CONVERGENCE ANALYSIS OF GMRES FOR THE SUPG DISCRETIZED CONVECTION-DIFFUSION MODEL PROBLEM

J. LIESEN[†] AND Z. STRAKOŠ[‡]

Abstract. When GMRES [25] is applied to streamline upwind Petrov Galerkin (SUPG) discretized convection-diffusion problems [17], [4], [19], it typically exhibits an initial period of slow convergence followed by a faster decrease of the residual norm. Ernst conjectured in [13] that the duration of the initial phase is governed by the number of steps needed for boundary information to pass from the inflow boundary across the (discretized) domain following the longest streamline of the velocity field. He also illustrated that for these practical problems eigenvalues alone might give misleading information about convergence. He focused in his analysis on the field of values. Using the eigendecomposition of the discretized operator, Fischer, Ramage, Silvester and Wathen analyzed in [14] the choice of parameters in the SUPG discretization and their relation to convergence of GMRES. Since the analyses in [13] and [14] are based on the discretized operator only, they can not explain the length of the initial period of slow convergence which depends on the right hand side of the linear system (and hence on the boundary conditions).

In this paper we concentrate on a model problem on the unit square with Dirichlet boundary conditions, a regular grid and a constant velocity field parallel to one of the axes. Instead of the eigendecomposition of the block tridiagonal system matrix with symmetric blocks as in [14] we consider the simultaneous diagonalization of the symmetric matrix blocks. With permutation of unknowns this results in a block diagonal system matrix with unsymmetric tridiagonal blocks, a structure that was also employed in [5], [6], [8], [9]. Applying results from [18] we 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, SUPG discretization, GMRES, rate of convergence, ill conditioned eigenvectors, nonnormality, tridiagonal Toeplitz matrices

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

1. Introduction.

Convergence of modern iterative methods for solving *linear* algebraic systems (such as Krylov subspace methods) represents a complicated *nonlinear* problem. In classical stationary iterative methods (such as SOR) the description of convergence is (in fact) linearized by focusing not on the transient period but on the asymptotic convergence factors, see the pioneering and fundamental work of Young [29], [30] and Varga [28]. The goal of preconditioned Krylov subspace methods is to achieve sufficiently accurate approximate solution in a reasonable number of steps. Apart from the lucky situation when a good preconditioner leads to very fast convergence, the asymptotic approach to convergence analysis of Krylov subspace methods, though still useful, can not dominate. A more detailed description of convergence must be based not on a single number (such as the 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]),

[†] 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, USA, supported by the US 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 and by the Ministry of Education of the Czech Republic under project MSM242200002.

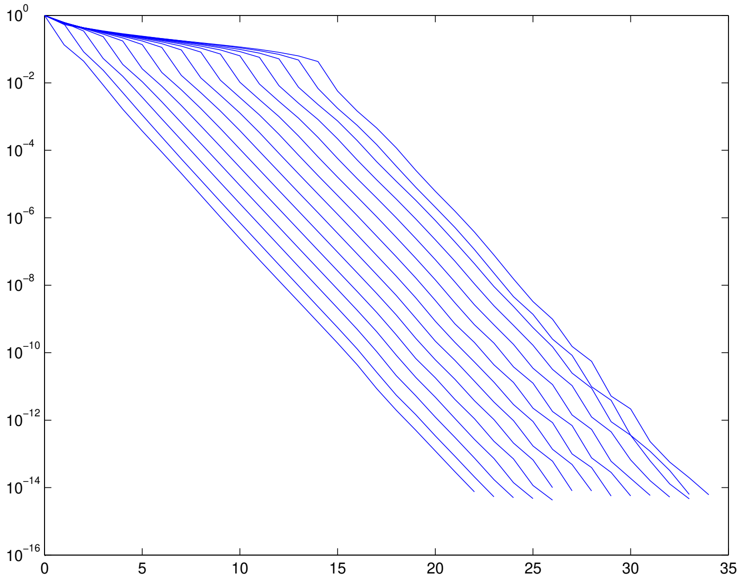


Fig. 1.1. *Typical GMRES convergence (measured by the relative residual norm) when applied to SUPG discretized convection-diffusion problems. Different behavior corresponds to the same discretized operator but to different boundary conditions.*

the matrix eigenvalues answer practical questions about convergence of (symmetric) Krylov subspace methods. If the system matrix is unsymmetric and 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 problems, and their solution with GMRES [25]. 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 has been proved in [15] and [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 showed, however, an example of a convection-diffusion problem discretized via the streamline upwind Petrov Galerkin (SUPG) method for which the eigenvalues alone gave indeed misleading information about convergence [13]. 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 GMRES behavior for the model problem specified in Section 2.1 of our paper with different right hand sides specified in Section 2.3 is illustrated in Fig. 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. The presence of an initial phase of slow convergence as illustrated in Fig. 1.1 proves the necessity of including into the convergence analysis the particular right hand side of the linear system (and hence the boundary conditions of the PDE). For an example of a similar philosophy considering the role of right hand sides, besides eigenvalues and eigenvectors, in *complete* stagnation of GMRES we refer to [31].

Here we focus, as in [8], [9], on a convection-diffusion model problem with a constant velocity field parallel to one of the axes and with Dirichlet boundary conditions. Eigenvalues and eigenvectors of the corresponding SUPG discretized operator are known analytically. It seems therefore natural to exploit the corresponding eigenexpansion of the initial residual, 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 was clearly formulated by Trefethen in [27, p. 384] (see also [23]):

> "The difficulty with nonnormal matrices and operators goes beyond the need for a Jordan canonical form instead of a diagonalization if a matrix lacks a complete set of eigenvectors. Any use of eigenvalues to derive physical predictions relies on an implicit transformation to eigenvector coordinates. If the matrix is normal, this transformation is unitary – a rotation or a reflection. If it is far from normal, however, the change to eigenvector coordinates may involve an extreme distortion of the state space. In the new coordinates, the physics of the system may become strangely complicated. A typical state of the system may be a superposition of huge eigenfunction components that nearly cancel, and the evolution over time intervals of scientific interest may be determined by how this pattern of cancellation evolves, rather than by the growth or decay of the individual eigenfunctions. In other words, there may be no good scientific reason for attempting to analyze the problem in terms of eigenvalues and eigenvectors."

Apart from the time evolution, the model problem used throughout our paper represents an illuminative illustration of this quote. In convection-diffusion problems the initial period of possible slow convergence is of primary importance, and the eigenvalues combined with the ill-conditioned eigenvectors do not represent a proper tool for analyzing it.

In some papers, see [14] and [10], the SUPG discretized model operator is presented as a block tridiagonal matrix with symmetric tridiagonal Toeplitz blocks. Based on this representation we use, instead of the operator's eigendecomposition, the simultaneous eigendecomposition of its tridiagonal Toeplitz submatrices. With a proper reordering of unknowns the resulting N^2 by N^2 system with a block tridiagonal matrix with diagonal blocks can be decomposed into N independent N by N systems with *unsymmetric* tridiagonal Toeplitz matrices. The last formulation of the SUPG discretized convection-diffusion model problem has been used by Elman and Ramage [8], [9], and by Eiermann and Ernst [5], [6] who derive it directly from the continuous problem. It will prove especially convenient in our analysis of the initial phase of convergence.

We would like to point out the following: The convergence of GMRES is usually measured by the residual norm. Though we do not consider this an ideal measure of convergence, we use it for simplicity also in this paper. We believe that measuring convergence of iterative methods represents a complicated issue which needs a thorough reconsideration. This is beyond our purpose here. Nevertheless, we stress at least one simple, still sometimes overlooked, point. Unless a nonzero initial guess x_0 is available that contains useful information about the solution x, for example a choice of 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, all measured by the relative residual norm. For examples see [20, relation (2.8), and the discussion of Figures 7.9 and 7.10]. Any such choice of x_0 is, however, useless. Throughout this paper we always use $x_0 = 0$.

The paper is organized as follows: Section 2 specifies the model problem and summarizes properties of the block tridiagonal discretized system. In Section 3 we describe the correspondence between the standard formulation of the discretized system and the Kronecker product formulation used by Eiermann and Ernst. The initial period of slow convergence is analyzed in Section 4. In Section 5 we investigate the subsequent phase of convergence. In Section 6 we return to the discussion of the eigenvalue decomposition, and in Section 7 we present numerical experiments. Concluding remarks close the paper.

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,

$$-\nu\, \nabla^2 u + w \cdot \nabla u \;=\; f, \quad \text{in } \Omega = (0,1) \times (0,1), \qquad u = g \text{ on } \partial\Omega. \tag{2.1}$$

Here the scalar-valued function $u(x, y)$ represents the concentration of the transported quantity, $w = [w_x, w_y]^T$ the velocity field and ν the scalar diffusion parameter. This model problem has been used and studied in many publications, see, e.g. [2], [10], [13], [14] and, in particular, [8], [9]. We are interested in the *convection-dominated* case, i.e. we assume $\|w\| \gg \nu$ in (2.1). We consider a bilinear finite element Galerkin discretization on a regular grid with square elements of the size $h \times h$,

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

where N represents the number of inner nodes along each side.

Nonphysical oscillations present in the discrete Galerkin solution for the *mesh Peclet number*

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

greater than one can be suppressed by adding some stabilization terms, which modifies the bilinear form and right hand side functionals. Here we consider the *streamline upwind Petrov Galerkin* (SUPG) discretization, see [17], [4], [19], [13, equations (2.4) and (2.5)] and [14, equations (7) and (8)]. Then the stabilization 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 $\hat{\delta}$ represents a stabilization parameter. When (2.1) is convection dominated, then $\hat{\delta}$ is typically chosen as

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

where $\delta > 0$ is a *tuning parameter*. The choice of δ is discussed in [24, Remark 3.34, p. 234]. If piecewise linear finite elements are used for a one-dimensional constant coefficient problem, then

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

yields the exact solution at the node points, see [17], [4, Section 2.4], [24, Section I 2.1.3]. A similarly optimal choice of δ for two or more dimensional problems is unknown. 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). The authors of [14] try to correlate such choice with a fast convergence of GMRES for the resulting linear system.

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 also 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 analysis of GMRES convergence in our paper to the discretized problems with the value $\delta = \delta_*$ with

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

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

2.1. SUPG discretized operator.

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

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

where $A_d = \langle \nabla\phi_j, \nabla\phi_i \rangle$, $A_c = \langle w \cdot \nabla\phi_j, \phi_i \rangle$ respectively $A_s = \langle w \cdot \nabla\phi_j, w \cdot \nabla\phi_i \rangle$ represent the diffusion, the convection respectively the stabilization term, and $\phi_1, \ldots, \phi_{N^2}$ are the finite element basis functions. Both A_d and A_s are symmetric positive definite while A_c is skew symmetric, see [14, identity (11)]. For constant velocity fields w the constituent matrix stencils can be found in [14, formulas (12)–(14)].

Throughout this paper we concentrate on a special case of the *vertical wind* $w = [0, 1]^T$. Then the constituent stencil

$$\begin{tikzcd} m_4 & m_3 & m_4 \\ m_2 & m_1 \arrow[l] \arrow[r] \arrow[u] \arrow[d] \arrow[ul] \arrow[ur] \arrow[dl] \arrow[dr] & m_2 \\ m_6 & m_5 & m_6 \end{tikzcd} \tag{2.7}$$

has numerical values

$$\begin{tikzcd}
-\frac{\nu}{3}+\frac{h}{12}(1-2\delta) & -\frac{\nu}{3}+\frac{h}{3}(1-2\delta) & -\frac{\nu}{3}+\frac{h}{12}(1-2\delta) \\
-\frac{\nu}{3}+\frac{\delta h}{3} & \frac{8}{3}\nu+\frac{4}{3}\delta h \arrow[ul] \arrow[u] \arrow[ur] \arrow[l] \arrow[r] \arrow[dl] \arrow[d] \arrow[dr] & -\frac{\nu}{3}+\frac{\delta h}{3} \\
-\frac{\nu}{3}-\frac{h}{12}(1+2\delta) & -\frac{\nu}{3}-\frac{h}{3}(1+2\delta) & -\frac{\nu}{3}-\frac{h}{12}(1+2\delta)\,.
\end{tikzcd} \tag{2.8}$$

Using the natural *horizontal line ordering* of unknowns, see also [10, Section 3.1, Fig. 1], A has with the vertical wind the form of a block tridiagonal matrix with symmetric tridiagonal Toeplitz blocks,

$$A \;=\; A(h,\nu,\delta) \;=\; \mathrm{tridiag}(M_3,M_1,M_2) \;\in\; \mathbf{R}^{N^2\times N^2}, \tag{2.9}$$

where the $N\times N$ real symmetric Toeplitz blocks

$$\begin{aligned} M_1 &= \mathrm{tridiag}(m_2,m_1,m_2)\,, \\ M_2 &= \mathrm{tridiag}(m_4,m_3,m_4)\,, \\ M_3 &= \mathrm{tridiag}(m_6,m_5,m_6)\,, \end{aligned} \tag{2.10}$$

have the entries specified in (2.7)–(2.8).

Writing the coefficient matrix in the form

$$A \;=\; \langle(\nu I+\hat{\delta} w w^T)\nabla\phi_j,\nabla\phi_i\rangle \;+\; \langle w\cdot\nabla\phi_j,\phi_i\rangle\,, \tag{2.11}$$

we have for $w=[0,1]^T$ the 'effective' diffusivity tensor

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

Using Kronecker products, the matrix A can be written as a sum of two terms accounting for the diffusion in the direction $[1,0]^T$, respectively for the diffusion, convection and stabilization in the direction $[0,1]^T$,

$$A \;=\; \nu M\otimes K \;+\; ((\nu+\delta h)K+C)\otimes M\,, \tag{2.12}$$

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

$$\begin{aligned} M &= \frac{h}{6}\,\mathrm{tridiag}(1,4,1)\,, \\ K &= \frac{1}{h}\,\mathrm{tridiag}(-1,2,-1)\,, \\ C &= \frac{1}{2}\,\mathrm{tridiag}(-1,0,1)\,, \end{aligned} \tag{2.13}$$

are the $N\times N$ mass, stiffness and gradient matrices of the one dimensional constant coefficient model problem discretized on a uniform mesh using linear elements.

2.2. Spectral analysis of the symmetric Toeplitz blocks. 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

$$\omega_j \;=\; 2\cos(jh\pi), \quad j=1,\ldots,N\,. \tag{2.14}$$

Furthermore, the corresponding normalized eigenvectors are given by

$$u_j \;=\; (h/2)^{-1/2}\left[\sin(jh\pi),\ldots,\sin(Njh\pi)\right]^T, \quad j=1,\ldots,N, \tag{2.15}$$

see, e.g., [26, pp. 113–115]. The matrix of eigenvectors $U=[u_1,\ldots,u_N]$ is orthonormal and symmetric. Consequently, the matrices M_1, M_2 and M_3, cf. (2.10), as well as M and K, cf. (2.13), are simultaneously diagonalizable by the symmetric orthonormal matrix U, so that

$$\begin{aligned} M_1 &= U\operatorname{diag}(\lambda_{1:N})\,U, & \lambda_j &\equiv m_1+m_2\,\omega_j, \quad j=1,\ldots,N,\\ M_2 &= U\operatorname{diag}(\mu_{1:N})\,U, & \mu_j &\equiv m_3+m_4\,\omega_j, \quad j=1,\ldots,N,\\ M_3 &= U\operatorname{diag}(\gamma_{1:N})\,U, & \gamma_j &\equiv m_5+m_6\,\omega_j, \quad j=1,\ldots,N, \end{aligned} \tag{2.16}$$

and

$$M \;=\; \frac{h}{6}\,U\operatorname{diag}\left(4+\omega_{1:N}\right)\,U, \qquad K \;=\; \frac{1}{h}\,U\operatorname{diag}\left(2-\omega_{1:N}\right)\,U\,. \tag{2.17}$$

In the remainder of this subsection we will analyze the numerical values of λ_j, μ_j and γ_j, $j=1,\ldots,N$, as defined in (2.16). These can be rewritten as

$$\begin{aligned} 3\lambda_j &= 2\delta h\left(2+\frac{\omega_j}{2}\right)+2\nu\left(4-\frac{\omega_j}{2}\right),\\ -3\mu_j &= \delta h\left(2+\frac{\omega_j}{2}\right)+\nu(1+\omega_j)-\frac{h}{2}\left(2+\frac{\omega_j}{2}\right),\\ -3\gamma_j &= \delta h\left(2+\frac{\omega_j}{2}\right)+\nu(1+\omega_j)+\frac{h}{2}\left(2+\frac{\omega_j}{2}\right). \end{aligned} \tag{2.18}$$

We first analyze their signs.

Lemma 2.1. *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 (2.18) satisfy*

$$\lambda_j \;>\; 0 \;>\; \gamma_j. \tag{2.19}$$

Furthermore, for all $j=1,\ldots,N$ the value of μ_j defined in (2.18) satisfies

$$sign(\mu_j) \;=\; sign(f(j)-\delta), \quad where \quad f(j) \;\equiv\; \delta_* \;+\; \frac{1-\omega_j/2}{P_h(4+\omega_j)}, \tag{2.20}$$

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 $\delta=\delta_*$ in (2.18), then $\mu_j \;>\; 0$ for all $j=1,\ldots,N$.

Proof. Considering the relations (2.18) we first note that since $-2<\omega_j<2$ 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 (2.20). Obviously, the left hand side of this equality has the same sign as μ_j, which proves (2.20). When $\delta = \delta_*$ in (2.18), then

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

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

We will next analyze the moduli of 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 (2.18) are equivalent to

$$\begin{aligned} 3\lambda_j &= h\left(2 + \frac{\omega_j}{2}\right) + 4\nu\left(1 - \frac{\omega_j}{2}\right), \\ 3\mu_j &= \nu\left(1 - \frac{\omega_j}{2}\right), \\ -3\gamma_j &= h\left(2 + \frac{\omega_j}{2}\right) - \nu\left(1 - \frac{\omega_j}{2}\right). \end{aligned} \tag{2.21}$$

Straightforward manipulations give the following result.

Lemma 2.2. *Let, as above, $w = [0, 1]^T$ ($\|w\| = 1$), and $\delta = \delta_*$. Then*

$$\frac{|\lambda_j|}{|\gamma_j|} = 1 + 5\left(2P_h\,\frac{4+\omega_j}{2-\omega_j} - 1\right)^{-1}, \tag{2.22}$$

$$\frac{|\mu_j|}{|\gamma_j|} = \left(2P_h\,\frac{4+\omega_j}{2-\omega_j} - 1\right)^{-1}, \qquad j = 1, \ldots, N. \tag{2.23}$$

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

$$\frac{|\lambda_j|}{|\gamma_j|} \;\approx\; 1 \;\gg\; \frac{|\mu_j|}{|\gamma_j|}. \tag{2.24}$$

For a moderate P_h the ratios (2.22) and (2.23) 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, (2.24) 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 2.3. In Figs. 2.1 and 2.2 we show typical examples of the magnitudes of λ_j, γ_j and μ_j. For Fig. 2.1 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 (2.24) 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 Fig. 2.2. Here $P_h = 312.5$ so that (2.24) holds for all $j = 1, \ldots, 15$. □

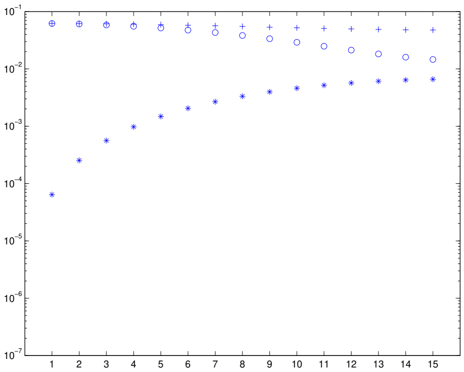


FIG. 2.1. *λ_j (+), μ_j (*) and $|\gamma_j|$ (o) for $j = 1, \ldots, 15$ and $h = 1/16$, $\nu = 0.01$, $\delta = \delta_* = 0.34$.*

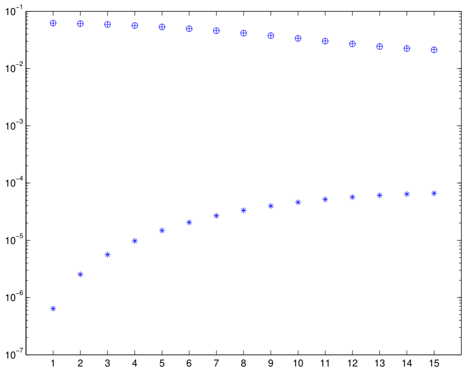


FIG. 2.2. *λ_j (+), μ_j (*) and $|\gamma_j|$ (o) for $j = 1, \ldots, 15$ and $h = 1/16$, $\nu = 0.0001$, $\delta = \delta_* = 0.4984$.*

2.3. Structure of the right hand sides.

We now study the structure of the right hand side vectors b in the linear systems $Ax = b$ arising from the SUPG discretization of (2.1). For simplicity we will assume that $f = 0$, so that the entries of b are *completely determined by the boundary condition* $u = g$ on $\partial\Omega$.

As above, we use the natural horizontal line ordering of unknowns. We partition the right hand side vector b of the length N^2 into N blocks of the length N each, i.e.

$$b \;=\; [b^{(1)T}, \ldots, b^{(N)T}]^T, \tag{2.25}$$

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

$$B_H \;=\; \left[\begin{array}{ccccc} b_{1,1} & b_{1,2} & \cdots & b_{1,N-1} & b_{1,N} \\ b_{2,1} & 0 & \cdots & 0 & b_{2,N} \\ \vdots & \vdots & & \vdots & \vdots \\ b_{N-1,1} & 0 & \cdots & 0 & b_{N-1,N} \\ b_{N,1} & b_{N,2} & \cdots & b_{N,N-1} & b_{N,N} \end{array}\right]. \tag{2.26}$$

The entries of B_H can easily be computed using (2.7)–(2.8). In the experiments presented in this paper we use the following examples.

EXAMPLE 2.4. Following the set of problems introduced by Raithby [21], the authors of [9], [10], [14] use boundary conditions that are discontinuous at the inflow boundary,

$$u(x,0) \;=\; u(1,y) \;=\; 1, \qquad \text{for} \quad 1/2 < x \le 1 \;\text{ and }\; 0 \le y < 1, \tag{2.27}$$

$$u(x,y) \;=\; 0, \qquad \text{elsewhere on } \;\partial\Omega. \tag{2.28}$$

Hence the first column of B_H has nonzero entries given by

$$b_{\lfloor N/2\rfloor,1} = -m_6 \;=\; \frac{\nu}{3} + \frac{h}{12}(1+2\delta)\,,$$

$$b_{\lfloor N/2\rfloor+1,1} = -(m_6 + m_5) \;=\; \frac{2}{3}\nu + \frac{5}{12}h(1+2\delta)\,,$$

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

while

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

are the remaining nonzero entries of B_H. □

Example 2.5. We also consider nonzero boundary conditions only on (a part of) the right side boundary of the unit square. Their solutions have characteristic layers on (a part of) the right side of the domain. 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,

$$u(1,y)=1 \qquad \text{for} \quad k/(N+1)\le y<1, \quad k=0,1,\ldots,N-1, \tag{2.29}$$

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

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

For other values of k the entries of B_H can be computed analogously. □

We use Example 2.5 because it helps to make the main points in our analysis in Section 4 clear. Example 2.4, which is more realistic from a practical point of view, is considered in the analysis in Sections 5 and 6, and also for additional numerical illustrations in Section 7.

3. Basic transformations and structure of the linear algebraic system.

With the vertical wind $w=[0,1]^T$ and the natural horizontal line ordering of unknowns the matrix A in (2.6) is block tridiagonal with symmetric tridiagonal Toeplitz blocks M_1, M_2 and M_3, see (2.9) or, equivalently, (2.12). Instead of focusing on spectral decomposition of A, which is for our model problem typically highly ill-conditioned, we will use, as in [8], [9], the simultaneous diagonalization of the symmetric Toeplitz blocks, which is orthonormal. Reordering of unknowns in the transformed system (which in the original system $Ax=b$ corresponds to reordering of unknowns along the vertical lines, i.e., parallel with the direction of the wind) then results in a system matrix which is block diagonal with unsymmetric tridiagonal Toeplitz blocks. For the sake of smooth readability we will first use the standard form (2.9) of A used frequently in the literature. Then we will show that the whole matter simplifies and gains some elegance when the Kronecker product form (2.12), used in the convection-diffusion context first (to our knowledge) by Eiermann [5], is exploited.

3.1. Standard form. Let $D_\lambda = \mathrm{diag}(\lambda_{1:N})$, $D_\mu = \mathrm{diag}(\mu_{1:N})$, $D_\gamma = \mathrm{diag}(\gamma_{1:N})$, cf. (2.16). The simultaneous diagonalization of the tridiagonal blocks of the system matrix (2.9) results from the following orthogonal transformation,

$$\begin{aligned}(I \otimes U^T)\, A\, (I \otimes U) &= (I \otimes U)\, A\, (I \otimes U) \\ &= (I \otimes U)\, \mathrm{tridiag}(M_3, M_1, M_2)\, (I \otimes U) \\ &= \mathrm{tridiag}(D_\gamma, D_\lambda, D_\mu)\,.\end{aligned} \tag{3.1}$$

Thus, using (3.1), the system $Ax = b$, corresponding to the vertical wind $w = [0, 1]^T$ and the horizontal line ordering of unknowns (perpendicular to the direction of the wind), is transformed to

$$\mathrm{tridiag}(D_\gamma, D_\lambda, D_\mu)\ [(I \otimes U)x] \ = \ [(I \otimes U)b]\ .$$

We partition, similarly to the partitioning of b in (2.25), the vector x of the length N^2 into N blocks of the length N each, and form, similarly to B_H in (2.26), a corresponding N by N matrix X_H,

$$x \ = \ [x^{(1)T}, \ldots, x^{(N)T}]^T\,, \tag{3.2}$$

$$X_H \ = \ [x^{(1)}, \ldots, x^{(N)}]\,. \tag{3.3}$$

Then $x = \mathrm{vec}(X_H)$, $b = \mathrm{vec}(B_H)$, cf. [16, Definition 4.2.9] for the definition of the vec operator, and the resulting linear algebraic system can be written as

$$\mathrm{tridiag}(D_\gamma, D_\lambda, D_\mu)\, \mathrm{vec}(U X_H) \ = \ \mathrm{vec}(U B_H)\,. \tag{3.4}$$

In the previous notation the index H indicates the horizontal line ordering of unknowns. The system (3.4) is block tridiagonal with diagonal blocks.

Consider now the permutation of the rows and columns of $\mathrm{tridiag}(D_\gamma, D_\lambda, D_\mu)$ corresponding to the natural vertical line ordering of unknowns. This ordering, which is parallel to the direction of the wind, can be conveniently described using the block matrices X_V and B_V analogously to (3.3) and (2.26) in the horizontal case. Clearly,

$$\mathrm{vec}(X_V) \ = \ \mathrm{vec}(X_H^T) \ = \ P\, \mathrm{vec}(X_H), \qquad \mathrm{vec}(B_V) \ = \ \mathrm{vec}(B_H^T) \ = \ P\, \mathrm{vec}(B_H)\,,$$

where

$$P = [I \otimes e_1, \ldots, I \otimes e_N], \quad P = P^T, \quad P^2 = I \tag{3.5}$$

represents the permutation matrix which transforms the horizontal line ordering into the vertical line ordering, see [16, Theorem 4.3.8]. With the (orthogonal) transformation of the basis represented by P, (3.4) gets the form

$$T\, \mathrm{vec}(X_H^T U) \ = \ \hat{b}\,, \qquad \hat{b} \ \equiv \ \mathrm{vec}(B_H^T U)\,, \tag{3.6}$$

$$T \ \equiv \ \mathrm{diag}(T_{1:N}), \quad T_j \ \equiv \ \mathrm{tridiag}(\gamma_j, \lambda_j, \mu_j), \ \ j = 1, \ldots, N\,. \tag{3.7}$$

This N^2 by N^2 system represents nothing but N independent linear systems of the size N by N,

$$T_j\ [X_H^T u_j] \ = \ \hat{b}^{(j)}\,, \qquad \hat{b}^{(j)} \ \equiv \ B_H^T u_j\,, \quad j = 1, \ldots, N\,. \tag{3.8}$$

Summarizing, with the permutation of unknowns represented by the permutation matrix P, the system (3.4) decomposes into N systems of the form (3.8) for the unknowns $X_H^T u_j$, which, for a given j, represent the jth Fourier coefficients of the blocks $x^{(1)}, \ldots, x^{(N)}$ in the basis $u_1, \ldots, u_N$, see [8], [9]. This connection to the Fourier transformation in the direction perpendicular to the wind will appear in a very natural way in the following subsection.

3.2. Kronecker product form. With (2.12), and using the mixed-product property of the Kronecker product, see, e.g., [16, Lemma 4.2.10], (3.1) is equivalent to

$$(I \otimes U)\, A\, (I \otimes U) \;=\; \nu M \otimes (UKU) + ((\nu + \delta h)\, K + C) \otimes (UMU)\,.$$

Thus, the Fourier transformation $(I \otimes U)x = \mathrm{vec}\,(UX_H)$ of the blocks of unknowns $x^{(1)}, \ldots, x^{(N)}$ in the direction perpendicular to the wind means diagonalization of the matrices K and M on the right hand side of the Kronecker products in (2.12). Then

$$\text{(3.9)}\; P\,(I \otimes U)\, A\, (I \otimes U)\, P \;=\; \nu(UKU) \otimes M + (UMU) \otimes ((\nu + \delta h) K + C) \;=\; T\,,$$

and (3.8) immediately follows.

With the *vertical line ordering* of the original unknowns (parallel to the direction of the wind) the discretized system can be written as $(PAP)\, Px = Pb$. The block diagonalization of the matrix

$$PAP \;=\; \nu K \otimes M + M \otimes ((\nu + \delta h) K + C)$$

then gives

$$(U \otimes I)\, PAP\, (U \otimes I) \;=\; T\,,$$

and (3.8) immediately follows. The equivalence of both approaches can easily be seen from rewriting

$$(P(I \otimes U)\, A\, (I \otimes U) P)\, (P(I \otimes U)x) \;=\; P(I \otimes U)\, b\,,$$

which summarizes the approach with using the horizontal line ordering (perpendicular to the wind direction), as

$$(P(I \otimes U)P)\, (PAP)\, (P(I \otimes U)P)\, ((P(I \otimes U)P)\, Px) \;=\; (P(I \otimes U)P)\, Pb\,,$$

which, considering that $P(I \otimes U)P = U \otimes I$, summarizes the approach with using the vertical line ordering (parallel to the wind direction). The last approach has been used, apart from rotating the domain by $\frac{\pi}{2}$, by Ernst in [13] and by Eiermann and Ernst in [6].

4. Initial period of slow convergence.

As indicated in the Introduction, when GMRES is applied to linear systems $Ax = b$ resulting from the SUPG discretization of (2.1), it typically exhibits an initial period of slow convergence . This behavior is illustrated in Fig. 1.1, which shows the relative GMRES residual norms, $\|r_n\|/\|r_0\|$, for the *fixed* discretized operator $A = A(1/16, 0.01, 0.34)$, cf. (2.9), and the 15 *different* right hand side vectors b resulting from the boundary conditions (2.29). The kth boundary condition corresponds to the initial period of slow convergence lasting for $N - 1 - k$ steps. In this section we will analyze why this happens. As explained above, we restrict our discussion to the choice $\delta = \delta_*$.

The structure and numerical entries of the system matrices A, and the structure of the corresponding right hand side vectors b, were described in Section 2. Section 3 then showed that the system $Ax = b$ can be orthogonally transformed into (3.6)–(3.7). Consequently, we obtain (with $x_0 = 0$) the following lower bound on the GMRES residual norms,

$$\|r_n\|^2 = \min_{p\in\pi_n} \| p(A)\, b \|^2 \tag{4.1}$$

$$= \min_{p\in\pi_n} \| p(T)\, \hat{b} \|^2 \tag{4.2}$$

$$= \min_{p\in\pi_n} \sum_{j=1}^{N} \| p(T_j)\, \hat{b}^{(j)} \|^2 \tag{4.3}$$

$$\geq \sum_{j=1}^{N} \min_{p\in\pi_n} \| p(T_j)\, \hat{b}^{(j)} \|^2 \,. \tag{4.4}$$

In the step from (4.1) to (4.2) we exploit orthogonality of the transformation from $Ax = b$ to (3.6)–(3.7). The next step from (4.2) to (4.3) reflects the fact that the system (3.6)–(3.7) decomposes into N independent systems of the form (3.8). Finally, (4.4) 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 (3.8). Since each of these systems is of the order N, the lower bound (4.4) is equal to zero (and hence useless) for $n = N$ (and possibly even earlier). However, when there is at least one system (3.8) for which GMRES shows an initial period of slow convergence, the lower bound (4.4) shows that GMRES for the coupled system (3.6)–(3.7) 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 the 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 (4.4) 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] and [10]. In the following we will use results from our paper [18] which is devoted to this subject. We are, however, not going to present all details of the analysis here – for these and for references to other related work we refer an interested reader to [18]. We first present two numerical experiments illustrating (4.2)–(4.4).

Experiment 4.1. Using the parameter values $h = 1/16$, $\nu = 0.01$ and $\delta = \delta_* = 0.34$, we set up a linear system of the form (3.6)–(3.7). For the right hand side we use the boundary conditions (2.29) 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 Fig. 4.1. We also apply GMRES independently to each of the $N = 15$ linear systems (3.8), and plot the resulting squared residual norms by the dashed lines in Fig. 4.1. The labels on these dashed lines correspond to the indices $j = 1, \ldots, 15$ of the individual systems (3.8). The plus signs show the sums of the individual dashed curves, i.e. the lower bound (4.4). Fig. 4.2 shows a 3D plot of the computed solution. □

Experiment 4.2. We use the same parameters as in Experiment 4.1, but for the computation of the right hand side we here use (2.29) with $k = 7$. Figs. 4.3 and 4.4 show the results analogous to Figs. 4.1 and 4.2. □

As in Fig. 1.1, the initial phase of slow GMRES convergence in Figs. 4.1 and 4.3 lasts $N - k - 1$ steps (14 steps for $k = 0$ and 7 steps for $k = 7$). We will show below why such initial behavior of GMRES is *typical* for our problem class and the

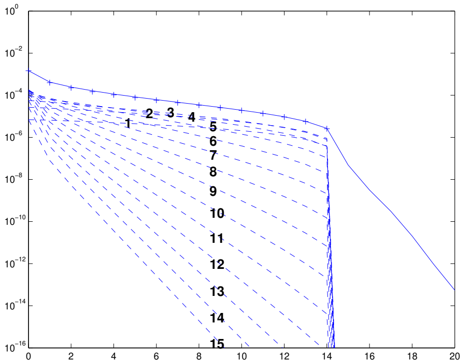


FIG. 4.1. *Squared GMRES residual norms for (3.6)–(3.7) with $\hat{b}$ from (2.29) with $k = 0$ (solid) and for each system (3.8) individually (dashed), and the lower bound (4.4) (+). System parameters are $h = 1/16$, $\nu = 0.01$, $\delta = \delta_* = 0.34$.*

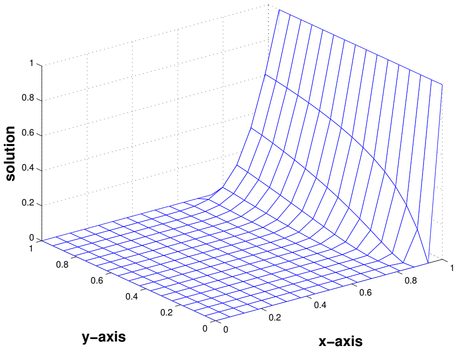


FIG. 4.2. *The solution corresponding to Experiment 4.1, i.e. the boundary conditions (2.29) with $k = 0$.*

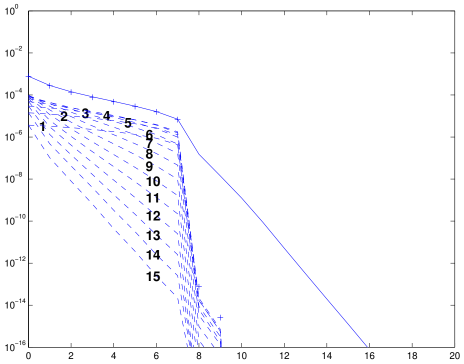


FIG. 4.3. *Results analogous to Fig. 4.1, but with $k = 7$ in (2.29).*

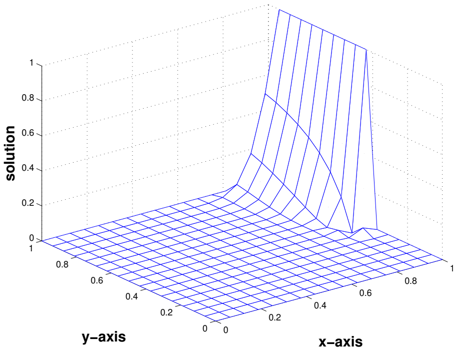


FIG. 4.4. *The solution corresponding to Experiment 4.2, i.e. the boundary conditions (2.29) with $k = 7$.*

boundary conditions (2.29) (using these boundary conditions appears convenient for explanation of the basic idea of our analysis; extension to other boundary conditions is straightforward). The parameters h, ν, and δ chosen in both Experiments 4.1 and 4.2 yield the matrices T_j, $j = 1, \ldots, N$, with the absolute values of the entries γ_j, λ_j, μ_j shown in Fig. 2.1. Apparently, the slow initial convergence occurs only for the individual systems (3.8) with a small index j, when (2.24) holds. Finally, we observe that throughout the initial phase of slow convergence the lower bound (4.4) is in both experiments very sharp.

Skipping details, our results in [18] 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 (3.8), and denote

$$(4.5) \qquad \hat{b}^{(j)} \;=\; B_H^T u_j \;\equiv\; [\rho_1^{(j)}, \ldots, \rho_N^{(j)}]^T\,, \qquad \tau_j \;\equiv\; \frac{\lambda_j}{\gamma_j}\,, \quad \text{and} \quad \zeta_j \;\equiv\; \frac{\mu_j}{\gamma_j}\,.$$

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 [18, Theorem 3.2],

$$\|\hat{r}_n^{(j)}\| = \min_{p\in\pi_n} \|p(T_j)\,\hat{b}^{(j)}\| \tag{4.6}$$

$$= \left\| [1, -\tau_j, \ldots, (-\tau_j)^n]\, \left[\hat{b}^{(j)}, (S^T+\zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T+\zeta_j S)^n\,\hat{b}^{(j)}\right]^+ \right\|^{-1} \tag{4.7}$$

$$\geq \left(\sum_{m=0}^{n} |\tau_j|^{2m}\right)^{-\frac{1}{2}} \sigma_{\min}\left(\left[\hat{b}^{(j)}, (S^T+\zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T+\zeta_j S)^n\,\hat{b}^{(j)}\right] \right). \tag{4.8}$$

Here X^+ denote 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 (4.7) can be simplified. Let

$$\left[\hat{b}^{(j)}, (S^T+\zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T+\zeta_j S)^{N-l}\,\hat{b}^{(j)}\right]^T \equiv [O, R_j] + \zeta_j\, P_j, \tag{4.9}$$

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 [18, 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 [18]). Assuming that $|\zeta_j|\|R_j^{-1}P_j\| < 1$ holds, [18, Theorem 3.3 and Theorem 2.1] give

$$\|\hat{r}_{N-l}^{(j)}\| = \min_{p\in\pi_{N-l}} \|p(T_j)\,\hat{b}^{(j)}\| \tag{4.10}$$

$$= \left\| \left([O, I] + \zeta_j R_j^{-1} P_j\right)^+ R_j^{-1} \left[1, -\tau_j, \ldots, (-\tau_j)^{N-l}\right]^T \right\|^{-1} \tag{4.11}$$

$$\geq \left(\,1 - |\zeta_j|\|R_j^{-1}P_j\|\,\right) \left(\sum_{m=0}^{N-l} |\tau_j|^{2m}\right)^{-\frac{1}{2}} \sigma_{min}(R_j)\,. \tag{4.12}$$

Consequently, for $|\zeta_j|\,\|R_j^{-1}P_j\|$ significantly smaller than one and $|\tau_j| \approx 1$, see (2.24), the first and the second factor in the lower bound (4.12) are typically not

j	$\lvert\tau_j\rvert$	$\lvert\zeta_j\rvert$	$\lvert\zeta_j\rvert\,\|R_j^{-1}P_j\|$	$(\sum_{m=0}^{14}\lvert\tau_j\rvert^{2m})^{-\frac{1}{2}}$	$\sigma_{min}(R_j)$	$(4.12)/\|\hat{b}^{(j)}\|$
1	1.0052	0.0010	0.0247	0.2489	0.0003	0.0318
2	1.0209	0.0042	0.0981	0.2216	0.0007	0.0262
3	1.0481	0.0096	0.2180	0.1785	0.0010	0.0182
4	1.0881	0.0176	0.3812	0.1260	0.0013	0.0102
5	1.1431	0.0286	0.5846	0.0752	0.0015	0.0041
6	1.2162	0.0432	0.8312	0.0368	0.0016	0.0008
7	1.3116	0.0623	1.1505	0.0145	0.0017	$*$
8	1.4348	0.0870	1.6373	0.0046	0.0018	$*$
9	1.5925	0.1185	2.4596	0.0012	0.0017	$*$
10	1.7923	0.1585	3.8905	0.0002	0.0016	$*$
11	2.0409	0.2082	6.4713	4.0e-5	0.0015	$*$
12	2.3392	0.2678	11.2601	6.2e-6	0.0013	$*$
13	2.6735	0.3347	19.9683	9.7e-7	0.0010	$*$
14	3.0033	0.4007	33.8919	1.9e-7	0.0007	$*$
15	3.2564	0.4513	49.8737	6.3e-8	0.0003	$*$

TABLE 4.1

Numerical values of the quantities in (4.12) corresponding to Experiment 4.1. The stars ($*$) in the rightmost column indicate that $|\zeta_j|\,\|R_j^{-1}P_j\| \geq 1$, so that (4.12) is not applicable.

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 the Experiments 4.1 and 4.2. Fig. 4.5 shows the absolute values of the entries of the right hand side vectors in Experiment 4.1. 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 (4.12) with $l = 1$. The corresponding numerical values of the factors in (4.12) are shown in Table 4.1. The parameters chosen in Experiment 4.1 yield a moderate mesh Peclet number, $P_h = 3.125$ (cf. Experiment 2.3). Hence (2.24) 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 (4.12) is applicable. The rightmost column of Table 4.1 shows that in the first $N-1 = 14$ steps GMRES makes little progress for the individual systems (3.8) corresponding to $j = 1, 2, 3, 4$. Consequently, the slow initial convergence of GMRES when applied the coupled system (3.6)–(3.7), as well as to the original system $Ax = b$, lasts (at least) for 14 steps. This is clearly visible in Fig. 4.1.

We now explain some subtle points illustrated by the Experiment 4.2. The components of the right hand side vectors $\hat{b}^{(j)}$ are shown in Fig. 4.6. Since $\rho_6^{(j)}$ is the first nonzero entry in each $\hat{b}^{(j)}$, $j = 1, \ldots, 15$, we are tempted to apply (4.12) with $l = 6$ $(N - l = 9)$. However, note that since

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

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 (4.12) are very small for all j. This corresponds to the GMRES residual norms $\|\hat{r}_9^{(j)}\|$ for the individual systems (3.8) in Fig. 4.3.

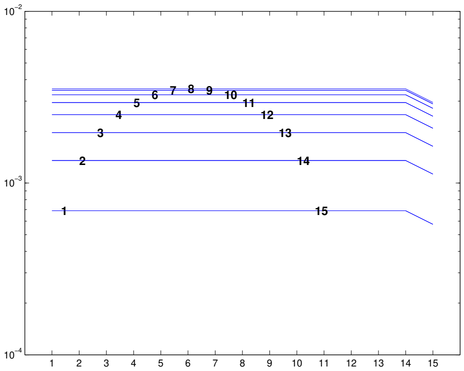


FIG. 4.5. *Absolute values of the entries in the right hand side vectors $\hat{b}^{(j)}$, $j = 1, \ldots, 15$, used Experiment 4.1.*

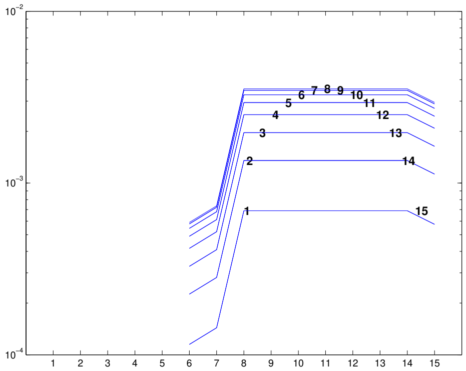


FIG. 4.6. *Absolute values of the entries in the right hand side vectors $\hat{b}^{(j)}$, $j = 1, \ldots, 15$, used Experiment 4.2.*

Since our analysis cannot be based on using (4.12) and the step $N - l$, we turn to the lower bound (4.8), which is applicable for all $n = 0, 1, \ldots, N - l$. The values of $|\tau_j|$ given in Table 4.1 are valid also for the Experiment 4.2. Hence for small j the first factor in (4.8) does not decrease significantly. Moreover, as shown in Fig. 4.7, the second factor in (4.8),

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

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 each of the vectors $\hat{b}^{(j)}$, $j = 1, \ldots, 15$. The bound (4.8) then implies that for small j the GMRES residual norms for T_j and $\hat{b}^{(j)}$ converge slowly for the first 7 steps, which is precisely what we observe in Fig. 4.3. Further numerical experiments are given in Section 7.

In summary, the presence of (at least) one system (3.8) satisfying (2.24), with l representing the index of the first significant entry of the corresponding right hand side, prevents fast convergence of GMRES for the coupled system (3.6)–(3.7), and therefore also for the original system $Ax = b$, for the initial $N - l$ steps. As shown in Section 2.2 the relation (2.24) 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. It remains to find why from the step $N - l + 1$ GMRES converges for this problem with an increased rate.

5. Acceleration of convergence. Let us recall the essence of the preceeding text. Using (4.1)–(4.3), the convergence analysis of GMRES for A and b arising from the SUPG discretization of the convection-diffusion problem (2.1) is transformed to the convergence analysis of GMRES for N coupled N by N tridiagonal systems (3.8). If at least one of the tridiagonal blocks satisfies (2.24), then (4.4) together with (4.8) and (4.12) proves that the GMRES convergence for the whole N^2 by N^2 system is slow for the first $N - l$ steps, where l is the first significant (in the meaning quantitatively described above) entry of the corresponding transformed block right hand side in (3.8). This result is based on [18], which relates the GMRES behavior for

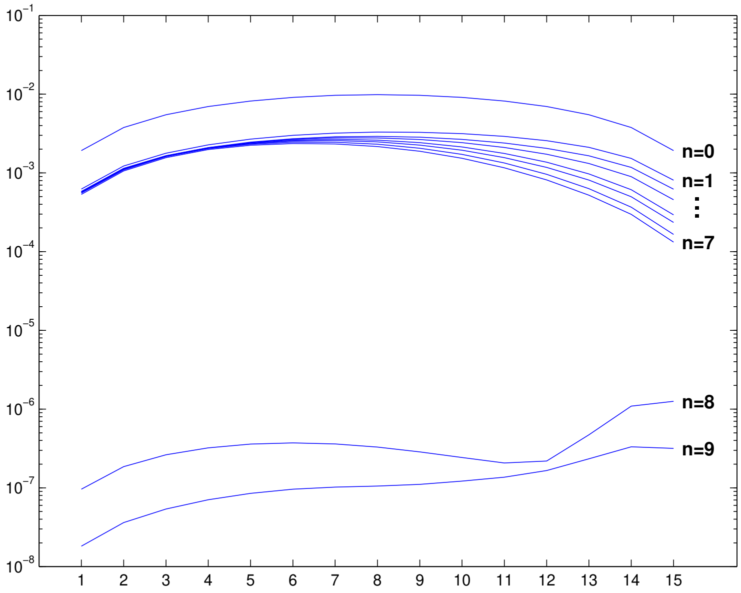


FIG. 4.7. $\sigma_{min}\left(\left[\hat{b}^{(j)}, (S^T + \zeta_j S)\,\hat{b}^{(j)}, \ldots, (S^T + \zeta_j S)^n\,\hat{b}^{(j)}\right]\right)$, *cf. (4.8), for* $j = 1, \ldots, 15$ *and* $n = 0, \ldots, 9$, *corresponding to Experiment 4.2.*

tridiagonal Toeplitz matrices to the GMRES behavior for Toeplitz lower bidiagonals (scaled Jordan blocks). Please note that the result is based on a simple fact – when GMRES behaves poorly for a single N by N system from (3.8), it must behave poorly for the whole coupled N^2 by N^2 system.

Any quantitative description of a possible acceleration of convergence *after the step* $N - l$ is extremely difficult. In order to avoid misunderstandings, we wish to stress that here we mean by 'acceleration of convergence' the behavior immediately succeeding the period of slow convergence, i.e. the behavior in the steps $N - l + 1, N - l + 2, \ldots$. Similarly as in the other parts of this paper, asymptotic convergence bounds based on the operator (system matrix) offer only very little help (if any) in solving this question.

In order to understand the difficulty we consider, for simplicity, a block diagonal matrix consisting of N lower bidiagonal Toeplitz blocks (scaled Jordan blocks) of size N by N, which all correspond to the same eigenvalue λ. Let the corresponding block right hand sides of length N have their first significant entries in the lth positions. When for at least one of the individual Toeplitz blocks the subdiagonal entry is close in magnitude to λ, then Section 4 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, it 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 since finding the exact solution will only take one additional step.

In the case (3.8) we do have N coupled systems like in the preceeding simple example. However,

- the coupled systems are tridiagonal (not bidiagonal) Toeplitz;
- the minimal polynomial of the coupled system can not be reduced to the minimal polynomial of a single N by N block, because the eigenvalues of the different blocks are different.

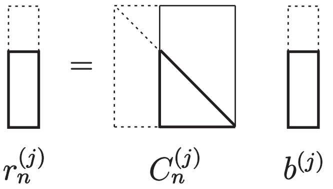
If the superdiagonal of T could be considered "a perturbation" of its lower bidiagonal part, i.e. if (2.24) holds, then the first difficulty could (even quantitatively) be overcome. A possible approach for that could be based on [18, Theorem 3.1], which describes the explicit mapping from $\hat{b}$ ($x_0 = 0$) to r_n,

$$r_n \;=\; p_n(T)\,\hat{b} = \Big[(p_n(T_1)\,\hat{b}^{(1)})^T, \ldots, (p_n(T_N)\,\hat{b}^{(N)})^T\Big]^T . \tag{5.1}$$

For each $j = 1, \ldots, N$, the matrix $C_n^{(j)}$ representing the mapping from $\hat{b}^{(j)}$ to $r_n^{(j)}$, $r_n^{(j)} = C_n^{(j)}\hat{b}^{(j)}$, has no more than $2n+1$ nonzero diagonals. A simple extension of the referenced theorem shows that the subdiagonals $c_n^{(-d)}$ and the superdiagonals $c_n^{(d)}$ of $C_n^{(j)}$ are related by the formula (here we omit the superscript and index j)

$$c_n^{(d)} \;=\; (\zeta_j)^d\, c_n^{(-d)}, \quad \zeta_j = \frac{\mu_j}{\gamma_j}\,.$$

Thus, if (2.24) holds, meaning $|\zeta_j| \ll 1$, the relevance of superdiagonals decreases, in comparison to subdiagonals, with powers of ζ_j with the distance from the main diagonal. Consider now $\hat{b}^{(j)}$, $j = 1, \ldots, N$, with l being the index of the first significant entry in any of them. Then the only columns of $C_n^{(j)}$, $j = 1, \ldots, N$, which become significant are those from l to N (the other columns are in the formula for $r_n^{(j)}$ multiplied by zeros or small quantities),



with the lower triangular part (including the diagonal) dominating in magnitude the upper trapezoidal part. Therefore, from this point of view, one can only expect acceleration of convergence after the dominating part is completely filled, i.e. after the step $N-l$. Until this step the lower triangular part gets new nonzero subdiagonals at every step, which makes any convergence acceleration difficult.

This point of view, however, does not consider the second difficulty, namely the differences between the individual blocks T_j and $\hat{b}^{(j)}$ (the latter being in our problem unimportant). The acceleration of convergence *assumes* that the entries in the mappings $C_j^{(n)}$ uniformly (for *all* j, with some possible little variations) and substantially decrease beginning from the step $N-l+1$. This, however, depends on how much the eigenvalues of the individual blocks differ, see [18, relation (3.7)]. The last point can not, to our opinion, be easily quantified. We illustrate the interplay of both factors in the following experiment.

Experiment 5.1. Consider Experiment 2.3, see Figs. 2.1 and 2.2. For the large mesh Peclet number $P_h = 312.5$, condition (2.24) is perfectly satisfied for all j

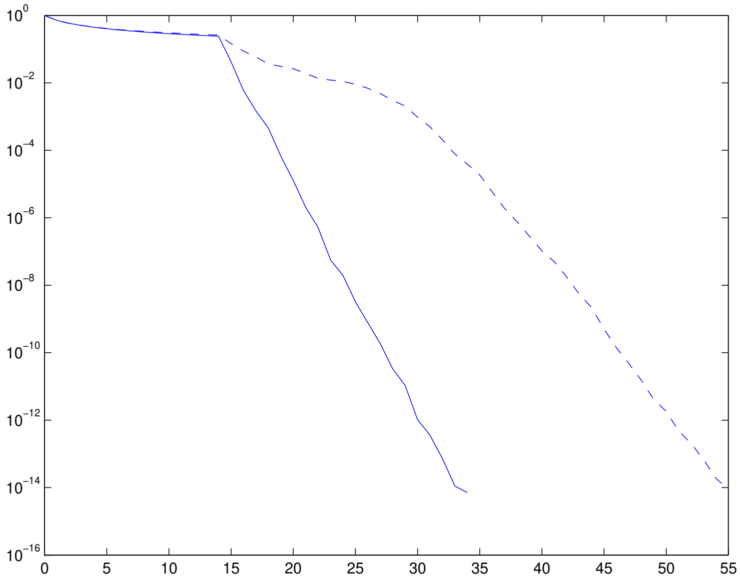


FIG. 5.1. *Relative GMRES residual norms for the systems (3.6)–(3.6) with $h = 1/16$, $\nu = 0.01$ (solid) and $\nu = 0.0001$ (dashed), the respective values $\delta = \delta_*$ ($\delta = 0.34$ for $\nu = 0.01$ and $\delta = 0.4984$ for $\nu = 0.0001$), and the boundary conditions (2.27)–(2.28).*

(cf. Fig. 2.2) and, with our argument above, the coupled system (3.6)–(3.7) can indeed, with a small inaccuracy, be considered as N coupled lower bidiagonal systems. 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 can not be compensated for by smaller differences between the individual γ_j respectively μ_j. In this case we can therfore expect that in the few steps following 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 Fig. 5.1 which compares the relative GMRES residual norms $\|r_n\|/\|r_0\|$ for these two systems corresponding to the right hand sides from the boundary conditions (2.27)–(2.28). □

6. Eigendecomposition and convergence analysis.

In this section we focus on the eigendecomposition of the system matrix A (given by (2.6), (2.9)), and give some details on the complicated relationship between the eigenvalues of A and the convergence of GMRES.

As A is unitarily similar to the block tridiagonal matrix T with tridiagonal Toeplitz blocks T_j,

$$A \;=\; (I \otimes U)\, P\, \mathrm{diag}(T_{1:N})\; P\, (I \otimes U)\,, \tag{6.1}$$

see (3.9), the existence and form of its eigendecomposition are determined by (6.1) and the existence and form of the eigendecompositions of 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 2.1, then the N distinct eigenvalues of T_j are given by

$$\sigma_{jk} \;=\; \lambda_j + (\gamma_j \mu_j)^{1/2}\, \omega_k\,, \quad \omega_k \;\equiv\; 2\cos(kh\pi)\,, \quad k = 1, \ldots, N\,, \tag{6.2}$$

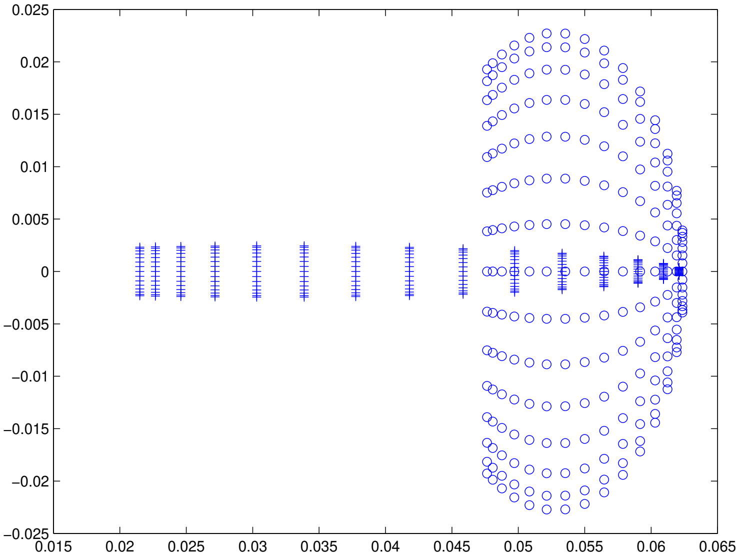


FIG. 6.1. *Eigenvalues σ_{jk}, $j,k = 1,\ldots,15$, cf. (6.2), of the matrices $A(1/16, 0.01, 0.34)$ (o) and $A(1/16, 0.0001, 0.4984)$ (+), which correspond to λ_j, μ_j and γ_j, $j = 1,\ldots,15$, shown in Figs. 2.1 and 2.2, respectively.*

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., [26, pp. 113–115]. Clearly, when $|\zeta_j| = |\mu_j/\gamma_j| \ll 1$, the eigenvectors of T_j are ill conditioned.

Using (6.1), the N^2 eigenvalues of A obviously are the values σ_{jk} for $j,k = 1,\ldots,N$. Furthermore, elementary algebra using (3.5) and the mixed-product property of the Kronecker product shows that a unit norm eigenvector corresponding to σ_{jk} is

$$(6.3) \qquad y_{jk} \;=\; \chi_{jk}\,[\Delta_j u_k] \otimes u_j\,, \quad j,k = 1,\ldots,N\,,$$

where $\chi_{jk} \equiv \|[\Delta_j u_k] \otimes u_j\|^{-1}$. We denote the resulting eigenvector matrix of A by

$$(6.4) \qquad Y \;\equiv\; [y_{11}, y_{12},\ldots,y_{1N},\ldots,y_{N1},y_{N2},\ldots,y_{NN}] \;\equiv\; [Y_1,\ldots,Y_N]\,.$$

EXPERIMENT 6.1. We use (6.2) to compute the eigenvalues of the matrices $A(15, 0.01, 0.34)$ and $A(15, 0.0001, 0.4984)$, cf. (2.9), and plot these values in Fig. 6.1. As explained in Section 5, for both these matrices and the respective b from the boundary conditions (2.27)–(2.28) the slow initial convergence of GMRES lasts 14 $(N-1)$ steps, cf. Fig. 5.1. The subsequent convergence is much faster for $\nu = 0.01$, i.e. $P_h = 3.125$. This fact is in Section 5 related to the spectrum (more accurately, to the differences between *the real parts* of the individual eigenvalues). However, it is not at all obvious from the shapes of the spectra shown in Fig. 6.1. □

Since A is diagonalizable we could have based our convergence analysis of GMRES in the previous sections on its eigendecomposition. In particular, we could have applied the standard GMRES convergence bound [25, Proposition 4],

$$\begin{aligned} \|r_n\| &= \min_{p\in\pi_n} \|p(A)r_0\| \\ &\le \kappa(Y) \min_{p\in\pi_n} \max_{j,k=1,\dots,N} |\, p(\sigma_{jk})\,|\, \|r_0\|\,, \end{aligned} \tag{6.5}$$

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 (6.3) 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., [23]). Such nonnormality makes the direct application of (6.5) rather useless for proving well-justified 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 (for unpublished notes in this direction see [11], [12]).

In the following we will describe some details about the ill-conditioning of Y. First, note that

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

which gives the following.

Proposition 6.2. *The eigenvectors of A in the ordering given by (6.4) 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}, $k = 1, \dots, N$, *within* each block Y_j, $j = 1, \dots, N$, and that

$$\kappa(Y) = \max_{j=1,\dots,N} \kappa(Y_j)\,. \tag{6.6}$$

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

It follows from (6.3) that the conditioning of the block Y_j depends on the scaling imposed by the matrix Δ_j. Specifically, $\kappa(Y_j)$ 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 $\kappa(Y_j)$, $j = 1, \dots, N$, in the following table[1].

[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 (6.3) for the eigenvectors of A. Hence (6.6) does not hold in our finite precision computation.

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

In summary, the most ill-conditioned blocks Y_j in our example correspond to the tridiagonal Toeplitz systems in (3.8) that satisfy (2.24), and that are responsible for the initial phase of slow GMRES convergence. Thus, the eigendecomposition at least reveals which blocks are the most troublesome for the GMRES convergence.

7. Additional numerical experiments.

In this section we show some additional experiments to further illustrate our theoretical considerations.

EXPERIMENT 7.1. We first consider the boundary conditions defined by (2.27)–(2.28). We use the same parameters as in Experiment 4.1, i.e. $h = 1/16$, $\nu = 0.01$, and $\delta = \delta_* = 0.34$. The 3D plot of the computed solution, shown in Fig. 7.2, shows the discontinuity at the inflow boundary, and indicates the boundary layer at the outflow boundary which is caused by the boundary condition $u(x, 1) = 0$.

Fig. 7.1 shows the squared GMRES residual norms for the resulting system (3.6)–(3.7) (solid), and each individual system (3.8) (dashed lines). The labels on the dashed lines correspond to the indices $j = 1, \dots, 15$ of the individual systems (3.8). The lower bound (4.4) is plotted by the plus signs. As in Experiment 4.1, shown in Fig. 4.1, the lower bound closely approximates the GMRES residual norms over the whole initial period of slow convergence.

Note that because of the choice of h, ν and δ the only difference between the systems (3.6)–(3.7) in this experiment and in Experiment 4.1 is in the right hand sides. Fig. 7.3 shows a 3D plot of the absolute values of the coefficients in the right hand side vectors for each individual system (3.8). Each line in Fig. 7.3 corresponds to one vector $\hat{b}^{(j)} \equiv [\rho_1^{(j)}, \dots, \rho_N^{(j)}]$, $j = 1, \dots, 15$. Because of the structure of the boundary conditions (2.27)–(2.28), most of the right hand side vectors satisfy $|\rho_1^{(j)}| \gg |\rho_k^{(j)}|$ for $k = 2, \dots, 15$. Hence, in the notation of Section 4, $\rho_1^{(j)}$ is for most j (particularly for small j) the first significant entry. As explained in Section 4, this leads to $N - 1 = 14$ steps of slow initial convergence. In addition to that, this experiment confirms the subtle point about insignificant leading entries in $\hat{b}^{(j)}$. For the individual system (3.8) with $j = 11$ the initial phase of GMRES lasts only $N - 2 = 13$ steps, cf. Fig. 7.1. This corresponds to the fact that $\rho_2^{(11)}$, *not* $\rho_1^{(11)}$, is the first significant entry in $\hat{b}^{(11)}$, cf. Fig. 7.3. The index $j = 11$ is the only one for which this correspondence can be observed here. □

EXPERIMENT 7.2. In Fig. 7.4 we show the relative GMRES residual norms for the linear systems (3.6)–(3.7) with boundary conditions from (2.27)–(2.28), and for the parameters $h = 1/51$, $\nu = 0.001, 0.0005, 0.00025$, and the respective values of $\delta = \delta_*$. As predicted in Section 4, the initial phase of slow GMRES convergence in each case lasts $N - 1 = 49$ steps. In addition, similar to the observations made in Section 5, the speed of convergence decreases in the subsequent phase with increasing

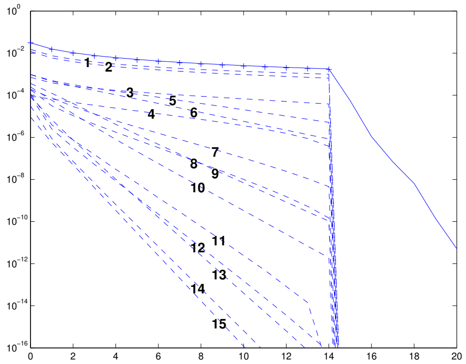


FIG. 7.1. *Squared GMRES residual norms for (3.6)–(3.7) with $\hat{b}$ from (2.27)–(2.28) (solid) and for each system (3.8) individually (dashed), and the lower bound (4.4) (+). System parameters are $h = 1/16$, $\nu = 0.01$, and $\delta = \delta_* = 0.34$.*

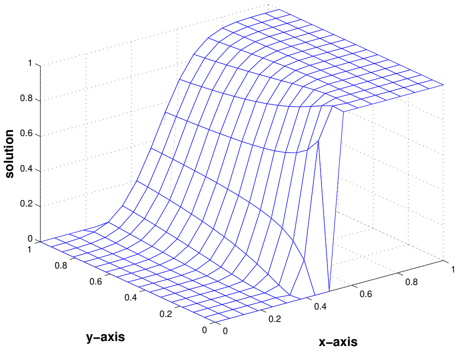


FIG. 7.2. *The solution corresponding to Experiment 7.1, i.e. the boundary conditions (2.27)–(2.28).*

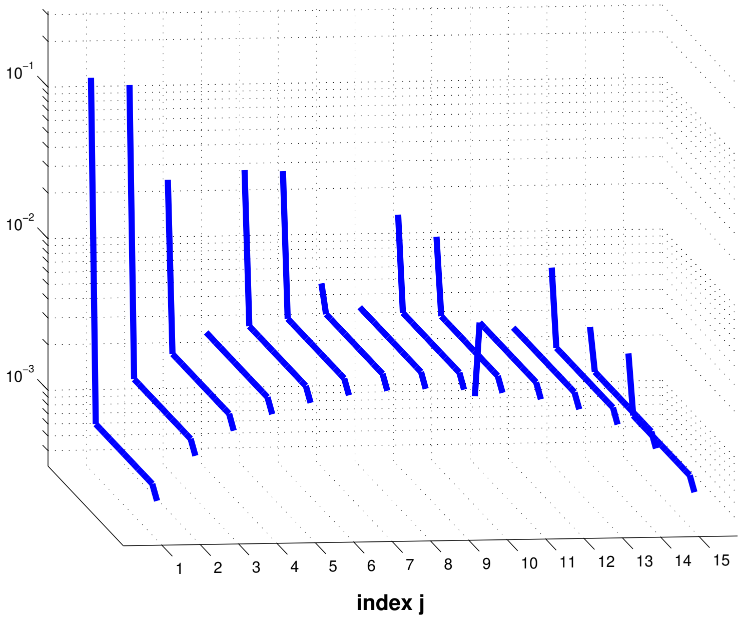


FIG. 7.3. *Absolute values of the entries in the right hand side vectors $\hat{b}^{(j)} = B_H^T u_j$, $j = 1, \ldots, 15$, corresponding to $h = 1/16$, $\nu = 0.01$, $\delta = \delta_* = 0.34$, and the boundary conditions (2.27)–(2.28).*

mesh Peclet number P_h; here $P_h = 9.8, 19.6$, and 39.2, for $\nu = 0.001, 0.0005$, and 0.00025, respectively. For smaller ν (not shown here) we observe that the GMRES residual norm curves do not differ much from the curve for $\nu = 0.00025$. □

8. Concluding remarks. This paper is devoted to the convergence analysis of GMRES applied to the SUPG discretized convection-diffusion model problem with dominating convection. Though the eigenvalues and eigenvectors of the discretized operator are known analytically, their use for for convincing and clear description of

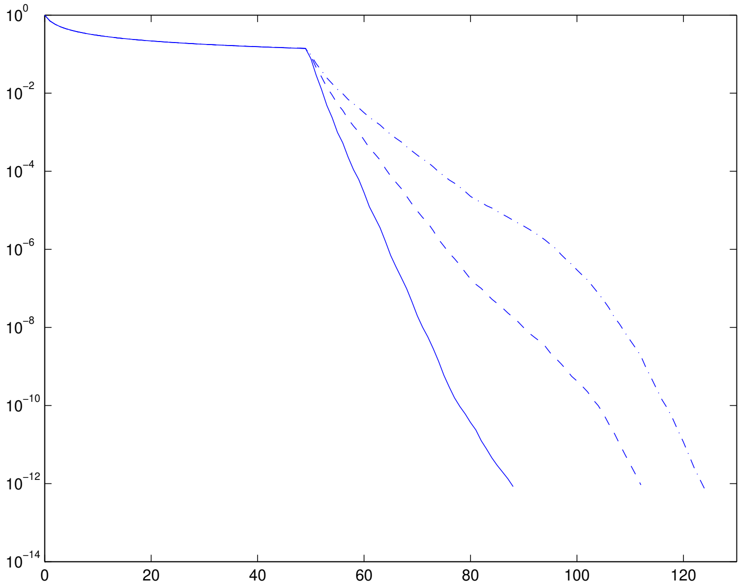


FIG. 7.4. *Relative GMRES residual norms for the linear systems (3.6)–(3.7) (solid line) with right hand sides from (2.27)–(2.28) for the parameters* $h = 1/51$, $\nu = 0.001$ *(solid),* 0.0005 *(dashed),* 0.00025 *(dash-dot), and the respective* $\delta = \delta_*$.

convergence is very difficult. The transformation to the eigenvector coordinates is highly ill-conditioned. Therefore any analysis based on it must involve a rather complicated pattern of cancellation of potentially huge components of the initial residual (right hand side) in the individual eigenspaces, otherwise the results are quantitatively useless. Instead of following this technically complicated and physically unnatural way, we propose another idea.

Assume that a linear algebraic system can be transformed using a well-conditioned transformation of the basis to a new system with a structure, not necessarily with a diagonal system matrix, for which the GMRES convergence can easily be analyzed. 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.

In our paper the transformation is orthonormal and the transformed system is block diagonal with tridiagonal Toeplitz blocks. The GMRES convergence for individual tridiagonal Toeplitz systems is then analyzed by making a link to the GMRES convergence for scaled Jordan blocks, which is possible due to the dominance of convection over diffusion in the model problem. This approach clearly describes the relationship between the boundary conditions in the model problem and the initial phase of slow GMRES convergence for the discretized algebraic system. These results rely heavily on the simple special properties of the original PDE model problem. They can perhaps be considered as an example of a successful convergence analysis revealing, as a byproduct, a possibly very complicated relationship of the eigeninformation and GMRES convergence. We believe, however, that it is worth to examine the idea of well-conditioned transformation to some easy-to-analyze structure also in more general context.

Our results can be extended to a 3D model problem described by Ramage [22] as well as to other separable second order PDEs on rectangular domains. From the other side, 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, the convergence behavior would not change much. The approach can be extended to structures with more nonzero diagonals. Some of the ideas presented in the paper can be used for analysis of the effects of diagonal scaling and possibly other preconditioning. We hope to return to some of these points in our future 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 wish to thank also Mark Embree for sharing his unpublished notes [11] and [12] and Howard Elman, Dianne O'Leary and Nick Trefethen for helpful discussions.

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*, ETNA, 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. FENOMECH '81, Part I (Stuttgart, 1981).

[5] M. Eiermann, *Semiiterative Verfahren für nichtsymmetrische lineare Gleichungssysteme*, Habilitationsschrift, Universität Karsruhe, (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 in 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*, Mathematics of Computation, 72 (2001), pp. 263–288.

[9] ———, *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] ———, *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*, Computer Methods in Applied Mechanics and Engineering, 179 (1999), pp. 179–195.

[15] A. Greenbaum and Z. Strakoš, *Matrices that generate the same Krylov residual spaces*, in Recent Advances in Iterative Methods, G. H. Golub, A. Greenbaum, and M. Luskin, eds., vol. 60 of The IMA volumes in mathematics and its applications, Springer-Verlag, New York, 1994, pp. 95–118.

[16] R. A. Horn and C. R.Johnson, *Topics in matrix analysis*, Cambridge University Press, 1991.

[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), vol. 34 of AMD, Amer. Soc. Mech. Engrs. (ASME), New York, 1979, pp. 19–35.

[18] J. Liesen and Z. Strakoš, *Convergence of GMRES for tridiagonal Toeplitz matrices*, submitted to SIAM J. Matrix Anal. Appl., (2003).

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

[20] C. Paige and Z. Strakoš, *Residual and backward error bounds in minimum residual krylov subspace methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1899–1924.

[21] G. D. Raithby, *Skew upstream differencing schemes for problems involving fluid flow*, Computer Methods in Applied Mechanics and Engineering, 9 (1976), pp. 153–164.

[22] A. Ramage, *A note on parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems in three dimensions*, Tech. Rep. 32, University of Strathclyde, Department of Mathematics, 1998.

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

[24] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical methods for singularly perturbed differential equations*, vol. 24 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1996.

[25] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[26] G. D. Smith, *Numerical solution of partial differential equations*, The Clarendon Press Oxford University Press, New York, second ed., 1978. Finite difference methods, Oxford Applied Mathematics and Computing Science Series.

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

[28] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[29] D. M. Young, *Iterative methods for solving partial difference equations of elliptic type*, PhD thesis, Harvard University, Cambridge, Massachusetts, 1950.

[30] D. M. Young, *Iterative solution of large linear systems*, Academic Press, New York, 1971.

[31] I. Zavorin, D. P. O'Leary, and H. C. Elman, *Complete stagnation of GMRES*, Linear Algebra Appl., 367 (2003), pp. 165–183.