SIAM J. SCI. COMPUT.
Vol. 26, No. 6, pp. 1989–2009

GMRES CONVERGENCE ANALYSIS FOR A CONVECTION-DIFFUSION MODEL PROBLEM*

J. LIESEN† AND Z. STRAKOŠ‡

Abstract. When GMRES [Y. Saad and M. H. Schultz, *SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 856–869] is applied to streamline upwind Petrov–Galerkin (SUPG) discretized convection-diffusion problems, it typically exhibits an initial period of slow convergence followed by a faster decrease of the residual norm. Several approaches were made to understand this behavior. However, the existing analyses are solely based on the matrix of the discretized system and they do not take into account any influence of the right-hand side (determined by the boundary conditions and/or source term in the PDE). Therefore they cannot explain the length of the initial period of slow convergence which is right-hand side dependent.

We concentrate on a frequently used model problem with Dirichlet boundary conditions and with a constant velocity field parallel to one of the axes. Instead of the eigendecomposition of the system matrix, which is ill conditioned, we use its orthogonal transformation into a block-diagonal matrix with nonsymmetric tridiagonal Toeplitz blocks and offer an explanation of GMRES convergence. We show how the initial period of slow convergence is related to the boundary conditions and address the question why the convergence in the second stage accelerates.

Key words. convection-diffusion problem, streamline upwind Petrov–Galerkin discretization, GMRES, rate of convergence, ill-conditioned eigenvectors, nonnormality, tridiagonal Toeplitz matrices

AMS subject classifications. 65F10, 65F15, 65N22, 65N30

DOI. 10.1137/S1064827503430746

1. Introduction.

Krylov subspace methods such as GMRES [28] are typically used to solve very large linear algebraic systems. The goal is to find a sufficiently accurate approximate solution in a number of steps that is significantly less than the system dimension. Consequently, the convergence analysis of these methods must focus particularly on the early stages of the iteration, i.e., on the *transient* rather than the *asymptotic* behavior. This makes the methods' analysis a complicated nonlinear problem, which must be based not on a single number (such as the so-called asymptotic convergence factor) but on correspondingly more complex characteristics of the problem. If the system matrix is symmetric, then except for some special right-hand sides corresponding to some particular boundary conditions and/or outer forces, see, e.g., [3], the matrix eigenvalues answer practical questions about the convergence behavior of Krylov subspace methods. If the system matrix is nonsymmetric or, more generally, nonnormal, then the situation is much less clear.

In this paper we are interested in a particular example of the latter. We study linear algebraic systems $Ax = b$ arising from discretization of convection-diffusion

*Received by the editors July 3, 2003; accepted for publication (in revised form) September 7, 2004; published electronically June 16, 2005.

http://www.siam.org/journals/sisc/26-6/43074.html

†Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft. Part of this work was done while the author was a postdoctoral research assistant at the Center for Simulation of Advanced Rockets, University of Illinois at Urbana-Champaign, Urbana, IL 61801, and was supported by the U.S. Department of Energy under grant DOE LLNL B341494.

‡Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vod. věží 2, 182 07 Prague, and Technical University Liberec, Hálkova 6, 461 17 Liberec, Czech Republic (strakos@cs.cas.cz). The work of this author was supported by the GA CR under grant 201/02/0595.

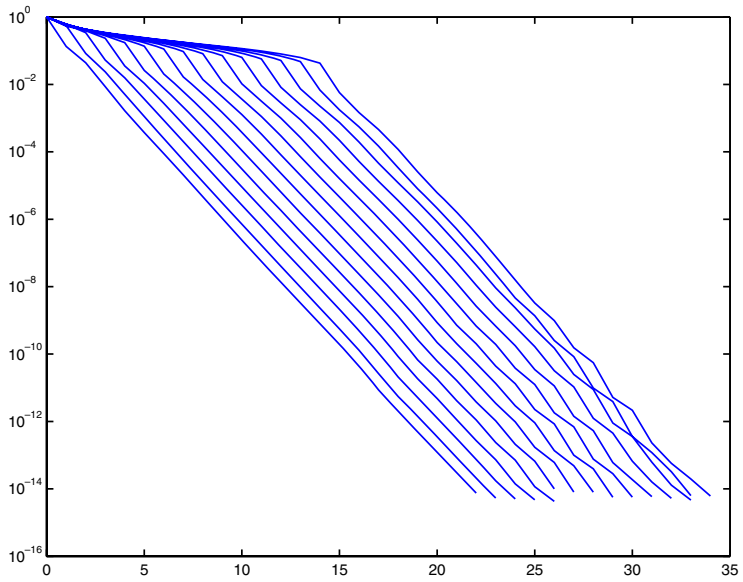

FIG. 1.1. *Relative GMRES residual norms for our SUPG discretized convection-diffusion model problem. Different behavior corresponds to the same discretized operator but to different boundary conditions.*

problems, and their solution with GMRES [28]. Starting from an initial guess x_0, this method computes the initial residual $r_0 = b - Ax_0$ and a sequence of iterates $x_1, x_2, \ldots$, so that the nth residual $r_n \equiv b - Ax_n$ satisfies

$$\|r_n\| \;=\; \|p_n(A)r_0\| \;=\; \min_{p\in\pi_n} \|p(A)r_0\|, \tag{1.1}$$

where π_n denotes the set of polynomials of degree at most n with value one at the origin.

It was shown in [16, 1] that GMRES can exhibit any nonincreasing convergence curve (of its residual norms) for a matrix having any eigenvalues. In these results the constructed matrix A and the right-hand side b are always related in a way which can hardly be interpreted in terms of any practical problem. Ernst [13] showed, however, an example of a convection-diffusion problem discretized via the streamline upwind Petrov–Galerkin (SUPG) method, for which the eigenvalues alone indeed give misleading information about convergence. He also observed, together with several other authors, see, e.g., [14], that GMRES applied to discretized convection-diffusion problems can exhibit an initial period of slow convergence followed by a faster decrease of the residual norm. Typical examples of such behavior are shown in Figure 1.1. Ernst 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 longest streamline of the velocity field. His conjecture and the example shown in Figure 1.1 clearly demonstrate the necessity of considering the particular right-hand side of the linear system (and hence the source term and the boundary conditions of the PDE) in the convergence analysis of GMRES.

The model problem studied in this paper is a convection-diffusion equation on the unit square with Dirichlet boundary conditions and with a constant velocity field parallel to one of the axes. We discretize the problem with the SUPG method based

on bilinear finite elements, an approach that has been frequently used before; see, e.g., [8, 9, 10, 13, 14]. Eigenvalues and eigenvectors of the discretized operator are known analytically. It seems therefore natural to exploit the eigenexpansion of the initial residual in the convergence analysis; see [14]. The eigenvector basis is, however, poorly conditioned; i.e., the system matrix is highly nonnormal. In such cases there is a reasonable doubt about using eigenvalues and eigenvectors in an analysis of convergence. This doubt was clearly formulated by Trefethen in [30, p. 384] (see also [26]). In fact, Trefethen concludes that if a matrix is far from normal, "there may be no good scientific reason for attempting to analyze the problem in terms of eigenvalues and eigenvectors." Our model problem represents an illuminative illustration of this viewpoint. Understanding the initial period of convergence is of primary importance, and the ill-conditioned eigendecomposition is not a proper tool for analyzing it.

Instead of the eigendecomposition we use an orthogonal transformation of the system matrix to a block-diagonal form with nonsymmetric tridiagonal Toeplitz blocks, a structure that was also employed in [5, 6, 8, 9]. Applying results from [19] we show how the initial period of slow convergence is related to the boundary conditions. We also address the question why the convergence in the second stage accelerates, although a quantitative understanding of this phenomenon still remains a subject of further work.

We have mentioned the necessity of considering the right-hand side b in our analysis. A careful reader might object that not b but the initial residual $r_0 = b - Ax_0$ enters the minimization condition (1.1). In fact, the initial residual, and therefore also the GMRES behavior, may depend significantly not only on b but also on the initial guess x_0. In order to clarify the practical role of x_0, we would like to stress the following simple, but sometimes overlooked point: Unless a nonzero initial guess x_0 is available that contains useful information about the solution x, for example an x_0 giving $\|r_0\| \leq \|b\|$, the choice $x_0 = 0$ should be preferred. Choosing a nonzero x_0 containing no useful information about x, e.g., choosing a random x_0, might create a completely "biased" r_0 with $\|r_0\| \gg \|b\|$. Such a choice potentially creates an illusion of a fast convergence to a high relative accuracy, measured by the relative residual norm. For examples see [22, relation (2.8), and the discussion of Figures 7.9 and 7.10]. Any such choice of x_0 is, however, useless. In this paper we always use $x_0 = 0$.

The paper is organized as follows. Section 2 specifies the model problem. Section 3 contains our analysis of the GMRES convergence behavior. Section 4 discusses the use of the eigenvalue decomposition for this purpose, and section 5 provides a concluding discussion.

Throughout the paper we assume exact arithmetic.

2. Specification of the model problem. In this paper we consider the following convection-diffusion model problem with Dirichlet boundary conditions:

$$-\nu\,\Delta u + w \cdot \nabla u \;=\; 0 \quad \text{in } \Omega = (0,1)\times(0,1), \qquad u = g \;\text{ on }\; \partial\Omega\,. \tag{2.1}$$

Here the scalar-valued function $u(\eta_1, \eta_2)$ represents the concentration of the transported quantity, $w = [w_1, w_2]^T$ the velocity field, and ν the scalar diffusion parameter. We are interested in the *convection-dominated* case; i.e., we assume $\|w\| \gg \nu$ in (2.1). For simplicity we use zero as the source term in the convection-diffusion equation. At the end of section 3.3 we explain how a nonzero source term would affect the findings of our investigation.

It is well known that in the convection-dominated case, the standard Galerkin finite element approximation to the solution of (2.1) suffers from nonphysical oscilla-

tions, see, e.g., [4, Fig. 2.1], [23, Fig. 8.3.1], and [8, Fig. 5.5]. Such oscillations can be avoided by using a stabilized Petrov–Galerkin finite element discretization instead.

In this paper we consider a particular instance of such discretization, namely, the SUPG method, introduced by Hughes and Brooks [17, 4]. More recent descriptions of the SUPG method can be found in [23, Chapter 8.3], [27, Chapter III.3.2], and [21, Chapter 5.5]. We specifically consider the SUPG method with bilinear finite elements on a regular grid with square elements of size $h \times h$,

$$h \;=\; (N+1)^{-1}\,,$$

where N represents the number of inner nodes along each side. In our case of dominating convection we assume that the *mesh Peclet number*

$$P_h \;\equiv\; \frac{h\|w\|}{2\nu} \tag{2.2}$$

is greater than one. The same model problem has been used and studied in many publications, see in particular [8, 9, 10, 13, 14]. Other finite element discretizations have also been considered; cf. [2] for a study concerning the piecewise linear case.

The stabilization in the SUPG method can be expressed as an additional diffusion term with the diffusivity tensor given by $\hat{\delta} w w^T$, which acts only in the direction of the flow. Here $\hat{\delta}$ denotes the *stabilization parameter*. If $P_h > 1$, then $\hat{\delta}$ is typically chosen as

$$\hat{\delta} \;=\; \frac{\delta h}{\|w\|}\,, \tag{2.3}$$

where $\delta > 0$ is a *tuning parameter*.

In case of piecewise linear finite elements for a one-dimensional constant coefficient problem, the choice

$$\delta_0 \;\equiv\; \frac{1}{2}\left(\coth(P_h) - \frac{1}{P_h}\right)\,, \quad \text{i.e.,} \quad \hat{\delta}_0 \;\equiv\; \frac{h}{2\|w\|}\left(\coth(P_h) - \frac{1}{P_h}\right)\,, \tag{2.4}$$

yields the exact solution at the node points; see, e.g., [17], [4, Section 2.4], and [27, Chapter I.2.1.3]. A similar optimal choice of δ for two or more dimensional problems is unknown; see [27, Remark III.3.34] for an informative discussion. Hence some authors use $\delta = \delta_0$ (cf. [13, equation (2.8)]) or $\delta \approx \delta_0$ (cf. [14, pp. 186–187]) also for the two-dimensional problem (2.1). By definition,

$$\coth(P_h) \;=\; \frac{e^{P_h} + e^{-P_h}}{e^{P_h} - e^{-P_h}}\,,$$

and hence the following simplified value,

$$\delta_* \;\equiv\; \frac{1}{2}\left(1 - \frac{1}{P_h}\right) \;<\; \delta_0\,, \tag{2.5}$$

is close to δ_0 even for moderate values of P_h. For example, if $P_h = 5$, then $\delta_* = 0.4$ and $\delta_0 \approx 0.40005$. The parameter δ_* is defined in [14, p. 187], where the authors note that $\delta_* \nearrow \delta_0$ as $P_h \to \infty$. Obviously this convergence is very rapid. In [9] the authors study the effects of the tuning parameter δ on the behavior of the solution with respect to the nonphysical oscillations. Their analysis gives a theoretical justification for the choice $\delta = \delta_*$.

Supported by [9] and for the sake of clarity of our exposition, we limit the GMRES convergence analysis in our paper to the discretized problems with the value $\delta = \delta_*$, giving

$$\hat{\delta}_* \;\equiv\; \frac{\delta_* h}{\|w\|}\,.$$

For different values of $\hat{\delta}$ the problem can be analyzed analogously.

2.1. The discretized operator. The coefficient matrix of the linear algebraic system resulting from the SUPG discretization of (2.1) described above can be written in the form

$$A \;=\; \nu A_d + A_c + \hat{\delta} A_s, \tag{2.6}$$

where $A_d = \langle \nabla\phi_j, \nabla\phi_i\rangle$, $A_c = \langle w\cdot\nabla\phi_j, \phi_i\rangle$, and $A_s = \langle w\cdot\nabla\phi_j, w\cdot\nabla\phi_i\rangle$ represent the diffusion, convection, and stabilization term, respectively. Here $\phi_1, \ldots, \phi_{N^2}$ denote the piecewise bilinear nodal basis functions, and $\langle\cdot,\cdot\rangle$ denotes the L^2 inner product on Ω.

In the following we consider the special case of the *vertical wind*

$$w \;=\; [0,1]^T\,.$$

Then both A_d and A_s are symmetric positive definite while A_c is skew-symmetric; see [14, p. 182]. Writing the coefficient matrix in the form

$$A \;=\; \langle (\nu I + \hat{\delta} w w^T)\nabla\phi_j, \nabla\phi_i\rangle \;+\; \langle w\cdot\nabla\phi_j, \phi_i\rangle\,, \tag{2.7}$$

the "effective" diffusivity tensor is given by

$$\nu I + \hat{\delta} w w^T \;=\; \begin{pmatrix} \nu & 0 \\ 0 & \nu + \hat{\delta} \end{pmatrix}, \qquad \hat{\delta} \;=\; \delta h\,.$$

Moreover, the constituent matrix stencil for A,

$$\begin{array}{ccccc} m_4 & & m_3 & & m_4 \\ & \nwarrow & \uparrow & \nearrow & \\ m_2 & \leftarrow & m_1 & \rightarrow & m_2 \\ & \swarrow & \downarrow & \searrow & \\ m_6 & & m_5 & & m_6 \end{array} \tag{2.8}$$

has numerical values

$$\begin{array}{ccccc} -\frac{\nu}{3} + \frac{h}{12}(1-2\delta) & & -\frac{\nu}{3} + \frac{h}{3}(1-2\delta) & & -\frac{\nu}{3} + \frac{h}{12}(1-2\delta) \\ & \nwarrow & \uparrow & \nearrow & \\ -\frac{\nu}{3} + \frac{\delta h}{3} & \leftarrow & \frac{8}{3}\nu + \frac{4}{3}\delta h & \rightarrow & -\frac{\nu}{3} + \frac{\delta h}{3} \\ & \swarrow & \downarrow & \searrow & \\ -\frac{\nu}{3} - \frac{h}{12}(1+2\delta) & & -\frac{\nu}{3} - \frac{h}{3}(1+2\delta) & & -\frac{\nu}{3} - \frac{h}{12}(1+2\delta) \end{array} \tag{2.9}$$

(see [14, formulas (12)–(14)] for the general form of the matrix stencil for A in case of a constant wind $w = [w_1, w_2]^T$).

Using the *vertical* line ordering of the unknowns, i.e., the ordering parallel to the direction of the wind, the N^2 by N^2 system matrix A_V takes the form

$$A_V \;=\; A_V(h,\nu,\delta) \;=\; \nu K \otimes M + M \otimes ((\nu + \delta h)K + C); \tag{2.10}$$

see, e.g., [5, Section 1.1] and [13, pp. 1081 and 1089]. Here

$$\begin{aligned} M &= \frac{h}{6}\,\mathrm{tridiag}(1,4,1)\,, \\ K &= \frac{1}{h}\,\mathrm{tridiag}(-1,2,-1)\,, \\ C &= \frac{1}{2}\,\mathrm{tridiag}(-1,0,1) \end{aligned} \tag{2.11}$$

are the N by N mass, stiffness, and gradient matrices of the one-dimensional constant coefficient convection-diffusion equation discretized on a uniform mesh using linear elements.

Next note that the eigenvalues of an N by N symmetric tridiagonal Toeplitz matrix $\mathrm{tridiag}(t_2, t_1, t_2)$ are given by $t_1 + t_2\omega_j$, where

$$\omega_j \;=\; 2\cos(jh\pi), \quad j = 1, \ldots, N\,. \tag{2.12}$$

Furthermore, the corresponding normalized eigenvectors are given by

$$u_j \;=\; (2h)^{1/2}\,[\,\sin(jh\pi), \ldots, \sin(Njh\pi)\,]^T, \quad j = 1, \ldots, N; \tag{2.13}$$

see, e.g., [29, pp. 113–115]. Consequently, the matrices M and K in (2.11) are simultaneously diagonalizable by the symmetric orthonormal matrix $U = [u_1, \ldots, u_N]$. The block diagonalization of the matrix A_V in (2.10) by the discrete sine transform then gives

$$(U \otimes I)A_V(U \otimes I) \;=\; \nu(UKU) \otimes M + (UMU) \otimes ((\nu + \delta h)K + C) \;\equiv\; T\,. \tag{2.14}$$

Elementary algebra shows that T is a block-diagonal matrix consisting of N *nonsymmetric* tridiagonal Toeplitz blocks T_j, each of the size N by N,

$$T = \mathrm{diag}(T_{1:N})\,, \tag{2.15}$$

where

$$T_j = \mathrm{tridiag}(\gamma_j, \lambda_j, \mu_j)\,, \quad j = 1, \ldots, N\,, \tag{2.16}$$

$$\lambda_j = m_1 + m_2\,\omega_j\,, \quad \mu_j = m_3 + m_4\,\omega_j\,, \quad \gamma_j = m_5 + m_6\,\omega_j \tag{2.17}$$

(cf. (2.8)–(2.9) for the definition of $m_1, \ldots, m_6$).

For completeness we mention that instead of the vertical line ordering used in (2.10), some authors have considered the *horizontal* line ordering; see, e.g., [10, 14]. In this case the resulting coefficient matrix A_H takes the form

$$A_H \;=\; A_H(h, \nu, \delta) \;=\; \nu M \otimes K \;+\; ((\nu + \delta h)K + C) \otimes M\,. \tag{2.18}$$

The matrix A_H is of the form $\mathrm{tridiag}(M_3, M_1, M_2)$, with N by N symmetric tridiagonal Toeplitz blocks given by

$$\begin{aligned} M_1 &= \mathrm{tridiag}(m_2, m_1, m_2)\,, \\ M_2 &= \mathrm{tridiag}(m_4, m_3, m_4)\,, \\ M_3 &= \mathrm{tridiag}(m_6, m_5, m_6)\,. \end{aligned} \tag{2.19}$$

Of course, the two approaches are equivalent. The orthogonal transformation

$$(I \otimes U)\, A_H\, (I \otimes U) \;=\; \mathrm{tridiag}(D_\gamma, D_\lambda, D_\mu)\,,$$

where $D_\lambda = \mathrm{diag}(\lambda_{1:N})$, $D_\mu = \mathrm{diag}(\mu_{1:N})$, $D_\gamma = \mathrm{diag}(\gamma_{1:N})$, and a permutation of the unknowns yields the matrix T in (2.14). Using the symmetric permutation matrix

$$P \;\equiv\; [I \otimes e_1, \ldots, I \otimes e_N]\,, \qquad P^2 = I\,,$$

which transforms the horizontal line ordering into the vertical line ordering and vice versa, the equivalence of the discretized systems $A_H x_H = b_H$ and $A_V x_V = b_V$ can be easily seen from the relation

$$A_H x_H \;=\; b_H \qquad \Longleftrightarrow \qquad \underbrace{(PA_HP)}_{=A_V}\,\underbrace{(Px_H)}_{=x_V} \;=\; \underbrace{Pb_H}_{=b_V}\,.$$

2.2. Structure of the right-hand sides. We now discuss the structure of the right-hand side vectors b_V in the linear system corresponding to the matrix A_V in (2.10). Due to the zero source term in (2.1) the entries of b_V are *completely determined by the Dirichlet boundary condition* $u = g$ on $\partial\Omega$.

We partition the vector b_V of the length N^2 into N blocks of the length N each,

$$b_V \;=\; [b^{(1)T}, \ldots, b^{(N)T}]^T\,, \tag{2.20}$$

where the jth block corresponds to the jth *vertical* layer of the mesh. We then form the N by N matrix $B_V \equiv [b^{(1)}, \ldots, b^{(N)}]$ which has the following general nonzero structure:

$$B_V \;=\; \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,N-1} & b_{1,N} \\ b_{2,1} & 0 & \cdots & 0 & b_{2,N} \\ \vdots & \vdots & & \vdots & \vdots \\ b_{N-1,1} & 0 & \cdots & 0 & b_{N-1,N} \\ b_{N,1} & b_{N,2} & \cdots & b_{N,N-1} & b_{N,N} \end{bmatrix}\,. \tag{2.21}$$

The entries of B_V can easily be computed using (2.8)–(2.9). In the experiments presented in this paper we use the following examples.

Example 2.1. Following the set of problems introduced by Raithby [24], the authors of [9, 10, 14] use boundary conditions that are discontinuous at the inflow boundary,

$$u(\eta_1, 0) \;=\; u(1, \eta_2) \;=\; 1 \qquad \text{for} \quad 1/2 < \eta_1 \le 1 \;\text{ and }\; 0 \le \eta_2 < 1, \tag{2.22}$$

$$u(\eta_1, \eta_2) \;=\; 0 \qquad \text{elsewhere on } \partial\Omega. \tag{2.23}$$

Hence the first row of B_V has nonzero entries given by

$$\begin{aligned} b_{1,\lfloor N/2\rfloor} = -m_6 \;&=\; \frac{\nu}{3} + \frac{h}{12}(1+2\delta)\,, \\ b_{1,\lfloor N/2\rfloor+1} = -(m_6+m_5) \;&=\; \frac{2}{3}\nu + \frac{5}{12}h(1+2\delta)\,, \\ b_{1,\lfloor N/2\rfloor+j} = -(2m_6+m_5) \;&=\; \nu + \frac{h}{2}(1+2\delta), \quad j = 2, \ldots, N - (\lfloor N/2\rfloor + 1)\,, \\ b_{1,N} = -(2m_6+m_5+m_2+m_4) \;&=\; \frac{5}{3}\nu + \frac{5}{12}h(1+2\delta)\,, \end{aligned}$$

while

$$b_{j,N} = -(m_6 + m_2 + m_4) \;=\; \nu\,, \quad j = 2, \ldots, N-1,$$
$$b_{N,N} = -(m_6 + m_2) \;=\; \frac{2}{3}\nu + \frac{1}{12}h(1-2\delta)$$

are the remaining nonzero entries of B_V. ∎

Example 2.2. We also consider nonzero boundary conditions only on (a part of) the right side boundary of the unit square. Specifically, we equally divide the y-direction of the unit square into $N+1$ parts according to the N internal nodes of the mesh. This gives rise to the following N boundary conditions:

$$u(1, \eta_2) = 1 \qquad \text{for} \quad k/(N+1) \le \eta_2 < 1, \quad k = 0, 1, \ldots, N-1, \tag{2.24}$$

and $u(\eta_1, \eta_2) = 0$ elsewhere on $\partial\Omega$. The resulting matrices B_V have nonzero entries only in their last columns. For example, in the case $k = 7$ the nonzero entries of B_V are given by

$$b_{j,N} = 0, \quad j = 1, \ldots, 5,$$
$$b_{6,N} = \frac{1}{3}\nu - \frac{1}{12}h(1-2\delta), \quad b_{N,7} \;=\; \frac{2}{3}\nu - \frac{1}{12}h(1+2\delta),$$
$$b_{j,N} = \nu, \quad j = 8, \ldots, N-1,$$
$$b_{N,N} = \frac{2}{3}\nu + \frac{1}{12}h(1-2\delta).$$

For other values of k the entries of B_V can be computed analogously. ∎

3. GMRES convergence analysis. As indicated in the introduction, when GMRES is applied to linear systems resulting from the SUPG discretization of (2.1), it typically exhibits an initial period of slow convergence followed by a faster decrease of the residual norms. This behavior is illustrated in Figure 1.1, which shows the relative GMRES residual norms, $\|r_n\|/\|b_V\|$ (here, and elsewhere, $x_0 = 0$), for $w = [0,1]^T$, the *fixed* discretized operator $A_V = A_V(1/16, 0.01, 0.34)$, cf. (2.10), and the 15 *different* right-hand side vectors b_V resulting from the boundary conditions (2.24). The kth boundary condition corresponds to the initial period of slow convergence lasting for $N-1-k$ steps. Our first goal in this section is to quantitatively analyze how this happens. We then address the question why the convergence speed of GMRES accelerates after the initial phase. As explained above, we restrict our discussion to $w = [0,1]^T$, and to the choice $\delta = \delta_*$.

3.1. Derivation of the basic lower bound. Consider a linear system

$$A_V x_V \;=\; b_V \tag{3.1}$$

that corresponds to the vertical line ordering of the unknowns. An orthogonal transformation by $(U \otimes I)$, cf. (2.14), yields

$$(U \otimes I)A_V(U \otimes I)\,[(U \otimes I)x_V] \;\equiv\; T\,[(U \otimes I)x_V] \;=\; (U \otimes I)b_V \;\equiv\; \hat{b}\,. \tag{3.2}$$

We partition the vector of unknowns x_V similarly as the right-hand side vector b_V in section 2.2, and denote the corresponding N by N matrix by X_V. Then the transformed system (3.2) decomposes into N linear systems of the size N by N,

$$T_j\,[X_V u_j] \;=\; \hat{b}^{(j)}\,, \qquad \hat{b}^{(j)} \;\equiv\; B_V u_j\,, \quad j = 1, \ldots, N. \tag{3.3}$$

Since the transformation of the original system (3.1) into the decomposed block-diagonal system represented by (3.3) is orthogonal, the GMRES residual norms for these two systems (and the corresponding initial guess) coincide. In particular, for $x_0 = 0$ we obtain the following lower bound on the GMRES residual norms when the algorithm is applied to the system (3.1):

$$\|r_n\|^2 = \min_{p\in\pi_n} \| p(A_V)\, b_V \|^2 \tag{3.4}$$

$$= \min_{p\in\pi_n} \| p(T)\, \hat{b} \|^2 \tag{3.5}$$

$$= \min_{p\in\pi_n} \sum_{j=1}^{N} \| p(T_j)\, \hat{b}^{(j)} \|^2 \tag{3.6}$$

$$\geq \sum_{j=1}^{N} \min_{p\in\pi_n} \| p(T_j)\, \hat{b}^{(j)} \|^2 . \tag{3.7}$$

In the step from (3.4) to (3.5) we exploit orthogonality of the transformation from (3.1) to (3.2). The next step from (3.5) to (3.6) reflects the decomposition (3.3). Note that the linear systems in (3.3) are for different values of j independent of each other, and hence can in principle be solved independently. This is not true, however, when GMRES is applied to (3.1). Then the individual approximations are *coupled together* by the global minimization problem (3.4)–(3.6). Finally, (3.7) bounds the squared GMRES residual norm from below by the sum of the squared GMRES residual norms when the algorithm is applied *independently* to each of the systems (3.3). Since each of these systems is of the order N, the lower bound (3.7) is equal to zero (and hence useless) for $n = N$, possibly even earlier. However, when there is at least one system (3.3) for which GMRES shows an initial period of slow convergence, the lower bound (3.7) shows that GMRES for the original system (3.1) also initially converges slowly for at least as many steps. This is, in a nutshell, the tool needed to understand the initial phase of convergence of GMRES applied to our SUPG discretized convection-diffusion model problem.

Each of the matrices T_j, $j = 1, \ldots, N$, is a nonsymmetric tridiagonal Toeplitz matrix. In order to evaluate (3.7) we have to analyze the behavior of GMRES for this class of matrices. This represents a peculiar problem on its own. Physically it can be interpreted, e.g., as analyzing the GMRES behavior for the discretized one-dimensional convection-diffusion problem with a constant wind, cf. [7, 10]. Below we will use results from our paper [19], which is devoted to this subject. Their application requires more details about the numerical values of the entries in the matrices T_j.

3.2. The entries of the matrices T_j. Each of the matrices T_j is of the form tridiag$(\gamma_j, \lambda_j, \mu_j)$, cf. (2.16)–(2.17). The numerical values of the entries can be found from the stencils (2.8)–(2.9). For simplicity, we rewrite these entries as

$$\begin{aligned} 3\lambda_j &= 2\delta h \left(2 + \frac{\omega_j}{2}\right) + 2\nu \left(4 - \frac{\omega_j}{2}\right), \\ -3\mu_j &= \delta h \left(2 + \frac{\omega_j}{2}\right) + \nu(1 + \omega_j) - \frac{h}{2} \left(2 + \frac{\omega_j}{2}\right), \\ -3\gamma_j &= \delta h \left(2 + \frac{\omega_j}{2}\right) + \nu(1 + \omega_j) + \frac{h}{2} \left(2 + \frac{\omega_j}{2}\right). \end{aligned} \tag{3.8}$$

We first analyze their signs.

LEMMA 3.1. *Let, as above,* $w = [0,1]^T$ *(*$\|w\| = 1$*). If the mesh Peclet number* (2.2) *satisfies* $P_h > 1$*, then for all* $j = 1, \ldots, N$ *the values* λ_j *and* γ_j *defined in* (3.8) *satisfy*

$$\lambda_j \ > \ 0 \ > \ \gamma_j. \tag{3.9}$$

Furthermore, for all $j = 1, \ldots, N$ *the value of* μ_j *defined in* (3.8) *satisfies*

$$\mathrm{sign}(\mu_j) \ = \ \mathrm{sign}(f(j) - \delta), \quad \textit{where} \quad f(j) \ \equiv \ \delta_* \ + \ \frac{1 - \omega_j/2}{P_h(4 + \omega_j)}\,, \tag{3.10}$$

so that μ_j *is negative, zero, or positive, if* δ *is larger than, equal to, or smaller than* $f(j)$*, respectively. In particular, if* $\delta = \delta_*$*, then* $\mu_j > 0$ *for all* $j = 1, \ldots, N$*.*

Proof. Considering the relations (3.8) we first note that since $-2 < \omega_j < 2$, see (2.12), we always have $\lambda_j > 0$. Next, if $P_h > 1$, then $h/2 - \nu > 0$, so that

$$-3\gamma_j \ > \ \delta h - \nu + \frac{h}{2} \ > \ \delta h \ > \ 0 \quad \Rightarrow \quad \gamma_j < 0.$$

An elementary computation yields

$$\frac{3\mu_j}{h\,(\,2 + \omega_j/2\,)} \ = \ f(j) - \delta,$$

where $f(j)$ is defined as in (3.10). Obviously, the left-hand side of this equality has the same sign as μ_j, which proves (3.10). If $\delta = \delta_*$, then

$$\mathrm{sign}(\mu_j) \ = \ \mathrm{sign}\left(\frac{1 - \omega_j/2}{P_h(4 + \omega_j)}\right),$$

which shows that in this case $\mu_j > 0$ for all $j = 1, \ldots, N$. ∎

We will next analyze the moduli of the ratios of the values λ_j, μ_j, and γ_j, $j = 1, \ldots, N$. Note that if $\|w\| = 1$, then $\delta_* = (h - 2\nu)/(2h)$. Thus for $\delta = \delta_*$ the relations (3.8) are equivalent to

$$\begin{aligned} 3\lambda_j &= h\left(2 + \frac{\omega_j}{2}\right) + 4\nu\left(1 - \frac{\omega_j}{2}\right), \\ 3\mu_j &= \nu\left(1 - \frac{\omega_j}{2}\right), \\ -3\gamma_j &= h\left(2 + \frac{\omega_j}{2}\right) - \nu\left(1 - \frac{\omega_j}{2}\right). \end{aligned} \tag{3.11}$$

Straightforward manipulations give the following result.

LEMMA 3.2. *Let, as above,* $w = [0,1]^T$ *(*$\|w\| = 1$*), and* $\delta = \delta_*$*. Then*

$$\frac{|\lambda_j|}{|\gamma_j|} = 1 + 5\,\left(2P_h\,\frac{4 + \omega_j}{2 - \omega_j} \ - \ 1\right)^{-1}, \tag{3.12}$$

$$\frac{|\mu_j|}{|\gamma_j|} = \left(2P_h\,\frac{4 + \omega_j}{2 - \omega_j} \ - \ 1\right)^{-1}, \qquad j = 1, \ldots, N\,. \tag{3.13}$$

Clearly, when $P_h \gg 1$, then for each $j = 1, \ldots, N$,

$$\frac{|\lambda_j|}{|\gamma_j|} \ \approx \ 1 \ \gg \ \frac{|\mu_j|}{|\gamma_j|}\,. \tag{3.14}$$

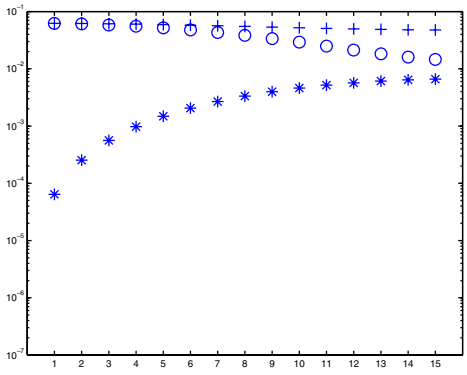

FIG. 3.1. λ_j (+), μ_j (∗), and $|\gamma_j|$ (o) for $j = 1, \ldots, 15$ and $h = 1/16$, $\nu = 0.01$, $\delta = \delta_* = 0.34$.

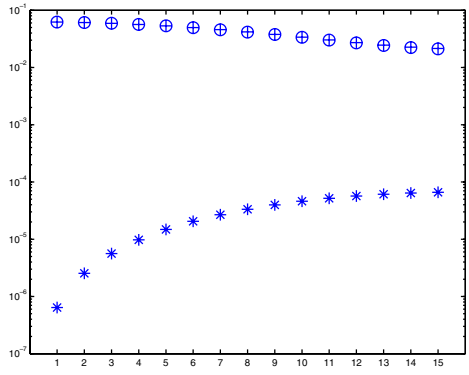

FIG. 3.2. λ_j (+), μ_j (∗), and $|\gamma_j|$ (o) for $j = 1, \ldots, 15$ and $h = 1/16$, $\nu = 0.0001$, $\delta = \delta_* = 0.4984$.

For a moderate P_h the ratios (3.12) and (3.13) depend more significantly on the index j. When $jh\pi \ll 1$, the expansion of the cosine function gives

$$\frac{4+\omega_j}{2-\omega_j} \;=\; \frac{6}{(jh\pi)^2} - 1 + \mathcal{O}(\,(jh\pi)^4\,)\,.$$

Hence for small indices j, (3.14) holds even for a moderate P_h. Since λ_j, γ_j, and μ_j depend linearly on δ, these considerations hold not only for $\delta = \delta_*$ but apply also whenever $\delta \approx \delta_*$.

Experiment 3.3. In Figures 3.1 and 3.2 we show typical examples of the magnitudes of λ_j, γ_j, and μ_j. For Figure 3.1 we use $h = 1/16$, $\nu = 0.01$, and $\delta = \delta_* = 0.34$, which are the same parameters as in [14, p. 186]. These yield a moderate mesh Peclet number, $P_h = 3.125$, so that (3.14) holds only for smaller indices j. To show results for a larger mesh Peclet number we choose $h = 1/16$, $\nu = 0.0001$, and $\delta = \delta_* = 0.4984$ for Figure 3.2. Here $P_h = 312.5$ so that (3.14) holds for all $j = 1, \ldots, 15$. ∎

3.3. Analysis of the initial phase. Having analyzed the entries in the matrices T_j we now come to our explanation of the initial phase of slow convergence. We first present two numerical experiments illustrating (3.5)–(3.7).

Experiment 3.4. Using the parameter values $h = 1/16$, $\nu = 0.01$, and $\delta = \delta_* = 0.34$, we set up a linear system of the form (3.2). For the right-hand side we use the boundary conditions (2.24) with $k = 0$. GMRES with the initial guess $x_0 = 0$ then produces the squared residual norms $\|r_n\|^2$ plotted by the solid line in Figure 3.3. We also apply GMRES independently to each of the $N = 15$ linear systems (3.3), and plot the resulting squared residual norms by the dashed lines in Figure 3.3. The labels on these dashed lines correspond to the indices $j = 1, \ldots, 15$ of the individual systems (3.3). The plus signs show the sums of the individual dashed curves, i.e., the lower bound (3.7). Figure 3.4 shows a three-dimensional plot of the computed solution. ∎

Experiment 3.5. We use the same parameters as in Experiment 3.4, but for the computation of the right-hand side we here use (2.24) with $k = 7$. Figures 3.5 and 3.6 show the results analogous to Figures 3.3 and 3.4. ∎

Our first observation in both experiments is that during the initial phase of slow convergence the lower bound (3.7) is very tight. Furthermore, as in Figure 1.1, the

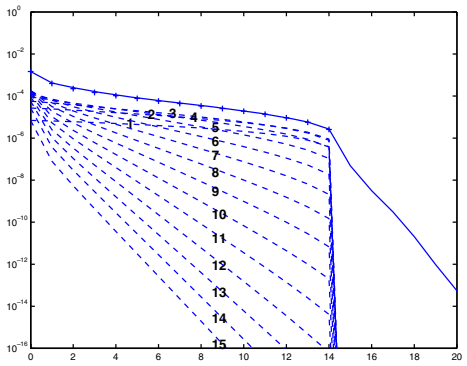

FIG. 3.3. *Squared GMRES residual norms for* (3.2) *with right-hand side from* (2.24) *with* $k = 0$ *(solid) and for each system* (3.3) *individually (dashed), and the lower bound* (3.7) *(+). System parameters are* $h = 1/16$, $\nu = 0.01$, $\delta = \delta_* = 0.34$.

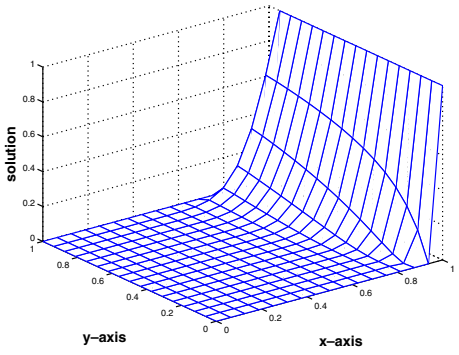

FIG. 3.4. *The solution corresponding to Experiment* 3.4, *i.e., the boundary conditions* (2.24) *with* $k = 0$.

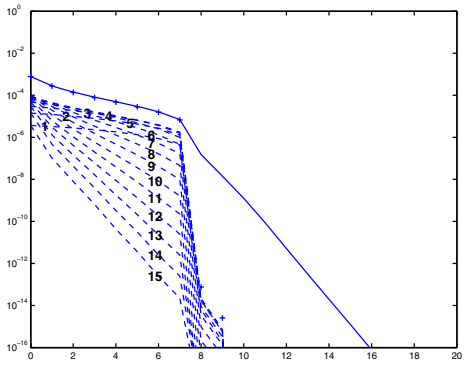

FIG. 3.5. *Results analogous to Figure* 3.3 *but with* $k = 7$ *in* (2.24).

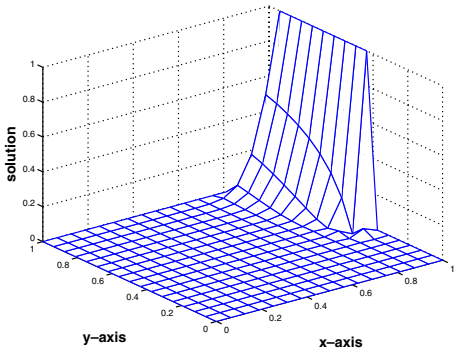

FIG. 3.6. *The solution corresponding to Experiment* 3.5, *i.e., the boundary conditions* (2.24) *with* $k = 7$.

initial phase in Figures 3.3 and 3.5 lasts $N-k-1$ steps (14 steps for $k = 0$ and 7 steps for $k = 7$). The parameters h, ν, and δ chosen in both Experiments 3.4 and 3.5 yield the matrices T_j, $j = 1, \ldots, N$, with the absolute values of the entries γ_j, λ_j, μ_j shown in Figure 3.1. Apparently, the slow initial convergence occurs only for the individual systems (3.3) with a small index j, when (3.14) holds. Using our results in [19], this observation can be understood and quantified.

Skipping details, our results in [19] about the convergence of GMRES for tridiagonal Toeplitz matrices can be summarized in the following way. Suppose that GMRES with $x_0 = 0$ is applied to a system of the form (3.3), and denote

$$\hat{b}^{(j)} \;=\; B_V u_j \;\equiv\; [\rho_1^{(j)}, \ldots, \rho_N^{(j)}]^T\,, \qquad \tau_j \;\equiv\; \frac{\lambda_j}{\gamma_j}\,, \quad \text{and} \quad \zeta_j \;\equiv\; \frac{\mu_j}{\gamma_j}\,. \tag{3.15}$$

Now suppose that $\rho_l^{(j)}$ is the first nonzero component of $\hat{b}^{(j)}$, and that GMRES applied to T_j and $\hat{b}^{(j)}$ does not terminate in the first $N - l$ steps (we exclude some very

peculiar circumstances under which $\hat{b}^{(j)}$ has less than $N-l$ nonzero components in the directions of the individual eigenvectors of T_j, and GMRES therefore terminates sooner). Then for $n = 0, 1, \ldots, N-l$ the GMRES residual norms for T_j and $\hat{b}^{(j)}$ satisfy, see [19, Theorem 3.2],

$$\|\hat{r}_n^{(j)}\| \;=\; \min_{p\in\pi_n} \|p(T_j)\,\hat{b}^{(j)}\| \tag{3.16}$$

$$\;=\; \left\| [1, -\tau_j, \ldots, (-\tau_j)^n\,] \left[\hat{b}^{(j)}, (S^T+\zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T+\zeta_j S)^n\,\hat{b}^{(j)}\right]^+ \right\|^{-1} \tag{3.17}$$

$$\;\geq\; \left(\sum_{m=0}^{n} |\tau_j|^{2m}\right)^{-\frac{1}{2}} \sigma_{\min}\left(\left[\hat{b}^{(j)}, (S^T+\zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T+\zeta_j S)^n\,\hat{b}^{(j)}\right] \right). \tag{3.18}$$

Here X^+ denotes the Moore–Penrose generalized inverse and $\sigma_{\min}(X)$ the smallest singular value of the matrix X, and $S = [0, e_1, \ldots, e_{N-1}]$ denotes the standard upward shift matrix.

For the iteration step $n = N-l$, the expression (3.17) can be simplified. Let

$$\left[\hat{b}^{(j)}, (S^T+\zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T+\zeta_j S)^{N-l}\,\hat{b}^{(j)}\right]^T \;\equiv\; [O, R_j] + \zeta_j\, P_j, \tag{3.19}$$

where O denotes the $N-l+1$ by $l-1$ zero matrix,

$$R_j \;\equiv\; \begin{bmatrix} \rho_l^{(j)} & \rho_{l+1}^{(j)} & \cdots & \rho_N^{(j)} \\ & \rho_l^{(j)} & \cdots & \rho_{N-1}^{(j)} \\ & & \ddots & \vdots \\ & & & \rho_l^{(j)} \end{bmatrix},$$

and the columns of the matrix P_j^T are given by

$$\zeta_j^{-1}\left\{(S^T+\zeta_j S)^m - (S^T)^m\right\}\hat{b}^{(j)}, \quad m = 0, 1, \ldots, N-l\,.$$

As shown in [19, Section 3.2], the norm of the mth column of P_j^T is bounded by $m\,\|\hat{b}^{(j)}\|\,(\,1 + \mathcal{O}(|\zeta_j|m)\,)$. Since we assume that $\rho_l^{(j)} \neq 0$, the square matrix R_j is nonsingular. Furthermore, R_j does not depend on ζ_j. Consequently, for $|\zeta_j|$ small enough, $|\zeta_j|\|R_j^{-1}P_j\| < 1$ (for details see [19]). Assuming that $|\zeta_j|\|R_j^{-1}P_j\| < 1$ holds, [19, Theorems 3.3 and 2.1] give

$$\|\hat{r}_{N-l}^{(j)}\| = \min_{p\in\pi_{N-l}} \|p(T_j)\,\hat{b}^{(j)}\| \tag{3.20}$$

$$= \left\| \left([O, I] + \zeta_j R_j^{-1}P_j\right)^+ R_j^{-1}\left[1, -\tau_j, \ldots, (-\tau_j)^{N-l}\,\right]^T \right\|^{-1} \tag{3.21}$$

$$\geq \left(\,1 - |\zeta_j|\|R_j^{-1}P_j\|\,\right)\left(\sum_{m=0}^{N-l} |\tau_j|^{2m}\right)^{-\frac{1}{2}} \sigma_{min}(R_j)\,. \tag{3.22}$$

Moreover, independently of the value $|\zeta_j|\|R_j^{-1}P_j\|$,

$$\|\hat{r}_{N-l}^{(j)}\| \;\leq\; \left(\,1 + |\zeta_j|\|R_j^{-1}P_j\|\,\right)(N-l+1)^{\frac{1}{2}}\,\|\hat{b}^{(j)}\|\left(\sum_{m=0}^{N-l} |\tau_j|^{2m}\right)^{-\frac{1}{2}}. \tag{3.23}$$

TABLE 3.1

Numerical values of the quantities in the bounds (3.22) *and* (3.23) *corresponding to Experiment* 3.4. *The stars* (∗) *indicate that* $|\zeta_j|\,\|R_j^{-1}P_j\| \geq 1$, *so that* (3.22) *is not applicable.*

j	$\lvert\tau_j\rvert$	$\lvert\zeta_j\rvert$	$\lvert\zeta_j\rvert\,\|R_j^{-1}P_j\|$	$\frac{1}{\sqrt{\sum \lvert\tau_j\rvert^{2m}}}$	$\sigma_{min}(R_j)$	$\frac{(3.22)}{\|\hat{b}^{(j)}\|}$	$\frac{\|\hat{r}_{N-1}^{(j)}\|}{\|\hat{b}^{(j)}\|}$	$\frac{(3.23)}{\|\hat{b}^{(j)}\|}$
1	1.0052	0.0010	0.0247	0.2489	0.0003	0.0318	0.2364	0.9879
2	1.0209	0.0042	0.0981	0.2216	0.0007	0.0262	0.1828	0.9424
3	1.0481	0.0096	0.2180	0.1785	0.0010	0.0182	0.1183	0.8420
4	1.0881	0.0176	0.3812	0.1260	0.0013	0.0102	0.0641	0.6740
5	1.1431	0.0286	0.5846	0.0752	0.0015	0.0041	0.0290	0.4613
6	1.2162	0.0432	0.8312	0.0368	0.0016	0.0008	0.0110	0.2609
7	1.3116	0.0623	1.1505	0.0145	0.0017	∗	0.0034	0.1208
8	1.4348	0.0870	1.6373	0.0046	0.0018	∗	0.0009	0.0468
9	1.5925	0.1185	2.4596	0.0012	0.0017	∗	0.0002	0.0155
10	1.7923	0.1585	3.8905	0.0002	0.0016	∗	3.5e-5	0.0045
11	2.0409	0.2082	6.4713	4.0e-5	0.0015	∗	5.7e-6	0.0012
12	2.3392	0.2678	11.2601	6.2e-6	0.0013	∗	8.5e-7	0.0003
13	2.6735	0.3347	19.9683	9.7e-7	0.0010	∗	1.3e-7	7.9e-5
14	3.0033	0.4007	33.8919	1.9e-7	0.0007	∗	2.7e-8	2.6e-5
15	3.2564	0.4513	49.8737	6.3e-8	0.0003	∗	8.9e-9	1.2e-5

The lower bound (3.22) is our primary concern. For $|\zeta_j|\,\|R_j^{-1}P_j\| \ll 1$, and $|\tau_j| \approx 1$, see (3.14), (3.15), the first and the second factor in this bound are typically not small, and the GMRES residuals for T_j and $\hat{b}^{(j)}$ can substantially decrease within the first $N-l$ steps only if R_j is highly ill conditioned.

For illustration we turn to Experiments 3.4 and 3.5. Figure 3.7 shows the absolute values of the entries of the right-hand side vectors in Experiment 3.4. Each solid line, except for the line representing $|\hat{b}^{(8)}|$, represents a pair of vectors $|\hat{b}^{(j)}|, |\hat{b}^{(N-j+1)}|$, $j = 1, \ldots, 7$. For all j, $\rho_1^{(j)}$ is the first nonzero entry in $\hat{b}^{(j)}$. We can therefore apply (3.22) with $l = 1$. The corresponding numerical values of the factors in (3.22) are shown in Table 3.1. The parameters chosen in Experiment 3.4 yield a moderate mesh Peclet number, $P_h = 3.125$ (cf. Experiment 3.3). Hence (3.14) with $|\zeta_j|$ sufficiently small holds only for $j = 1, 2, 3, 4$, and, to a lesser extend, for $j = 5, 6$. For these indices we have $|\zeta_j|\,\|R_j^{-1}P_j\| < 1$, so that (3.22) is applicable. The column "$(3.22)/\|\hat{b}^{(j)}\|$" shows that, in the first $N-1 = 14$ steps, GMRES makes little progress for the individual systems (3.3) corresponding to $j = 1, 2, 3, 4$. Consequently, the slow initial convergence of GMRES when applied to the coupled system (3.2), as well as to the original system (3.1), lasts (at least) for 14 steps. This is clearly visible in Figure 3.3. The two rightmost columns show the relative GMRES residual norms $\|\hat{r}_{N-1}^{(j)}\|/\|\hat{b}^{(j)}\|$, and the values of the upper bound (3.23) for the relative residual norms.

We now explain some subtle points illustrated by Experiment 3.5. The components of the right-hand side vectors $\hat{b}^{(j)}$ are shown in Figure 3.8. Since $\rho_6^{(j)}$ is the first nonzero entry in each $\hat{b}^{(j)}$, $j = 1, \ldots, 15$, we are tempted to apply (3.22) with $l = 6$ ($N - l = 9$). However, note that since

$$|\rho_6^{(j)}| \;\approx\; |\rho_7^{(j)}| \;\ll\; |\rho_8^{(j)}|\,, \tag{3.24}$$

the matrices R_j are ill conditioned ($\sigma_{min}(R_j) = \mathcal{O}(10^{-7})$) for all $j = 1, \ldots, 15$. Consequently, the values of the lower bound (3.22) are very small for all j. This is in agreement with the actual GMRES residual norms $\|\hat{r}_9^{(j)}\|$ for the individual systems (3.3) in Figure 3.5.

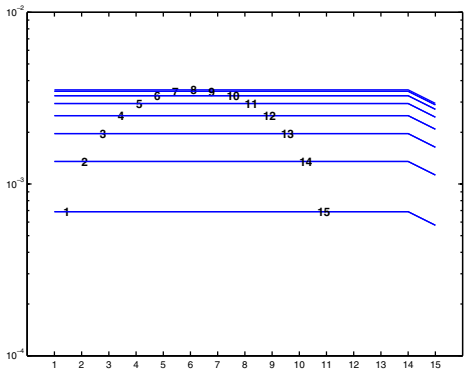

FIG. 3.7. *Absolute values of the entries in the right-hand side vectors* $\hat{b}^{(j)}$, $j = 1, \ldots, 15$, *used Experiment* 3.4.

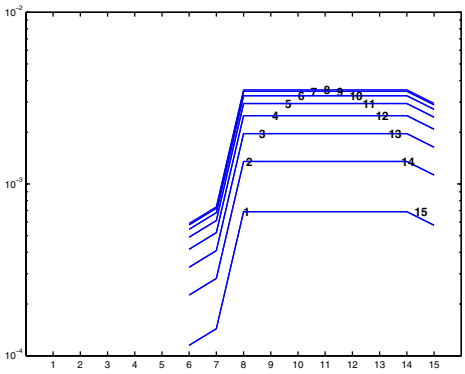

FIG. 3.8. *Absolute values of the entries in the right-hand side vectors* $\hat{b}^{(j)}$, $j = 1, \ldots, 15$, *used Experiment* 3.5.

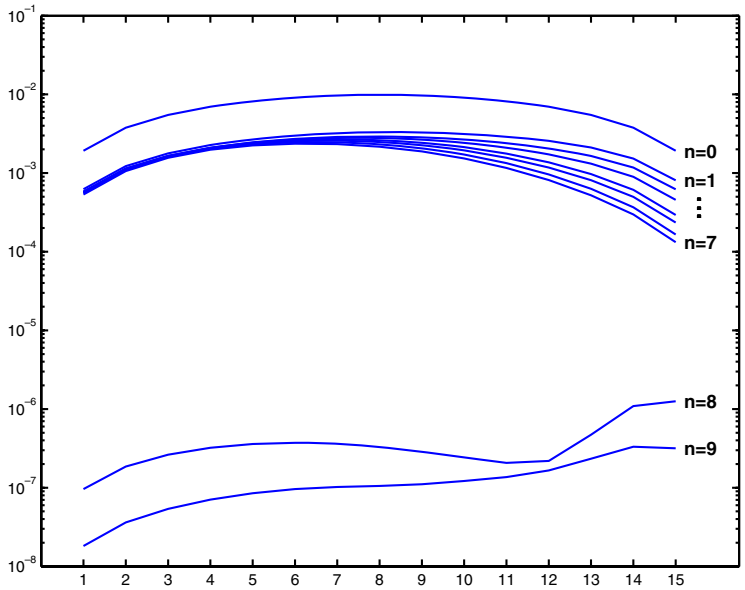

FIG. 3.9. $\sigma_{min}([\hat{b}^{(j)}, (S^T + \zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T + \zeta_j S)^n\,\hat{b}^{(j)}])$, *cf.* (3.18), *for* $j = 1, \ldots, 15$ *and* $n = 0, \ldots, 9$, *corresponding to Experiment* 3.5.

Since our analysis cannot be based on using (3.22) and the step $N - l$, we turn to the lower bound (3.18), which is applicable for all $n = 0, 1, \ldots, N - l$. The values of $|\tau_j|$ given in Table 3.1 are valid also for Experiment 3.5. Hence for small j the first factor in (3.18) does not decrease significantly. Moreover, as shown in Figure 3.9, the second factor in (3.18),

$$\sigma_{min}\left(\left[\hat{b}^{(j)}, (S^T + \zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T + \zeta_j S)^n\,\hat{b}^{(j)}\right]\right),$$

stays for all $j = 1, \ldots, 15$ on the order $\mathcal{O}(10^{-3})$, and thus close to $\mathcal{O}(\|\hat{b}^{(j)}\|)$ until $n = N - 8 = 7$. This corresponds to the fact that $\rho_8^{(j)}$ is the first *significant* entry

(in the quantitative sense of (3.24)) in each of the vectors $\hat{b}^{(j)}$, $j = 1, \ldots, 15$. The bound (3.18) then implies that for small j the GMRES residual norms for T_j and $\hat{b}^{(j)}$ converge slowly for the first seven steps, which is precisely what we observe in Figure 3.5. Further numerical illustrations of these subtleties can be found in [18, Section 7].

In summary, the presence of at least one system (3.3) with a tridiagonal Toeplitz matrix $T_j = \text{tridiag}(\gamma_j, \lambda_j, \mu_j)$ satisfying (3.14), i.e., with T_j close to the Jordan block $\text{tidiag}(1, 1, 0)$, and with l representing the index of the first significant entry of the corresponding right-hand side, prevents fast convergence of GMRES for the original system (3.1) for the initial $N - l$ steps. As shown in section 3.2, the relation (3.14) holds (whenever $\delta \approx \delta_*$) for small j when P_h is moderate, and for all j when P_h is large. Therefore the initial phase of slow convergence is *typical* for matrices arising from the SUPG discretization of the convection-diffusion model problem used in this paper. Our considerations also show that in case of a general nonzero source term in (2.1), we can *typically* expect that the initial phase lasts $N - 1$ steps, unless the source term has a special structure that gives leading zeros (or very small values) in the right-hand side vectors $\hat{b}^{(j)}$ analogously to the boundary conditions (2.24). We also point out that a nonzero initial guess x_0 that is not related to the problem (e.g., a "random" x_0) most likely leads to an initial phase lasting $N - 1$ steps, *regardless of the source term and the boundary conditions*. Such an x_0 clearly represents an unwise choice in this context (also cf. our general discussion in the introduction). We next ask why from the step $N - l + 1$ GMRES converges with an increased rate.

3.4. Acceleration of convergence. As explained in section 3.1, the lower bound (3.7) is useless for analyzing the convergence behavior after the step $N - 1$, possibly even earlier. Hence the above approach *cannot* be used for quantifying any possible acceleration of convergence after the initial phase. In fact, any quantification of that phenomenon appears to be difficult.

In order to illustrate the difficulties we consider, for simplicity, a block-diagonal matrix consisting of N lower bidiagonal Toeplitz blocks (scaled Jordan blocks) of size N by N, all corresponding to the *same* eigenvalue λ. Let the corresponding block right-hand sides of length N have their first nonzero entries in the lth positions (all being assumed significant in the quantitative sense above). When for at least one of the individual Toeplitz blocks the subdiagonal entry is close in magnitude to λ, then our analysis in section 3 shows that GMRES will for this block, and, consequently, for the whole system, converge slowly for $N - l$ steps. In step $N - l + 1$, however, GMRES will construct the minimal polynomial of the system matrix with respect to the given particular right-hand side, which is in this case equal to $(\lambda - z)^{N-l+1}$ (for details about GMRES and the minimal polynomial of a matrix see [1, Section 3] and the references given there). Hence in this case the acceleration of convergence after the initial phase will be maximal—finding the exact solution will only take one additional step.

In (3.1)–(3.3), however,

- the N diagonal blocks T_j are tridiagonal (not bidiagonal) Toeplitz;
- the minimal polynomial of the diagonal blocks generally differ from each other.

If the superdiagonal of T could be considered a "perturbation" of its lower bidiagonal part, i.e., if (3.14) holds for all j, then the first difficulty could (even quantitatively) be overcome. A possible approach for that could be based on [19, Theorem 3.1], which

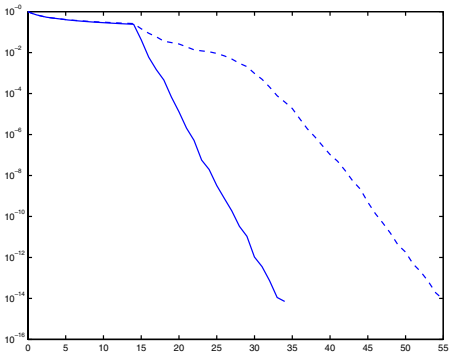

FIG. 3.10. *Relative GMRES residual norms for the systems* (3.2) *with* $h = 1/16$, $\nu = 0.01$ *(solid) and* $\nu = 0.0001$ *(dashed), the respective values* $\delta = \delta_*$ *(*$\delta = 0.34$ *for* $\nu = 0.01$ *and* $\delta = 0.4984$ *for* $\nu = 0.0001$*), and the boundary conditions* (2.22)–(2.23).

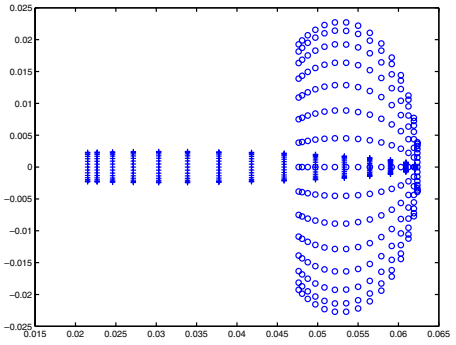

FIG. 3.11. *Eigenvalues* σ_{jk}, *for* $j, k = 1, \ldots, 15$, *cf.* (4.1), *of the matrices* $A_V(1/16, 0.01, \delta_*)$ *(*$\circ$*) and* $A_V(1/16, 0.0001, \delta_*)$ *(*$+$*), which correspond to the values of* λ_j, μ_j, *and* γ_j, *for* $j = 1, \ldots, 15$, *shown in Figures* 3.1 *and* 3.2, *respectively.*

describes the explicit mapping from $\hat{b}$ to r_n, given by $p_n(T)$,

$$r_n \;=\; p_n(T)\,\hat{b} = \Big[(p_n(T_1)\,\hat{b}^{(1)})^T, \ldots, (p_n(T_N)\,\hat{b}^{(N)})^T\Big]^T \;. \tag{3.25}$$

The second difficulty appears to be more challenging to resolve. As suggested by (3.6), a noticeable acceleration of convergence can only occur if all terms $\|p_n(T_j)\hat{b}^{(j)}\|$ of significant value decrease (with some possible little variations) from some step onwards. This point cannot, to our opinion, be easily quantified. In our model problem we observe that the acceleration of convergence is slower at higher Peclet numbers for problems whose parameters are otherwise the same. This is illustrated by the following experiment.

Experiment 3.6. Consider Experiment 3.3, see Figures 3.1 and 3.2. For the large mesh Peclet number $P_h = 312.5$, condition (3.14) is satisfied for all j (cf. Figure 3.2) and, with our argument above, the system matrices T_j in (3.3) can indeed, with a small inaccuracy, be considered as N lower bidiagonal matrices. However, the differences between the eigenvalues λ_j, μ_j, and γ_j of the individual Toeplitz blocks T_j, $j = 1, \ldots, N$, are slightly more pronounced for $P_h = 312.5$ than for $P_h = 3.125$. This is mainly due to the larger differences between the individual λ_j, which cannot be compensated for by smaller differences between the individual γ_j, respectively μ_j. In this case we can therefore expect that after the step $N - l$ the acceleration of convergence for the larger mesh Peclet number $P_h = 312.5$ will be much less pronounced than for the moderate mesh Peclet number $P_h = 3.125$. This is illustrated in Figure 3.10 which compares the relative GMRES residual norms $\|r_n\|/\|\hat{b}\|$ for the two systems corresponding to the right-hand sides from the boundary conditions (2.22)–(2.23). Note that the observed behavior is not at all obvious from the shapes of the corresponding spectra shown in Figure 3.11. ∎

4. Eigendecomposition and GMRES convergence analysis. In this section we refine the analysis based on eigendecomposition of the system matrix A_V in (2.10). It is easy to see from (2.14)–(2.17) that the existence and form of this eigendecomposition are determined by the existence and form of the eigendecompositions

of the matrices T_j, $j = 1, \ldots, N$. If $\gamma_j \mu_j \neq 0$, which is for the case $P_h > 1$ and $\delta = \delta_*$ guaranteed by Lemma 3.1, then the N distinct eigenvalues of T_j are given by

$$\sigma_{jk} \;=\; \lambda_j + \mu_j \zeta_j^{-1/2}\, \omega_k\,, \quad \zeta_j = \frac{\mu_j}{\gamma_j}\,, \quad \omega_k \;=\; 2\cos(kh\pi)\,, \quad k = 1, \ldots, N\,, \tag{4.1}$$

with the corresponding normalized eigenvectors given by

$$\nu_{jk}\, \Delta_j\, u_k\,, \quad k = 1, \ldots, N\,,$$

where $\Delta_j \equiv \mathrm{diag}(\zeta_j^{-1/2}, \ldots, \zeta_j^{-N/2})$ and $\nu_{jk} \equiv \|\Delta_j\, u_k\|^{-1}$; see, e.g., [29, pp. 113–115]. Clearly, when $|\zeta_j| = |\mu_j/\gamma_j| \ll 1$, the eigenvectors of T_j are ill conditioned.

Obviously, the N^2 eigenvalues of A_V are the values σ_{jk} for $j, k = 1, \ldots, N$. Furthermore, the mixed product property of the Kronecker product shows that a unit norm eigenvector corresponding to σ_{jk} is

$$y_{jk} \;=\; \chi_{jk}\, u_j \otimes [\Delta_j u_k]\,, \quad \chi_{jk} \;\equiv\; \|u_j \otimes [\Delta_j u_k]\|^{-1}\,, \quad j, k = 1, \ldots, N\,. \tag{4.2}$$

We denote the resulting eigenvector matrix of A_V by

$$Y \;\equiv\; [y_{11}, y_{12}, \ldots, y_{1N}, \ldots, y_{N1}, y_{N2}, \ldots, y_{NN}] \;\equiv\; [Y_1, \ldots, Y_N]\,. \tag{4.3}$$

Since A_V is diagonalizable we could have based our convergence analysis of GMRES on its eigendecomposition. In particular, we could have applied the standard GMRES convergence bound [28, Proposition 4],

$$\begin{aligned} \|r_n\| &= \min_{p \in \pi_n} \|p(A_V) b_V\| \\ &\leq \kappa(Y) \min_{p \in \pi_n} \max_{j,k=1,\ldots,N} |\, p(\sigma_{jk})\,|\, \|b_V\|\,, \end{aligned} \tag{4.4}$$

where $\kappa(Y) = \sigma_{\max}(Y)/\sigma_{\min}(Y)$ denotes the condition number of Y. However, as noted in [13, 14], the term $\kappa(Y)$ in this bound is typically very large. For example, when $h = 1/16$, $\nu = 0.01$, and $\delta = \delta_* = 0.34$, then a Matlab computation using (4.2) yields $\kappa(Y) = 2.1207e + 17$. The ill-conditioning of Y is not an oddity of our specific model problem, but corresponds to the general strong nonnormality of discretized convection-diffusion operators, particularly for mesh Peclet numbers greater than one; see, e.g., [26]. Such nonnormality makes the direct application of (4.4) rather complicated for proving well-justified quantitative conclusions about the GMRES convergence for discretized convection-diffusion problems. Still, it can be useful to look at the eigendecomposition in relation to the particular right-hand side and study the behavior of the individual components in the GMRES computation [11, 12]. It might also be useful to consider worst-case bounds (for a related discussion see [20]) in some cases, in particular bounds based in the polynomial numerical hull and related techniques, see [15].

We continue with some details of the eigenstructure of A_V. Note that

$$y_{jk}^T\, y_{il} \;=\; \chi_{jk}\chi_{il}\, (u_j^T u_i) \otimes (u_k^T \Delta_j \Delta_i\, u_l) \;=\; 0 \quad \text{for } \; j \neq i\,,$$

which gives the following.

PROPOSITION 4.1. *The eigenvectors of A_V in the ordering given by* (4.3) *form mutually orthogonal blocks, i.e., $Y_j^T Y_i = 0$ for $j \neq i$.*

The proposition implies that the conditioning of Y is fully determined by the conditioning of the eigenvectors y_{jk}, $k = 1, \ldots, N$, *within* each block Y_j, $j = 1, \ldots, N$, and that

$$\kappa(Y) = \max_{j=1,\ldots,N} \kappa(Y_j)\,. \tag{4.5}$$

In particular, if the eigenvectors within each block were mutually orthogonal, which is equivalent to $\Delta_j = I$ for all $j = 1, \ldots, N$, i.e., to $A_V = A_V^T$, then $\kappa(Y) = 1$.

It follows from (4.2) that $\kappa(Y_j)$ is large whenever Δ_j is far from the identity matrix, meaning that $|\zeta_j|$ must be either very large or very small. In our application $|\zeta_j| < 1$, with $|\zeta_j| \ll 1$ (at least) for small indices j. For these indices $\kappa(Y_j)$ is very large, and it is maximal for the minimal $|\zeta_j|$.

For numerical illustration we use the parameters $h = 1/16$, $\nu = 0.01$, and $\delta = \delta_* = 0.34$ and give the resulting values of $|\zeta_j|$ and $\kappa(Y_j)$, $j = 1, \ldots, N$, in the following table.[1]

j	$\lvert\zeta_j\rvert$	$\kappa(Y_j)$		j	$\lvert\zeta_j\rvert$	$\kappa(Y_j)$
1	0.0010	7.2672e+16		9	0.1185	3.4121e+06
2	0.0042	2.8020e+16		10	0.1585	4.3948e+05
3	0.0096	1.5523e+14		11	0.2082	6.4019e+04
4	0.0176	2.2296e+12		12	0.2678	1.0790e+04
5	0.0286	7.4153e+10		13	0.3347	2.2326e+03
6	0.0432	4.0925e+09		14	0.4007	6.2599e+02
7	0.0623	3.1392e+08		15	0.4513	2.6995e+02
8	0.0870	3.0166e+07				

In summary, the most ill-conditioned blocks Y_j in our example correspond to the tridiagonal Toeplitz systems in (3.3) that satisfy (3.14), and that are responsible for the initial phase of slow GMRES convergence. Thus, the eigendecomposition reveals which blocks are the most troublesome for the GMRES convergence.

5. Concluding discussion. This paper is devoted to the convergence analysis of GMRES applied to an SUPG discretized convection-diffusion model problem with dominating convection. The eigendecomposition of the discretized operator is known analytically, but the transformation to the eigenvector coordinates is highly ill conditioned. Therefore any analysis based on it, which aims at describing the initial stage of convergence, must involve a rather complicated pattern of cancellation of potentially huge components of the initial residual in the individual eigenspaces. Instead of following this technically complicated and physically unnatural approach, we propose another idea.

Assume that a linear algebraic system can be transformed using a well-conditioned transformation to a new system with a structure of the matrix, not necessarily diagonal, for which the GMRES convergence can be more easily understood. Then the geometry of the space is not significantly distorted by the transformation, and using the particular structure of the transformed system we can describe the GMRES convergence for the original problem.

[1] Note that excessive ill-conditioning of Y_j particularly for small indices j leads to round-off errors even when we use the analytic formulas (4.2) for the eigenvectors of A. Hence (4.5) does not hold in our finite precision computation.

In our application we use an orthonormal similarity transformation, and the transformed system is block diagonal with nonsymmetric tridiagonal Toeplitz blocks. Our approach clearly describes the relationship between the boundary conditions in the model problem and the initial phase of slow GMRES convergence for the linear algebraic system. Our results reveal, as a by-product, a possibly very complicated relationship of the eigeninformation and GMRES convergence.

Our results can be extended to a three-dimensional model problem described by Ramage [25] as well as to other separable second-order PDEs on rectangular domains. Note that the fact that the tridiagonal blocks of our transformed system were Toeplitz is not of any particular importance for the character of the GMRES convergence. If we perturbed the nonzero constant diagonals of Toeplitz blocks so that they were nonconstant but the relation between the magnitudes of the diagonals was still approximately preserved, then the convergence behavior would not change much. Application of the idea of using well-conditioned transformations to some easy-to-use structure in a more general context will be a subject of further work.

Acknowledgments. We thank Michael Eiermann and Oliver Ernst for sharing their unpublished notes [6] and for very stimulating discussions and advice about the subject matter of this paper. We also wish to thank Mark Embree for sharing his unpublished notes [11, 12], Howard Elman, Anne Greenbaum, Dianne O'Leary, and Nick Trefethen for helpful discussions, and the two anonymous referees for numerous suggestions that helped to improve the paper.

REFERENCES

[1] M. Arioli, V. Pták, and Z. Strakoš, *Krylov sequences of maximal length and convergence of GMRES*, BIT, 38 (1998), pp. 636–643.

[2] O. Axelsson, V. Eijkhout, B. Polman, and P. Vassilevski, *Incomplete block-matrix factorization iterative methods for convection-diffusion problems*, BIT, 29 (1989), pp. 867–889.

[3] B. Beckermann and A. Kuijlaars, *Superlinear CG convergence for special right-hand sides*, Electron. Trans. Numer. Anal., 14 (2002), pp. 1–19.

[4] 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.

[5] M. Eiermann, *Semiiterative Verfahren für nichtsymmetrische lineare Gleichungssysteme*, Habilitationsschrift, Universität Karlsruhe, Karlsruhe, Germany, 1989.

[6] M. Eiermann and O. G. Ernst, *GMRES and Jordan Blocks*, unpublished notes, 2002.

[7] H. C. Elman and M. P. Chernesky, *Ordering effects on relaxation methods applied to the discrete one-dimensional convection-diffusion equation*, SIAM J. Numer. Anal., 30 (1993), pp. 1268–1290.

[8] H. C. Elman and A. Ramage, *A characterization of oscillations in the discrete two-dimensional convection-diffusion equation*, Math. Comp., 72 (2001), pp. 263–288.

[9] H. C. Elman and A. Ramage, *An analysis of smoothing effects of upwinding strategies for the convection-diffusion equation*, SIAM J. Numer. Anal., 40 (2002), pp. 254–281.

[10] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Iterative methods for problems in computational fluid dynamics*, in Iterative Methods in Scientific Computing (Hong Kong, 1995), Springer, Singapore, 1997, pp. 271–327.

[11] M. Embree, *GMRES Residual Behavior in a Lousy Basis*, unpublished notes, 2000.

[12] M. Embree, *GMRES Residual Behavior in the Eigenvector Basis: A Study of Two Practical Examples*, unpublished notes, 2000.

[13] O. G. Ernst, *Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1079–1101.

[14] B. Fischer, A. Ramage, D. Silvester, and A. Wathen, *On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 179 (1999), pp. 179–195.

[15] A. Greenbaum, *Some theoretical results derived from polynomial numerical hulls of Jordan blocks*, Electron. Trans. Numer. Anal., submitted.

[16] A. Greenbaum and Z. Strakoš, *Matrices that generate the same Krylov residual spaces*, in Recent Advances in Iterative Methods, IMA Vol. Math. Appl. 60, G. H. Golub, A. Greenbaum, and M. Luskin, eds., Springer-Verlag, New York, 1994, pp. 95–118.

[17] T. J. R. Hughes and A. Brooks, *A multidimensional upwind scheme with no crosswind diffusion*, in Finite Element Methods for Convection Dominated Flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979), AMD 34, Amer. Soc. Mech. Engrs. (ASME), New York, 1979, pp. 19–35.

[18] J. Liesen and Z. Strakoš, *Convergence Analysis of GMRES for the SUPG Discretized Convection Diffusion Model Problem*, Technical report 26-2003, Technical University of Berlin, Institute of Mathematics, 2003; also available online from http://www.math.tu-berlin.de/preprints/abstracts/Report-26-2003.rdf.html.

[19] J. Liesen and Z. Strakoš, *Convergence of GMRES for tridiagonal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 233–251.

[20] J. Liesen and P. Tichý, *The worst-case GMRES for normal matrices*, BIT, 44 (2004), pp. 79–98.

[21] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London, 1996.

[22] C. C. Paige and Z. Strakoš, *Residual and backward error bounds in minimum residual Krylov subspace methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1898–1923.

[23] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.

[24] G. D. Raithby, *Skew upstream differencing schemes for problems involving fluid flow*, Comput. Methods Appl. Mech. Engrg., 9 (1976), pp. 153–164.

[25] A. Ramage, *A Note on Parameter Choice and Iterative Convergence for Stabilised Discretisations of Advection-Diffusion Problems in Three Dimensions*, Technical report 32, University of Strathclyde, Department of Mathematics, 1998.

[26] S. C. Reddy and L. N. Trefethen, *Pseudospectra of the convection-diffusion operator*, SIAM J. Appl. Math., 54 (1994), pp. 1634–1649.

[27] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin, 1996.

[28] 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.

[29] G. D. Smith, *Numerical Solution of Partial Differential Equations. Finite Difference Methods*, 2nd ed., Oxford Applied Mathematics and Computing Science Series, The Clarendon Press, Oxford University Press, New York, 1978.

[30] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.