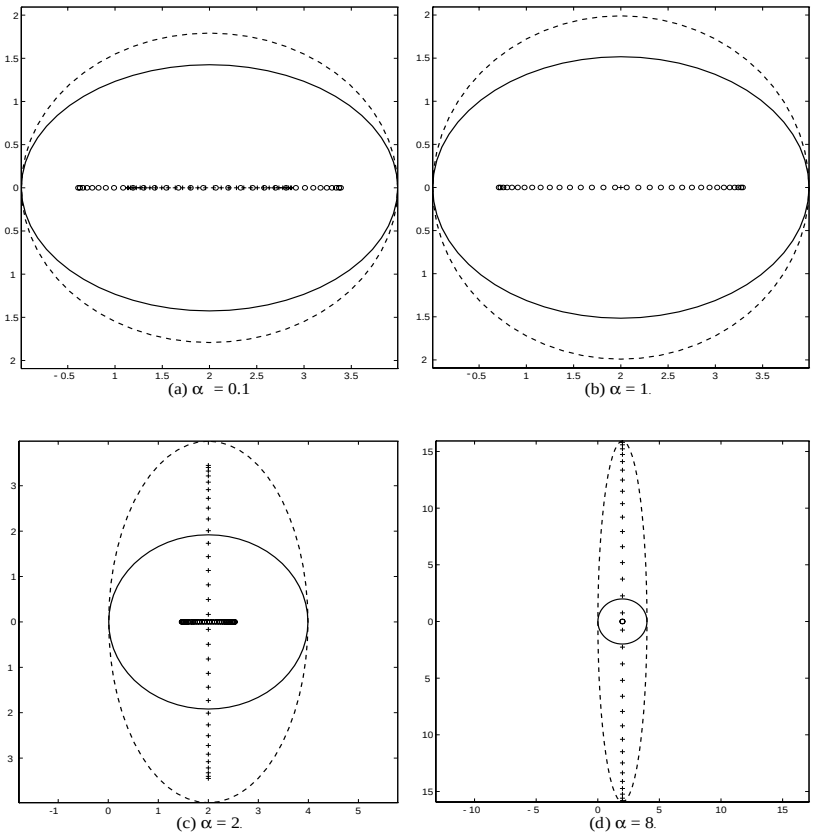

FIG. 4.1. $\Lambda(A)$ *(crosses),* $\Lambda(\tilde{A})$ *(circles),* $W(A)$ *(dashed line), and* $W(\tilde{A})$ *(solid line) for* $N = 32$ *and* $\alpha = 0.1$*,* 1*,* 2*, and* 8*.*

4.2. Field of values. From the results collected in the appendix, the field of values of the tridiagonal Toeplitz matrix $T = \text{tridiag}(-1-t, 2, -1+t)$ is the ellipse $\mathcal{E}_\rho(\sigma, \tau)$ with

$$\sigma = 2, \qquad \tau = 2\cos(\pi h)\sqrt{1-t^2}, \qquad \rho = \left|\frac{1+t}{1-t}\right|^{1/2}.$$

In particular, for $t < 1$ the focal line lies on the real axis, while for $t > 1$ it lies parallel to the imaginary axis. Thus, since $\tilde{\alpha} \in [0, 1)$ for all values of the grid Péclet number, the field of values of the stabilized matrix $\tilde{A}$ is always an ellipse with a real focal line. As derived in the appendix, the field of values of the one-dimensional convection-diffusion operator is a parabola, and for decreasing values of h the elliptical fields of values of A and $\tilde{A}$, suitably rescaled, approximate this parabola near its intersection with the real axis. In the convection-dominated limit $\alpha \to \infty$, the field of values degenerates to a circle centered at $z = 2$ with radius $r = 2\cos(\pi h)$. Figure 4.1 shows the spectra and fields of values for the discretization of the one-dimensional model problem on a mesh with 32 interior points and $\alpha = 0.1, 1, 2$, and 8.

4.3. Numerical experiments. The asymptotic convergence factors of the field of values and the convex hull of the spectrum of A and $\tilde{A}$ of dimension $N = 255$ for

TABLE 4.1
Asymptotic convergence factors, $\alpha = 0.1, 1, 2,$ *and* 8.

	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 2$	$\alpha = 8$
$W(A)$	0.9992	0.9999	1	1
$CH(\Lambda(A))$	0.9038	0	0.5773	0.8819
$W(\tilde{A})$	0.9992	0.9999	0.9999	0.9999
$CH(\Lambda(\tilde{A}))$	0.9041	0.3678	0.1353	3.354e-04

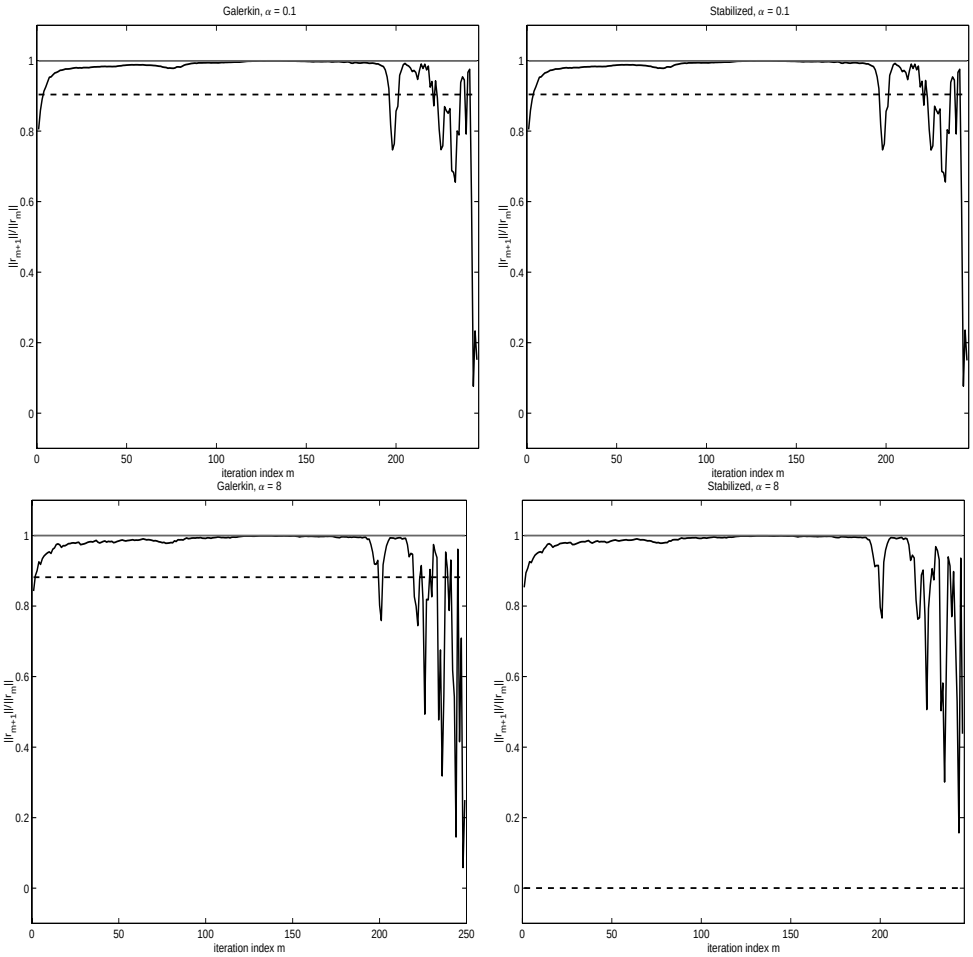

FIG. 4.2. *GMRES "lifespan" curves for* A *(left) and* $\tilde{A}$ *(right) for the cases* $\alpha = 0.1$ *and* $\alpha = 8$. *The solid vertical line represents the asymptotic convergence factor predicted by the field of values and the dashed vertical line represents that predicted by the spectrum.*

$\alpha = 0.1, 1, 2,$ and 8 are given in Table 4.1 to four digits. This table shows that, while the spectral information alone would predict considerably faster convergence for the stabilized discretization as soon as $\alpha > 1$, the fields of values of both discretization matrices indicate the same slow convergence rate of almost unity.

The behavior of GMRES for the Galerkin and the stabilized discretizations for the cases $\alpha = 0.1$ and $\alpha = 8$ can be seen in Figure 4.2. Here GMRES was applied to both systems, again of size $N = 255$, with zero initial guess and a random right-hand side.

While right-hand sides arising from the discretization of the boundary value problem resulted in similar behavior, we chose nonetheless to use a random right-hand side to make sure we were observing generic behavior of GMRES for these matrices. Rather than the usual plot of normalized residual norms $\|\boldsymbol{r}_m\|/\|\boldsymbol{r}_0\|$ in a logarithmic scale, we instead plot quotients $\|\boldsymbol{r}_{m+1}\|/\|\boldsymbol{r}_m\|$ of successive residual norms throughout the iteration history. Such *lifespan curves* were introduced by Nevanlinna [20] and give a more accurate view of the rate of convergence in different phases of the iteration. The solid vertical line represents the asymptotic rate of convergence predicted by the field of values and the dashed line represents that predicted by the spectrum. We observe that, after an initial phase of sublinear convergence (cf. [20]), the rate of convergence of GMRES during the linear phase is predicted remarkably well by the field of values, whereas the spectrum would have indicated a much too optimistic rate. For the case $\alpha = 8$, which is a very low Péclet number with regard to applications, the matrix $\tilde{A}$ is basically a Jordan block and the spectrum gives no information whatsoever regarding the convergence rate. Moreover, in all cases GMRES behaves almost identically for the Galerkin and the stabilized discretizations.

5. The two-dimensional case. In this section we consider two model problems on a rectangular domain. The first is a constant-coefficient problem with velocity at an angle to the coordinate axes; the second involves a semicircular velocity field. The numerical experiments will focus only on the stabilized problem since, for interesting grid Péclet numbers, it yields the only physically meaningful discretization.

5.1. First model problem. We consider the Dirichlet problem (2.1), (2.2) on the unit square $\Omega = (0,1) \times (0,1)$ with the constant coefficients $\boldsymbol{\kappa} = \kappa I_2$ and $\boldsymbol{a} = a(\cos\theta, \sin\theta)^T$ with $f = 0$. We discretize the problem using bilinear elements on a uniform rectangular mesh. The resulting stiffness matrices A and $\tilde{A}$ can then be written as the sum of Kronecker products

$$A = M \otimes \left(\kappa_{11} K + \frac{a_1 h}{2} C\right) + \left(\kappa_{22} K + \frac{a_2 h}{2} C\right) \otimes M - \frac{\kappa_{12} + \kappa_{21}}{4} C \otimes C$$

in terms of the matrices

$$M = \frac{1}{6}\ \text{tridiag}(1, 4, 1), \quad K = \ \text{tridiag}(-1, 2, -1), \quad C = \ \text{tridiag}(-1, 0, 1),$$

which are recognized as the mass, stiffness, and gradient matrices of the discretization of the one-dimensional constant-coefficient/uniform mesh model problem using linear elements. The diffusivity tensor $\boldsymbol{\kappa}$ is given by $\boldsymbol{\kappa} = \kappa I$ in the Galerkin case and by $\boldsymbol{\kappa} = \kappa I + \tau \boldsymbol{a}\boldsymbol{a}^T$ in the stabilized case. Using the definition (2.8) again for the stabilization parameter τ, we obtain, after scaling the system by $(\kappa(1 + \alpha\xi))^{-1}$, the coefficient matrix

$$\begin{aligned}\tilde{A} = M \otimes &\left(\frac{1 + \alpha\xi c^2}{1 + \alpha\xi} K + c \tanh(\alpha) C\right) \\ &+ \left(\frac{1 + \alpha\xi s^2}{1 + \alpha\xi} K + s \tanh(\alpha) C\right) \otimes M + \frac{sc}{2} \xi \tanh(\alpha) C \otimes C\end{aligned}$$

for the stabilized discretization, where $c = \cos\theta$, $s = \sin\theta$. The corresponding matrix for the Galerkin case is obtained by setting $\xi = 0$.

Although the Kronecker product representation of $\tilde{A}$ seems simple enough, we were unable to find a closed form representation for the eigenvalues, eigenvectors,

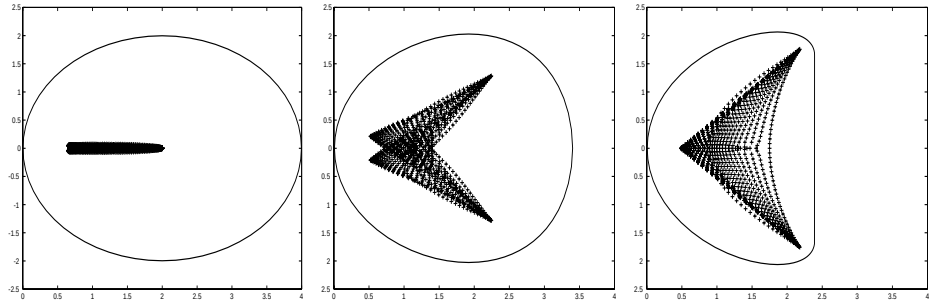

FIG. 5.1. $W(\tilde{A})$ *and* $\Lambda(\tilde{A})$ *for* $N = 31^2$, $\alpha = 1000$, *and* $\theta = 0$, $\pi/8$, *and* $\pi/4$.

and field of values of the discretization matrices of even this simple model problem without making some simplifications. One such simplification results from setting the flow angle θ to zero, which corresponds to flow in the direction of the x-axis (cf. also [11]). In this case $s = 0$, $c = 1$, and the resulting matrix $\tilde{A}_0$ is the sum of only two Kronecker products

$$\tilde{A}_0 = M \otimes (K + \tanh(\alpha)C) + \frac{\tanh(\alpha)}{\alpha} K \otimes M.$$

The first factors of the two terms are both symmetric tridiagonal Toeplitz matrices, and hence they share a common system of orthogonal eigenvectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ such that

$$M\boldsymbol{u}_j = \mu_j \boldsymbol{u}_j \quad \text{and} \quad K\boldsymbol{u}_j = \nu_j \boldsymbol{u}_j, \qquad j = 1, \ldots, n,$$

where the eigenvalues $\{\mu_j\}_{j=1}^n$ and $\{\nu_j\}_{j=1}^n$ and eigenvectors $\{\boldsymbol{u}_j\}_{j=1}^n$ are given in terms of the well-known trigonometric formulas

$$\begin{aligned} \mu_j &= \frac{1}{3}(2 + \cos(j\pi h), \\ \nu_j &= 2(1 - \cos(j\pi h)), \\ \text{and } (\boldsymbol{u}_j)_k &= \sin(jk\pi h), \qquad k = 1, \ldots, n. \end{aligned}$$

By using the properties of Kronecker products, we obtain that a vector $\boldsymbol{u}_j \otimes \boldsymbol{v}_j \in \mathbb{R}^{n^2}$ is an eigenvector of $\tilde{A}_0$ if the vector $\boldsymbol{v}_j \in \mathbb{R}^n$ is an eigenvector of the matrix

$$\mu_j \, (K + \tanh(\alpha)C) + \nu_j \frac{\tanh(\alpha)}{\alpha} M, \quad j = 1, \ldots n.$$

Each value of j yields a nonsymmetric tridiagonal Toeplitz matrix whose eigenvalues and eigenvectors can be determined using the results cited in the appendix.

For other flow angles, we resort to numerical computation. Figure 5.1 shows the computed eigenvalues and the boundary of $W(\tilde{A})$ for $N = 31^2$, $\alpha = 1000$, $\kappa = 1$ and flow angles $\theta = 0$, $\pi/8$, and $\pi/4$. In the second and third cases the eigenvalues were computed with MATLAB's `eig` function and the boundary of the field of values was computed by finding the extremal eigenvalues of the symmetric part of rotated matrices using the Lanczos algorithm combined with Chebyshev acceleration (cf. [1]). We note that, in terms of distance of the boundary $\partial W(\tilde{A})$ to the convex hull of the

TABLE 5.1
Asymptotic convergence factors of $W(\tilde{A})$ and $CH(\Lambda(\tilde{A}))$.

	$\theta = 0$	$\theta = \pi/8$	$\theta = \pi/4$
$W(\tilde{A})$	0.999	0.995	0.997
$CH(\Lambda(\tilde{A}))$	0.308	0.631	0.658

spectrum, the matrix seems farthest from normal in the case of flow along one of the coordinate axes.

Next, we compute the asymptotic convergence factors γ of both the convex hull $CH(\Lambda(\tilde{A}))$ of the spectrum and the field of values for the three cases depicted in Figure 5.1. Since these sets are both simply connected and do not contain the origin, γ may be calculated via the conformal mapping connection (3.3). To this end, we have used the Schwarz–Christoffel Toolbox (version 2.0) of Driscoll [5] to evaluate the exterior mapping Φ of $CH(\Lambda(\tilde{A}))$ and a polygonal approximation of $W(\tilde{A})$, respectively, at the origin. The results are shown in Table 5.1. As expected from the eigenvalue plot, the convergence factors of $CH(\Lambda(\tilde{A}))$ are much smaller than those of $W(\tilde{A})$, the latter being very close to unity due to the proximity of the field of values to the origin.

The convergence behavior of GMRES applied to the linear systems belonging to the three flow angles, $\alpha = 1000$ and $N = 31^2$ is shown in Figure 5.2. In all cases, a zero initial vector and a random right-hand side were used. The figures on the left show the normalized residual norms $\|\boldsymbol{r}_k\|/\|\boldsymbol{r}_0\|$ of GMRES applied to each of the three systems (solid line); the dashed lines indicate the linear convergence rates predicted by the convex hull of the spectrum and by the field of values, respectively. To make sure that the convex hull of the spectrum isn't overestimating the convergence rate, we have also included the residual curve of GMRES applied to the same system with $\tilde{A}$ replaced by a diagonal matrix with the same eigenvalues as $\tilde{A}$ (dotted line). The three plots on the right show the corresponding lifespan curves $\|\boldsymbol{r}_{k+1}\|/\|\boldsymbol{r}_k\|$. We observe two distinct phases of linear convergence in the residual curves of all three cases. The convergence rates of the first phase are slightly below the rate predicted by the field of values. Those of the second phase seem to approach the rate the spectrum would predict, in the absence of nonnormality, although the lifespan curves lie slightly below and above this rate in the first and third examples. The transition between the two phases occurs at iteration steps 35 and 44 for flow angles $\theta = \pi/8$ and $\theta = \pi/4$, respectively, which are upper bounds for the number of steps information would take to traverse the underlying finite element grid. For the case $\theta = 0$, however, this transition does not occur until step 52. Moreover, the transition is much more gradual than in the other two cases.

5.2. Second model problem. Our second model problem is a slight modification of a widely used test problem for discretizations of convection-diffusion equations (cf. [18, p. 10]). The domain is the rectangle $\Omega = (-1, 1) \times (0, 1)$, the diffusion tensor $\boldsymbol{\kappa}$ is the identity, and the incompressible velocity field is given by

$$\boldsymbol{a}(x, y) = 2a \begin{bmatrix} y(1 - x^2) \\ -x(1 - y^2) \end{bmatrix},$$

the semicircular flow pattern of which is shown in Figure 5.3. The parameter a can be used to vary the Péclet number. Along the resulting inflow boundary we impose

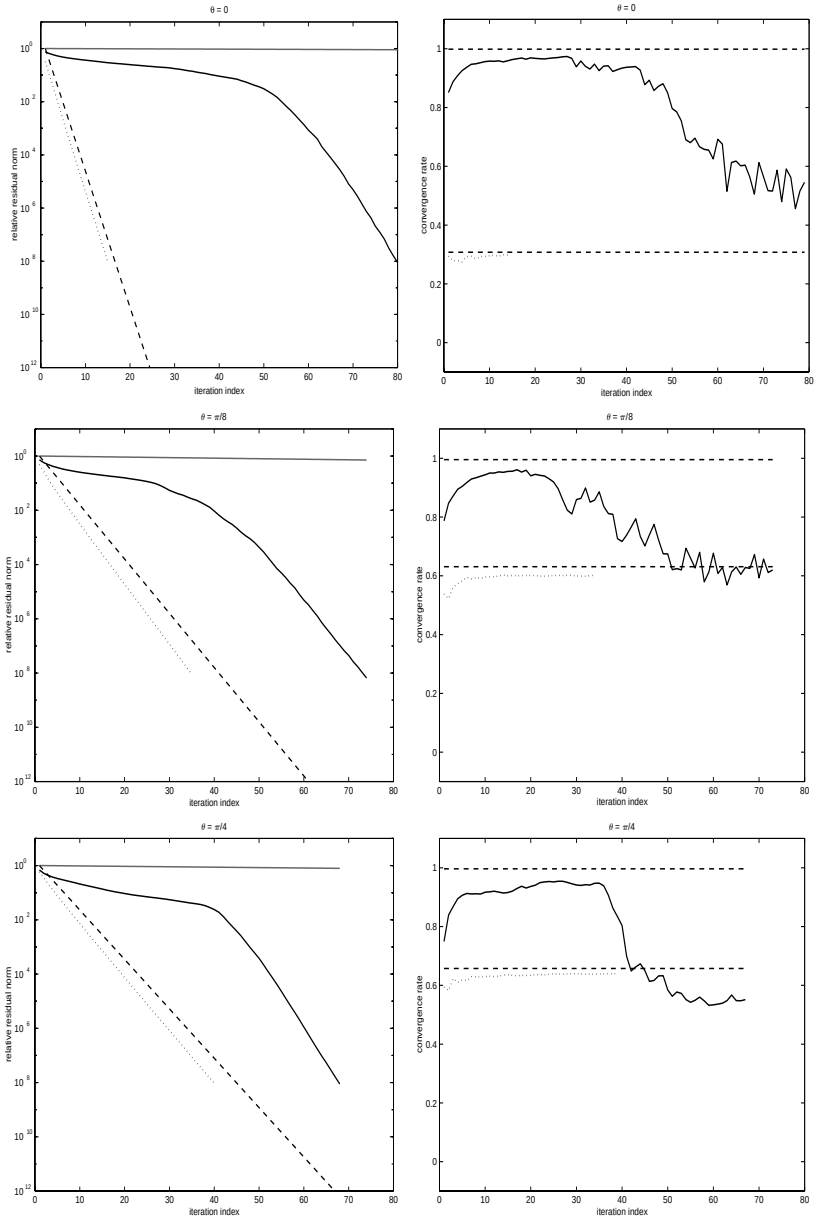

FIG. 5.2. *GMRES residual curves for* $\theta = 0, \pi/8,$ *and* $\pi/4$, $\alpha = 1000$ *and* $N = 31^2$.

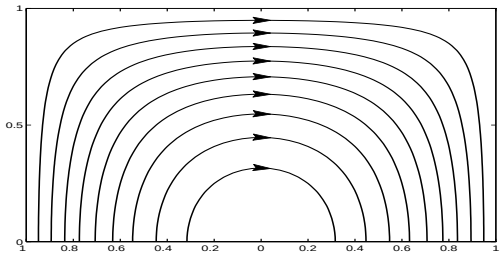

FIG. 5.3. *Velocity field of second model problem.*

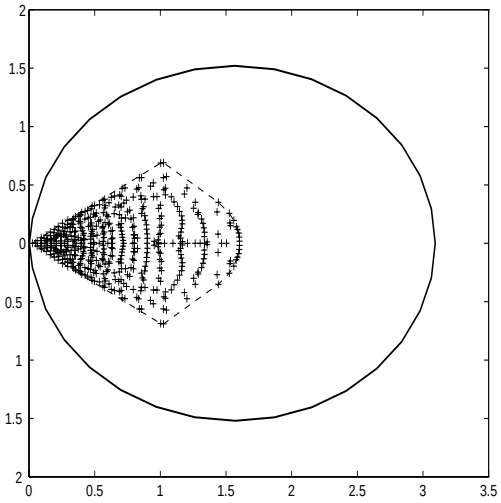

FIG. 5.4. *Field of values, spectrum, and the convex hull of the spectrum for the stabilized discretization of the second model problem on a* 32×16 *grid.*

the Dirichlet condition

$$u(x,0) = 1 + \tanh[\nu(2x+1)], \qquad -1 \le x \le 0,$$

in which another parameter ν determines the sharpness of the inflow profile. The remaining Dirichlet conditions are given by

$$\begin{aligned} &u(x,0) = 0, \qquad 0 \le x \le 1, \\ &u(x,y) = 1 + \tanh(\nu) \qquad \text{on the remaining portion of } \partial\Omega. \end{aligned}$$

Figure 5.4 shows the field of values, the spectrum and the convex hull of the spectrum for this problem with $a = 10^5$ discretized on a uniform mesh of 32×16 bilinear rectangular elements, which corresponds to a grid Péclet number of $\alpha = 6250$. The field of values is more than twice the size in diameter than the spectrum, so some nonnormality effects can be expected. Again, we have scaled the problem by $(\kappa(1 + \alpha\xi))^{-1}$.

In Figure 5.5, the solid line represents the GMRES residual norm curve for this problem. As before, the upper and lower dashed lines show the linear rates of convergence predicted by the asymptotic convergence factors of $W(\tilde{A})$ and $CH(\Lambda(\tilde{A})$, respectively. The dotted line is the residual curve of GMRES applied to a diagonal matrix $\tilde{D}$ with the same eigenvalues as $\tilde{A}$ using a zero initial guess and a random right-hand side. Since $\tilde{D}$ is a normal matrix, the convergence of GMRES is completely determined by $\Lambda(\tilde{D}) = \Lambda(\tilde{A})$. Again we observe two distinct phases of linear convergence. The rate in the first phase is somewhat overestimated by the field of values but noticeably smaller than the rate predicted by the spectrum. The convex hull of the spectrum also overestimates the rate in the second phase, but the residual curve of the diagonal matrix is seen to have the same rate as that observed in the second phase. The transition between these two phases takes place at iteration step 40, which is the roughly the number of iteration steps required for the profile prescribed at the inlet boundary to propagate across the mesh to the outflow boundary.

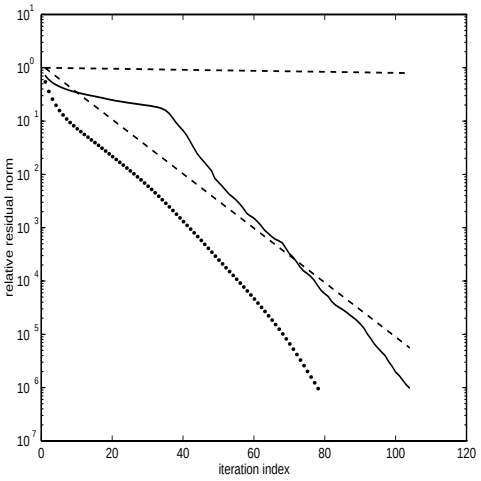

FIG. 5.5. *GMRES residual norms for the solution of the second model problem (solid), linear convergence rate predicted by $W(\tilde{A}$ (upped dashed) and $CH(\Lambda(\tilde{A}))$ (lower dashed), and residual norms for GMRES applied to diagonal matrix D with $\Lambda(D) = \Lambda(\tilde{A})$.*

6. Conclusions. We have tried to gain insight into the convergence behavior of residual minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion problems by studying various model problems. The one-dimensional results show that, while the stabilization results in a better discretization of the boundary value problem, it also results in a highly nonnormal discrete operator. The nonnormality can be characterized by the condition number of the eigenvector matrix, which grows exponentially with both the grid Péclet number and the grid size. A consequence of the high degree of nonnormality is that spectral information is virtually useless for assessing the convergence rate of Krylov subspace methods. We have also seen that the field of values is a viable alternative in this case. In the two-dimensional examples, we observed that, while less pronounced, nonnormality is still an issue. Its effect, as observed in computational experiments, is an initial rate of convergence governed essentially by the field of values, after which convergence governed by the spectrum takes over. It is conjectured that the duration of the initial phase is governed by the time it takes for boundary information to pass from the inflow boundary across the domain following the streamlines of the velocity field.

Appendix A. Ellipses and Chebyshev polynomials. In this section we collect some results pertaining to Chebyshev polynomials on ellipses in the complex plane. Chebyshev polynomials are often used to bound the convergence rate of Krylov subspace methods. For real intervals not containing the origin (and for ellipses with real foci "far enough away" from the origin [10]), scaled Chebyshev polynomials *are* the polynomials of least maximum modulus normalized at the origin. For general ellipses, the Chebyshev polynomials are still asymptotically optimal, i.e., they satisfy $\lim_{m\to\infty} \|p_m\|_\Omega^{1/m} = \gamma(\Omega)$. These results are well known [4, 17, 6] and we include them only for convenient reference.

We parameterize an ellipse in the complex plane by two complex numbers σ and τ such that the former represents the midpoint of the focal line and the two foci lie at $\sigma \pm \tau$. A third parameter $\rho > 1$ is used to parameterize the family of ellipses with these two foci ($0 < \rho < 1$ yields the same family again). The closed interior of an

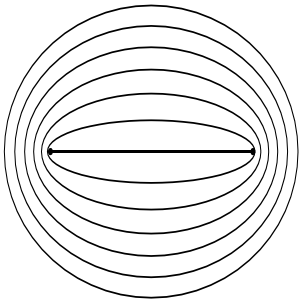

FIG. A.1. *Ellipses in the confocal family* $\mathcal{E}_\rho = \mathcal{E}_\rho(0,1)$, $\rho = 1, 1.25, \ldots, 2.5$.

ellipse can then be characterized in terms of the sum of the distances to the foci as the set

$$\mathcal{E}_\rho(\sigma,\tau) = \left\{ z \in \mathbb{C} : |z-\sigma+\tau| + |z-\sigma-\tau| \le |\tau| \left(\rho + \rho^{-1}\right) \right\}.$$

The boundary of the ellipse may be parameterized as

$$\partial\mathcal{E}_\rho(\sigma,\tau) = \left\{ z = \sigma + \frac{\tau}{2}\left(\rho e^{i\theta} + \rho^{-1} e^{-i\theta}\right) : \theta \in [0, 2\pi) \right\},$$

from which we immediately see that the two semiaxes, i.e., the largest and smallest distances from σ to the boundary of the ellipse, are $|\tau|(\rho \pm \rho^{-1})/2$. It is convenient to consider Chebyshev polynomials on the family of confocal ellipses $\mathcal{E}_\rho = \mathcal{E}_\rho(0,1)$ with foci located at ± 1. A given ellipse $\mathcal{E}_\rho(\sigma,\tau)$ is mapped onto the associated ellipse from this family having the same eccentricity (which is $|\tau|$) by the linear transformation $\zeta = (z-\sigma)/\tau$. The ellipses $\mathcal{E}_\rho$ for $\rho = 1, 1.25, \ldots, 2.5$ are shown in Figure A.1. For $\zeta \in [-1,1]$, the first-kind Chebyshev polynomials can be defined by

$$T_m(\zeta) = \cos(m \arccos\zeta), \quad m = 0, 1, \ldots.$$

These polynomials satisfy the recurrence relation

$$T_0 \equiv 1, \quad T_1(\zeta) = \zeta, \quad T_{m+1}(\zeta) = 2\zeta T_m(\zeta) - T_{m-1}(\zeta) \quad (m > 1) \tag{A.1}$$

and their zeros are all contained in the interval $(-1,1)$. The analysis of the Chebyshev polynomials is facilitated by introducing the Joukowski map

$$\zeta = \Psi(w) := \frac{1}{2}\left(w + \frac{1}{w}\right), \quad w \neq 0,$$

which maps the exterior $|w| > 1$ of the unit circle one-to-one onto $\mathbb{C}\backslash[-1,1]$. Each circle $|w| = \rho > 1$ is mapped to the ellipse $\mathcal{E}_\rho$. The inverse map Φ is given by

$$w = \Phi(\zeta) = \zeta + \sqrt{\zeta^2 - 1}, \quad \zeta \notin [-1,1],$$

in which that branch of the square root is chosen which results in $|\Phi(\zeta)| > 1$. (A method for selecting the proper branch is described in [13, p. 296].) Using the recurrence relation (A.1), one easily verifies that the Chebyshev polynomials satisfy

$$T_m(\zeta) = \frac{1}{2}(w^m + w^{-m}) = \frac{1}{2}[\Phi(\zeta)^m + \Phi(\zeta)^{-m}], \quad \zeta \notin [-1,1].$$

This also holds for $z \in [-1,1]$, but then the w associated with ζ is no longer unique. Setting $w = \rho e^{i\theta}$, this yields

$$T_m(\zeta) = \frac{1}{2}(\rho^m e^{im\theta} + \rho^{-m} e^{-im\theta}), \tag{A.2}$$

from which we conclude that the mth Chebyshev polynomial maps the ellipse $\mathcal{E}_\rho$ to the ellipse $\mathcal{E}_{\rho^m}$, which is covered m times. This implies that

$$\frac{1}{2}(\rho^m - \rho^{-m}) \le |T_m(\zeta)| \le \frac{1}{2}(\rho^m + \rho^{-m}), \quad \zeta \in \mathcal{E}_\rho,$$

or, equivalently, $|T_m(\zeta)| = (\rho^m + t_m \rho^{-m})/2$ with $|t_m| < 1$.

We obtain normalized Chebyshev polynomials p_m on arbitrary ellipses $\mathcal{E}_\rho(\sigma,\tau)$ not containing the origin by shifting and scaling the Chebyshev polynomials on $\mathcal{E}_\rho$:

$$p_m(z) := \frac{T_m(\zeta(z))}{T_m(\zeta(0))} = \frac{T_m((z-\sigma)/\tau)}{T_m(-\sigma/\tau)}, \quad m = 0, 1, \ldots. \tag{A.3}$$

Since $0 \notin \mathcal{E}_\rho(\sigma,\tau)$ implies $\zeta(0) \notin \mathcal{E}_\rho$ and since the zeros of T_m lie in $(-1,1)$, we have $T_m(\zeta(0)) \neq 0$.

We also define

$$\gamma = \gamma(\sigma,\tau) := \frac{1}{|\Phi(\zeta(0))|} = \frac{1}{|\Phi(\frac{-\sigma}{\tau})|}. \tag{A.4}$$

We can now describe the asymptotic properties of the residual polynomials defined by (A.3).

THEOREM A.1. *If σ and τ are arbitrary complex numbers such that $0 \notin \mathcal{E}_\rho(\sigma,\tau)$, then the residual polynomials defined in* (A.3) *satisfy*

$$\|p_m\|_{\mathcal{E}_\rho(\sigma,\tau)} \le (\rho^m + \rho^{-m}) \frac{\gamma^m}{1-\gamma^{2m}}. \tag{A.5}$$

The corresponding bound for $\|p_m\|_{[\sigma-\tau,\sigma+\tau]}$ is obtained by setting $\rho = 1$.

Proof. By their definition, the polynomials p_m satisfy

$$|p_m(z)| \le \frac{1}{2}(\rho^m + \rho^{-m}) \frac{1}{|T_m(\zeta(0))|}. \tag{A.6}$$

It is thus only necessary to examine $|T_m(\zeta(0))| = |T_m(-\sigma/\tau)|$. Since $\gamma^{-1} = |\Phi(\zeta(0))|$ lies outside the unit circle, its image under Ψ—i.e., $\zeta(0)$—lies on the boundary of the ellipse $\mathcal{E}_{\gamma^{-1}}$ and hence, by the mapping properties of the Chebyshev polynomials, $T_m(\zeta(0))$ lies in the ellipse $\mathcal{E}_{\gamma^{-m}}$. This implies

$$\frac{1}{2}(\gamma^{-m} - \gamma^m) \le |T_m(\zeta(0))| \le \frac{1}{2}(\gamma^{-m} + \gamma^m), \tag{A.7}$$

and, together with (A.6), yields (A.5). The assertion for the complex line segment is obtained in the same way by using the fact that $|T_m(\zeta)| \le 1|$ for $\zeta \in [-1,1]$. ∎

Two special cases allow the bounds in Theorem A.1 to be improved, as follows.

THEOREM A.2. *If, in addition to the assumptions of Theorem* A.1*, the foci of $\mathcal{E}_\rho(\sigma,\tau)$ and the origin are colinear, there holds*

$$\|p_m\|_{\mathcal{E}_\rho(\sigma,\tau)} \le (\rho^m + \rho^{-m}) \frac{\gamma^m}{1+\gamma^{2m}}. \tag{A.8}$$

If, in addition to the assumptions of Theorem A.1*, the line connecting* σ *and the origin is perpendicular to the focal line of* $\mathcal{E}_\rho(\sigma,\tau)$*, then*

$$\|p_m\|_{\mathcal{E}_\rho(\sigma,\tau)} \le (\rho^m + \rho^{-m}) \begin{cases} \frac{\gamma^m}{1+\gamma^{2m}} & \text{for } m \text{ even,} \\ \frac{\gamma^m}{1-\gamma^{2m}} & \text{for } m \text{ odd.} \end{cases} \tag{A.9}$$

The corresponding bounds for $\|p_m\|_{[\sigma-\tau,\sigma+\tau]}$ *are obtained by setting* $\rho = 1$*.*

Proof. Since $\gamma^{-1} = |\Phi(\zeta(0))|$, we can write $\Phi(\zeta(0)) = \gamma^{-1}e^{i\alpha}$, with α being the argument of $\Phi(\zeta(0))$. Using the relation (A.2), we obtain

$$|T_m(\zeta(0))| = \left|\frac{1}{2}(\gamma^{-m}e^{im\alpha} + \gamma^m e^{-im\alpha})\right|.$$

Thus,

$$|T_m(\zeta(0))| \le \frac{1}{2}\begin{cases} \gamma^{-m} + \gamma^m & \text{for } \arg(\Phi(\zeta(0))) = \frac{k}{m}\pi, \\ \gamma^{-m} - \gamma^m & \text{for } \arg(\Phi(\zeta(0))) = \frac{2k+1}{2m}\pi, \end{cases} \qquad k \in \mathbb{Z}.$$

If the foci are colinear or, equivalently, $-\sigma/\tau$ is real, this implies that also $\Phi(-\sigma/\tau)$ is real and hence its argument can be written in the form $k\pi/m$ for some suitable $k \in \{0,\dots,2m-1\}$. In the second special case $-\sigma/\tau$, and hence also $\Phi(\sigma/\tau)$, is pure imaginary. For even m, $\arg(\Phi(-\sigma/\tau))$ is equal to $\pi/2$ or $3\pi/2$ and may be written in the form $k\pi/m$. For odd m it can be written in the form $(2k+1)\pi/(2m)$. ∎

As a consequence of Theorem A.1, the asymptotic convergence factor of the residual polynomials p_m on the ellipse $\mathcal{E}_\rho(\sigma,\tau)$ is given by

$$\lim_{m\to\infty} \|p_m\|^{1/m}_{\mathcal{E}_\rho(\sigma,\tau)} = \rho\gamma.$$

Appendix B. Spectral properties and the field of values of tridiagonal Toeplitz matrices. For convenient reference, we include a derivation of the eigenvalues, eigenvectors and field of values of tridiagonal Toeplitz matrices. The presentation follows [6].

B.1. Eigensystem. It is sufficient to consider tridiagonal Toeplitz matrices with zero diagonal

$$T = \text{ tridiag}(\alpha, 0, \beta) \in \mathbb{C}^{n\times n},$$

since otherwise the spectrum is simply shifted by the diagonal term. For $\alpha\beta = 0$ the spectrum consists of one zero eigenvalue of algebraic multiplicity n and, unless $\alpha = \beta = 0$, geometric multiplicity one. Assuming $\alpha\beta \neq 0$, the matrix T may be symmetrized by the diagonal similarity transformation $D^{-1}TD = \eta T_1$, where

$$\eta = \eta(\alpha,\beta) = \sqrt{|\alpha\beta|}e^{\frac{i}{2}[\arg(\alpha)+\arg(\beta)]}$$

and $T_1 = \text{ tridiag}(1,0,1)$ and $D = \text{diag}(\delta, \delta^2, \dots, \delta^n)$ with

$$\delta = \delta(\alpha,\beta) = \sqrt{|\alpha/\beta|}e^{\frac{i}{2}[\arg(\alpha)-\arg(\beta)]}.$$

As is easily verified using basic trigonometric identities, the eigenvalues of T_1 are

$$\mu_j = 2\cos\left(\frac{j\pi}{n+1}\right) \qquad (j = 1,\dots,n) \tag{B.1}$$

with corresponding normalized orthogonal eigenvectors $\boldsymbol{u}_j = [u_{1j}, \ldots, u_{nj}]^T$ given by

$$u_{kj} = \sqrt{\frac{2}{n+1}} \sin\left(\frac{kj\pi}{n+1}\right) \qquad (k, j = 1, \ldots, n).$$

For T this results in the eigenvectors

$$\boldsymbol{t}_j = D\boldsymbol{u}_j = [\delta u_{1j}, \ldots, \delta^n u_{nj}]^T \qquad (j = 1, \ldots, n)$$

and corresponding eigenvalues $\tau_j = \eta\mu_j$ $(j = 1, \ldots, n)$. The eigenvector matrix is $V = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n] = DU$.

B.2. Field of values.

LEMMA B.1. *The field of values of the tridiagonal Toeplitz matrix* $T = \mathrm{tridiag}(\alpha, 0, \beta) \in \mathbb{C}^{n\times n}$ *consists of the closed interior of the ellipse*

$$z = c_n(\alpha e^{i\theta} + \beta e^{-i\theta}), \quad c_n = \cos(\pi/(n+1)), \quad 0 \le \theta < 2\pi. \tag{B.2}$$

Proof. We begin by showing that the field of values of the $n \times n$ Jordan block $J = \mathrm{tridiag}(0, 0, 1)$, which is connected with T by $T = \alpha J^T + \beta J$, is a closed disk centered at the origin. If $\zeta \in W(J)$, then there exists a vector $\boldsymbol{z} \in \mathbb{C}^n$ of unit norm such that $\zeta = \boldsymbol{z}^H J\boldsymbol{z}$. If a new vector $\boldsymbol{w}$ is defined by $w_j = e^{ij\phi}z_j$ $(j = 1, \ldots, n)$ with an arbitrary angle ϕ, then we have $\|\boldsymbol{w}\|_2 = 1$ and

$$\boldsymbol{w}^H J\boldsymbol{w} = \sum_{j=1}^{n-1} \overline{w_j} w_{j+1} = e^{i\phi} \sum_{j=1}^{n-1} \overline{z_j} z_{j+1} = e^{i\phi}\zeta.$$

This shows that, for any $\zeta \in W(J)$, the entire circle $|z| = |\zeta|$ must also belong to $W(J)$. Since the field of values is convex, this means that $W(J)$ must be a disk centered at the origin. The radius of this disk is obtained as the largest eigenvalue of $(J + J^H)/2$. This follows from the characterization of the field of values of a matrix A as the intersection of strips S_θ in the complex plane:

$$W(A) = \bigcap_{\theta\in[0,\pi)} S_\theta, \qquad S_\theta = \{z \in \mathbb{C} : \lambda_{\min}(H_\theta) \le \mathrm{Re}\ (e^{i\theta}z) \le \lambda_{\max}(H_\theta)\},$$

where H_θ is the Hermitian part $H_\theta := (A_\theta + A_\theta^H)/2$ of the rotated matrix $A_\theta = e^{i\theta}A$. By (B.1), the largest eigenvalue of $J + J^T$ is $c_n = \cos(\pi/(n+1))$, hence $W(J) = S(0, c_n)$.

Now assume $\omega \in W(T)$. This means there exists $\boldsymbol{w} \in \mathbb{C}^n$, $\|\boldsymbol{w}\|_2 = 1$, such that

$$\omega = \boldsymbol{w}^H T\boldsymbol{w} = \alpha\boldsymbol{w}^H J^H\boldsymbol{w} + \beta\boldsymbol{w}^H J\boldsymbol{w} = \alpha\overline{\zeta} + \beta\zeta$$

for some $\zeta \in W(J_n)$. Thus, each point in $W(T)$ has the form (B.2) and, conversely, any complex number of the form (B.2) lies in $W(T)$. ∎

We note that, in the notation defined in (3.4), Lemma B.1 states that $W(T) = \mathcal{E}_\rho(\sigma, \tau)$ with

$$\sigma = 0, \quad \tau = 2\cos(\pi/(n+1))\sqrt{|\alpha\beta|}e^{\frac{i}{2}[\arg(\alpha)-\arg(\beta)]}, \quad \text{and} \quad \rho = \sqrt{|\alpha/\beta|}.$$

Appendix C. Field of values and spectrum of the one-dimensional differential operator. For the one-dimensional case, we determine the field of values

$$W(L) = \left\{ \frac{a(u,u)}{(u,u)} : 0 \neq u \in V \right\}$$

of the differential operator with constant coefficients

$$L : V \to V', \quad Lu = -(\kappa u')' + au', \quad V = H_0^1(0,1)$$

associated with the Dirichlet problem considered as an unbounded linear operator on $L^2(0,1)$ with associated bilinear form $a(u,v) = (Lu,v)$. Of course, for an unbounded operator, the field of values will be an unbounded set in the complex plane. Since L^{-1} is a compact operator, its spectrum consists only of eigenvalues and these accumulate at zero. To determine $W(L)$, we thus proceed similarly as in the finite-dimensional case. We consider the rotated operators $L_\theta = e^{i\theta}L, \theta \in (-\pi/2, \pi/2)$ and determine the minimal eigenvalues of their Hermitian parts

$$H_\theta = \frac{1}{2}(L_\theta + L_\theta^*),$$

where the L^2-adjoint operator is characterized by

$$(L_\theta u, v) = (u, L_\theta^* v) \quad \forall u, v \in V.$$

Integration by parts yields $L_\theta^* v = e^{-i\theta}[-(\kappa v')' - av']$, which results in

$$H_\theta u = -\cos(\theta)\kappa u'' + i\sin(\theta)au'$$

for the Hermitian part. The eigenvalues of H_θ are determined by those values of $\lambda = \lambda(\theta)$ for which the boundary value problem

$$\begin{gathered} H_\theta u = -\cos(\theta)\kappa u'' + i\sin(\theta)au' = \lambda u, \quad x \in (0,1), \\ u(0) = u(1) = 0 \end{gathered}$$

possesses nontrivial solutions. Solutions of the form $u(x) = e^{\mu x}$ are determined by the solutions $\mu_\pm$ of the characteristic equation

$$-\cos(\theta)\kappa\mu^2 + ia\sin(\theta)\mu - \lambda = 0.$$

After imposing the homogeneous boundary condition, this results in

$$\lambda_k(\theta) = k^2\pi^2\kappa\cos(\theta) - \frac{a^2\sin^2(\theta)}{4\kappa\cos(\theta)}, \quad k \in \mathbb{Z}.$$

Since $k = 0$ results in a constant (zero) eigenfunction, and $\cos(\theta) > 0$ for $\theta \in (-\pi/2, \pi/2)$, the smallest eigenvalue of H_θ is thus obtained for $k = 1$ as

$$\lambda_{\min}(\theta) = \pi^2\kappa\cos(\theta) - \frac{a^2\sin^2(\theta)}{4\kappa\cos(\theta)}.$$

We obtain the field of values $W(L)$ as the intersection of all half-planes

$$S_\theta = \{z \in \mathbb{C} : \lambda_{\min}(\theta) \leq \mathrm{Re}(e^{i\theta}z)\}, \qquad \theta \in (-\pi/2, \pi/2).$$

By intersecting the boundaries of two half-planes H_{θ_1} and H_{θ_2} and taking the limit $\theta_2 \to \theta_1$, we obtain a parametrization of the boundary of $W(L)$, which turns out to be the parabola

$$x = \kappa \left(\pi^2 + \frac{y^2}{a^2} \right).$$

The eigenvalues of the operator L itself may be obtained analogously and are given by

$$\lambda_k = \kappa \left[\left(\frac{a}{2\kappa} \right)^2 + k^2 \pi^2 \right], \quad k = 1, 2, \ldots.$$

Acknowledgment. The author would like to thank Michael Eiermann for many helpful discussions.

REFERENCES

[1] T. Braconnier and N. J. Higham, *Computing the field of values and pseudospectra using the Lanczos method with continuation*, BIT, 36 (1996), pp. 422–440.
[2] A. N. Brooks and T. J. R. Hughes, *A multidimensional upwind scheme with no crosswind diffusion*, in Finite Element Methods for Convection Dominated Flows, T. J. R. Hughes, ed., ASME, New York, 1979, pp. 199–259.
[3] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.
[4] P. J. Davis, *Interpolation and Approximation*, 2nd ed., Dover Books on Advanced Mathematics 14, Dover, New York, 1975.
[5] T. A. Driscoll, *A MATLAB toolbox for Schwarz-Christoffel mapping*, ACM Trans. Math. Software, 22 (1996), pp. 168–186.
[6] M. Eiermann, *Semiiterative Verfahren für nichtsymmetrische lineare Gleichungssysteme*, Habilitationsschrift, Universität Karlsruhe, Karlsruhe, Germany, 1989.
[7] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.
[8] M. Eiermann, *Field of Values and Iterative Methods*, 1996, manuscript.
[9] M. Eiermann, W. Niethammer, and R. S. Varga, *A study of semiiterative methods for nonsymmetric systems of linear equations*, Numer. Math., 47 (1985), pp. 505–533.
[10] B. Fischer and R. Freund, *Chebyshev polynomials are not always optimal*, J. Approx. Theory, 65 (1991), pp. 261–272.
[11] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen, *On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 179 (1999). pp. 185–202.
[12] P. M. Gresho and R. L. Lee, *Don't suppress the wiggles—they're telling you something*, in Finite Element Methods for Convection Dominated Flows, Vol. 34, AMD/ASME, New York, 1979, pp. 37–61.
[13] P. Henrici, *Applied and Computational Complex Analysis* I, Wiley Classics Library, Wiley-Interscience, New York, 1988.
[14] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[15] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[16] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.
[17] T. A. Manteuffel, *Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration*, Numer. Math., 31 (1978), pp. 183–208.
[18] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Appl. Math. Math. Comput., Chapman and Hall, London, 1996.

[19] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.
[20] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.
[21] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[22] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin, 1996.
[23] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[24] Z. Strakoš, V. Pták, and A. Greenbaum, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.
[25] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Press, London, 1992.
[26] G. Zhou, *How accurate is the streamline diffusion finite element method?*, Math. Comp., 66 (1997), pp. 31–44.