Noname manuscript No.
(will be inserted by the editor)

Analysis of the second phase of the GMRES convergence for a convection-diffusion model problem

J. Duintjer Tebbens · J. Liesen · Z. Strakoš

the date of receipt and acceptance should be inserted later

Abstract It is well known that GMRES applied to linear algebraic systems arising from a convection-diffusion model problem that has been discretized by the streamline upwind Petrov-Galerkin (SUPG) method, typically displays two distinct phases of convergence: a slow initial phase followed by convergence acceleration in the second phase. This paper complements the known results on the length of the initial phase by analyzing how the acceleration in the second phase is related to the mesh Peclet number and the choice of the stabilization parameter in the SUPG discretization. The analysis is based on some new expressions and bounds for the GMRES residuals, which can be of general interest.

Keywords convection-diffusion problem, SUPG discretization, SDFEM discretization, GMRES, convergence bounds

Mathematics Subject Classification (2000) 65F10, 65F15, 65N22, 65N30

1 Introduction

We study the convergence of the GMRES method [20] when applied to discretizations of a convection-diffusion model problem of the form

$$-\nu\Delta u + w \cdot \nabla u = f \ \text{ in } \ \Omega = (0,1)\times(0,1), \quad u = g \ \text{ on } \ \partial\Omega, \tag{1.1}$$

The work of J. Duintjer Tebbens is a part of the Institutional Research Plan AV0Z10300504 and it was supported by the project M100300901 of the institutional support of the ASCR.

The work of J. Liesen was supported by the Heisenberg-Programm of the Deutsche Forschungsgemeinschaft.

The work of Z. Strakoš is a part of the research project MSM0021620839 and it was supported by the GACR grant 201/09/0917.

Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vod. věží 2, 182 07 Prague, Czech Republic, E-mail: tebbens@cs.cas.cz · Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, E-mail: liesen@math.tu-berlin.de · Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Prague, Czech Republic, E-mail: strakos@karlin.mff.cuni.cz

where $\nu > 0$ is a scalar diffusion coefficient, and w is a vector velocity field called the "wind". We are interested in the *convection dominated case* $\nu < \|w\|$. In this case special discretization techniques must be used to avoid non-physical oscillations of the numerical solution; see the books [16,19] or the more recent survey paper [23]. Here we consider a discretization by the popular streamline-diffusion finite element method (SDFEM), also called the streamline upwind Petrov-Galerkin (SUPG) method [1,9]. In this method the stabilization consists of an additional diffusion term of the form $\langle \hat{\delta} w w^T \nabla u, \nabla v \rangle$ that is added to the weak formulation of (1.1). Here $\hat{\delta} > 0$ is the *stabilization parameter*, and the diffusity tensor $\hat{\delta} w w^T$ means that the stabilization only acts in the direction of the flow (hence the name "streamline-diffusion" method). We use a *constant vertical wind* $w = [0, 1]^T$, and the SUPG method with bilinear finite elements on a regular grid with N inner nodes in each direction. The same problem has been studied in several publications; see, e.g., the papers [3,4,6,7] or the book [5, Chapter 3].

When GMRES is applied to the discretized model problem, the method's convergence measured by the residual norm typically exhibits two distinct phases: An initial phase of at most N steps where convergence is very slow, and a second phase of faster convergence (see the solid curves in Figure 3.2 below for a preview). Ernst [6, p. 1094] conjectured that the duration of the initial phase is governed by the time it takes for boundary information to pass from the inflow boundary across the domain following the streamlines of the velocity field, which was confirmed by the analysis in [15]. This analysis, however, does not apply to the second phase of convergence.

The convergence rate of GMRES in the second phase was addressed by Fischer et al. [7]. They investigated the asymptotic convergence factor related to the standard GMRES residual bound given, for a diagonalizable matrix $A = W \operatorname{diag}(\sigma_1, \ldots, \sigma_N) W^{-1}$ by

$$\frac{\|r_k\|}{\|r_0\|} \leq \kappa(W) \min_{p \in \Pi_k} \max_j |p(\sigma_j)|, \tag{1.2}$$

where Π_k denotes the set of polynomials of degree at most k having value one at zero, and $\kappa(W) = \|W\| \|W^{-1}\|$ is the condition number of the eigenvector matrix W. If S is any compact set that contains all eigenvalues (but excludes zero), the asymptotic convergence factor associated with S is given by

$$\rho(S) \equiv \lim_{k \to \infty} \left(\min_{p \in \Pi_k} \max_{z \in S} |p(z)| \right)^{1/k};$$

see [7]. In their analysis, Fischer et al. considered S being the smallest ellipse containing the eigenvalues. They showed numerically that the asymptotic convergence factor $\rho(S)$ is strongly correlated with the actual GMRES behavior for a wide range of stabilization parameters. In particular, they observed that "optimal stabilization" yields nearly the smallest asymptotic convergence factor as well as the fastest GMRES convergence.

Any analysis based on the asymptotic convergence factor discards the eigenvectors and their condition number. In case of a convection-dominated problem this number is typically very large, and it grows considerably when the stabilization parameter $\hat{\delta}$ approaches the optimal $\hat{\delta}_*$ (see equations (2.3) and (4.1) below for a preview). Although the convergence bound (1.2) can be pessimistic, it seems somewhat counterintuitive that optimal stabilization yields the nearly fastest GMRES convergence (for a discussion of this phenomenon in a slightly different context see [12,13]). Fischer et al. found

the correlation between the GMRES convergence and the asymptotic convergence factor "surprisingly good" [7, p. 191], but they did not analyze why the eigenvectors can be discarded in the analysis.

This paper continues and complements the work in [15] in two ways. First, we explain how the convergence acceleration of GMRES in the second phase depends (in case of optimal stabilization) on the mesh Peclet number (Section 3). Second, we demonstrate that the eigenvectors of the discretized operator have little influence on the convergence of GMRES, and we discuss the role of the eigenvalues (Section 4). Our investigation uses some new expressions and bounds for the GMRES residuals. They might be of general interest beyond their immediate application in the context of the considered model problem.

2 Properties of the discretized model problem

We will now summarize the properties of the SUPG discretization of the model problem (1.1) that are most relevant for the analysis in this paper. For further details on the discretization we refer to the detailed descriptions in [3, 4, 7, 15].

The SUPG discretization of (1.1) leads to a discrete operator of the form

$$\nu A_d + A_c + \hat{\delta} A_s, \tag{2.1}$$

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. The functions ϕ_j are the bilinear finite element nodal basis functions for the $N \times N$ grid with spacing $h = 1/(N+1)$, and $\langle \cdot, \cdot \rangle$ denotes the L^2 inner product on Ω.

We consider a convection-dominated problem with the wind given by $w = [0, 1]^T$. Then the *mesh Peclet number*

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

is typically greater than one, and the stabilization parameter $\hat{\delta}$ is chosen as

$$\hat{\delta} = \delta h, \tag{2.3}$$

where δ is a *tuning parameter* between zero and one. A possible choice for δ, which is optimal from the viewpoint of a 1D discretization, is

$$\delta_* = \frac{1}{2}\left(1 - \frac{1}{P_h}\right); \tag{2.4}$$

for details see [5, Section 3.2.2]. In the following we will refer to the stabilization with the tuning parameter δ_* as the *optimal SUPG-stabilization*.

Using a vertical ordering of the unknowns (i.e., parallel to the direction of the wind), and a discrete sine transformation, it can be shown that the discrete operator (2.1) of size $N^2 \times N^2$ is *orthogonally similar* to a block-diagonal matrix A_δ consisting of N nonsymmetric tridiagonal Toeplitz blocks $T_1, \ldots, T_N$,

$$A_\delta = \operatorname{diag}(T_1, \ldots, T_N), \quad T_j = \operatorname{tridiag}(\gamma_j, \lambda_j, \mu_j) \in \mathbb{R}^{N \times N}, \quad j = 1, \ldots, N.$$

We call the matrix A_δ the *SUPG-discretized operator*. The entries of the matrix T_j satisfy

$$3\lambda_j = 2\delta h c_j + 2\nu\left(6 - c_j\right), \tag{2.5}$$

$$-3\gamma_j = \left(\delta + \frac{1}{2}\right) h c_j + 2\nu\left(c_j - \frac{3}{2}\right), \tag{2.6}$$

$$-3\mu_j = \left(\delta - \frac{1}{2}\right) h c_j + 2\nu\left(c_j - \frac{3}{2}\right), \tag{2.7}$$

where $c_j = 2 + \cos j\pi h$. If $\mu_j\gamma_j \neq 0$ the matrix T_j is diagonalizable with the N distinct eigenvalues

$$\sigma_{jk} = \lambda_j + 2\sqrt{\mu_j\gamma_j}\cos kh\pi, \quad k = 1, \ldots, N, \tag{2.8}$$

and corresponding eigenvectors

$$w_{jk} = \sqrt{\frac{2}{N+1}}\left[\zeta_j^{-\frac{1}{2}}\sin kh\pi, \zeta_j^{-1}\sin 2kh\pi, \ldots, \zeta_j^{-\frac{N}{2}}\sin Nkh\pi\right]^T, \quad \zeta_j \equiv \frac{\mu_j}{\gamma_j}. \tag{2.9}$$

The condition number of the eigenvector matrix $W_j = [w_{j1}, \ldots, w_{jN}]$ is

$$\kappa(W_j) = \max\left\{|\zeta_j|^{\frac{1-N}{2}}, |\zeta_j|^{\frac{N-1}{2}}\right\}. \tag{2.10}$$

The matrix A_δ is diagonalizable if and only if all matrices T_j are diagonalizable. For given ν and N this happens except for a finite number of choices of δ (namely those giving $\mu_j\gamma_j = 0$ for some j); the corresponding block-diagonal eigenvector matrix $W_\delta = \mathrm{diag}\,(W_1, \ldots, W_N)$ then has the condition number

$$\kappa(W_\delta) = \max_{j=1,\ldots N}\kappa(W_j) = \max_{j=1,\ldots,N}\left|\frac{(\delta+\frac{1}{2})hc_j + 2\nu(c_j - \frac{3}{2})}{(\delta-\frac{1}{2})hc_j + 2\nu(c_j - \frac{3}{2})}\right|^{\frac{N-1}{2}}. \tag{2.11}$$

3 Analysis of the second phase of GMRES for optimal SUPG-stabilization

In this section we will analyze the second phase of the GMRES convergence for the case of optimal SUPG-stabilization, i.e., $\delta = \delta_*$.

3.1 General results for the GMRES residuals

We will start with some general results on the GMRES residuals and their norms. The first result introduces a commutative splitting.

Theorem 3.1 *Consider $A \in \mathbb{C}^{N\times N}$, $r_0 \in \mathbb{C}^N$. Let $B, C \in \mathbb{C}^{N\times N}$ be such that*

$$A = B + C, \quad \textit{where} \quad BC = CB, \;\; Cr_0 = \tau\, r_0 \tag{3.1}$$

for some scalar τ. Then for $k = 1, 2, \ldots$,

$$\mathrm{span}\{r_0, Ar_0, \ldots, A^k r_0\} = \mathrm{span}\{r_0, Br_0, \ldots, B^k r_0\},$$

and, in particular,

$$\mathrm{Rank}([r_0, Ar_0, \ldots, A^k r_0]) = \mathrm{Rank}([r_0, Br_0, \ldots, B^k r_0]).$$

Proof Since B and C commute,

$$A^\ell r_0 = (B+C)^\ell\, r_0 = \left(\sum_{j=0}^{\ell} \binom{\ell}{j} B^{\ell-j} C^j\right) r_0 = \left(\sum_{j=0}^{\ell} \binom{\ell}{j} \tau^j B^{\ell-j}\right) r_0, \qquad (3.2)$$

which proves the statement. □

In (3.1) the matrix A is split into two commuting matrices B and C, while the given vector r_0 is an eigenvector of one of these two matrices (here denoted as C). This is always possible; consider, e.g., $A = (A - \tau I) + \tau I$, which will be used in Section 3.2 below.

Let the kth GMRES residual for A and r_0, denoted as r_k^A, be nonzero (the matrix $[r_0, Ar_0, \ldots, A^k r_0]$ has full column rank). Then

$$(r_k^A)^T = \|r_k^A\|^2\, e_1^T\, [r_0, Ar_0, \ldots, A^k r_0]^+, \qquad (3.3)$$

where M^+ denotes the Moore-Penrose pseudoinverse of the matrix M, see [10,21,22] and [11, Theorem 2.1]. The proof of the following theorem is analogous to the proof of [14, Theorem 2.1]; we include it here for completeness.

Theorem 3.2 *Let $A \in \mathbb{C}^{N\times N}$, $r_0 \in \mathbb{C}^N$. If $k \geq 0$, the matrix $[r_0, Ar_0, \ldots, A^k r_0]$ has full column rank, and $A = B + C$, where B and C satisfy (3.1), then*

$$(r_k^A)^T = \|r_k^A\|^2\, [1, -\tau, \ldots, (-\tau)^k]\, [r_0, Br_0, \ldots, B^k r_0]^+. \qquad (3.4)$$

Proof For $k = 0$ the result is trivial. From (3.3)

$$(r_k^A)^T = \|r_k^A\|^2\, e_1^T [r_0, Ar_0, \ldots, A^k r_0]^+ \;\equiv\; \|r_k^A\|^2\, g_k^T. \qquad (3.5)$$

Since $[r_0, Ar_0, \ldots, A^k r_0]$ has full column rank,

$$g_k^T\, [r_0, Ar_0, \ldots, A^k r_0] = e_1^T. \qquad (3.6)$$

We next show inductively that

$$g_k^T [r_0, Br_0, \ldots, B^k r_0] = [1, -\tau, \ldots, (-\tau)^k]. \qquad (3.7)$$

From (3.6) and (3.2) we obtain

$$g_k^T r_0 = 1, \quad g_k^T A^\ell r_0 = g_k^T \left(\sum_{j=0}^{\ell} \binom{\ell}{j} \tau^j B^{\ell-j}\right) r_0 \;=\; 0, \quad 1 \leq \ell \leq k.$$

For $\ell = 1$ this yields $g_k^T B r_0 = -\tau$. The inductive step from $\ell - 1$ to ℓ gives

$$\begin{aligned} 0 = g_k^T \left(\sum_{j=0}^{\ell} \binom{\ell}{j} \tau^j B^{\ell-j}\right) r_0 \;&=\; g_k^T B^\ell r_0 + \sum_{j=1}^{\ell} \binom{\ell}{j} \tau^j (-\tau)^{\ell-j} \\ &=\; g_k^T B^\ell r_0 - (-\tau)^\ell + \sum_{j=0}^{\ell} \binom{\ell}{j} \tau^j (-\tau)^{\ell-j} \;=\; g_k^T B^\ell r_0 - (-\tau)^\ell, \end{aligned}$$

so that (3.7) indeed holds.

Note that r_k^A and hence g_k is a linear combination of the vectors $r_0, Ar_0, \ldots, A^k r_0$, and these vectors span the same space as $r_0, Br_0, \ldots, B^k r_0$ (cf. Theorem 3.1). Multiplying (3.7) from the right with $[r_0, Br_0, \ldots, B^k r_0]^+$ and using that the vector g_k is a linear combination of the columns of the matrix $[r_0, Br_0, \ldots, B^k r_0]$ yields an expression for g_k^T which, substituted in (3.5), proves (3.4). □

The following corollary gives useful bounds on $\|r_k^A\|$.

Corollary 3.1 *Using the notation and assumptions of Theorem 3.2, the norm of the GMRES residual r_k^A,*

$$\|r_k^A\| = \left\| \, [1, -\tau, \ldots, (-\tau)^k] \, [r_0, Br_0, \ldots, B^k r_0]^+ \right\|^{-1}, \tag{3.8}$$

can be bounded as

$$\omega_{min}([r_0, Br_0, \ldots, B^k r_0]) \left(\sum_{j=0}^{k} |\tau|^{2j} \right)^{-\frac{1}{2}} \leq \|r_k^A\| \leq \frac{\|[r_0, Br_0, \ldots, B^k r_0] \, v\|}{|\, [1, -\tau, \ldots, (-\tau)^k] \, v \,|}, \tag{3.9}$$

where $\omega_{min}([r_0, Br_0, \ldots, B^k r_0])$ denotes the smallest singular value of the given matrix, and $v \in \mathbb{C}^{k+1}$ is any vector satisfying $[1, -\tau, \ldots, (-\tau)^k] \, v \neq 0$.

Proof Equation (3.8) follows from taking norms on both sides of (3.4). We have assumed that $[r_0, Ar_0, \ldots, A^k r_0]$ has full rank, and hence $[r_0, Br_0, \ldots, B^k r_0]$ has full rank by Theorem 3.1. Therefore

$$\|[r_0, Br_0, \ldots, B^k r_0]^+\| = \omega_{min}([r_0, Br_0, \ldots, B^k r_0])^{-1},$$

and using this in (3.8) implies the lower bound in (3.9). For the upper bound, consider any $v \in \mathbb{C}^{k+1}$. Then by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left| \, [1, -\tau, \ldots, (-\tau)^k] \, v \right| \\
&= \left| \, [1, -\tau, \ldots, (-\tau)^k] \, [r_0, Br_0, \ldots, B^k r_0]^+ \, [r_0, Br_0, \ldots, B^k r_0] \, v \right| \\
&\leq \| \, [1, -\tau, \ldots, (-\tau)^k] \, [r_0, Br_0, \ldots, B^k r_0]^+ \| \, \|[r_0, Br_0, \ldots, B^k r_0] \, v\| \\
&= \|r_k^A\|^{-1} \, \|[r_0, Br_0, \ldots, B^k r_0] \, v\|.
\end{aligned}$$

If $[1, -\tau, \ldots, (-\tau)^k] \, v \neq 0$, a simple rearrangement yields the upper bound in (3.9), which completes the proof. □

Note that the lower bound in (3.9) is of practical use only when B is considerably simpler than A in the sense that $\omega_{min}([r_0, Br_0, \ldots, B^k r_0])$ is easy to evaluate.

3.2 The second phase of convergence

We will now consider the second phase of the GMRES convergence for the discretized convection-diffusion model problem with optimal SUPG-stabilization. The key observation is that, whenever $P_h \gg 1$, the entries $\lambda_j, \gamma_j, \mu_j$ of the matrices T_j on the block diagonal of the SUPG-discretized operator A_δ with $\delta = \delta_* = \frac{1}{2}(1 - \frac{1}{P_h})$ satisfy

$$|\lambda_j| \approx |\gamma_j| \gg |\mu_j|, \quad j = 1, \ldots, N; \tag{3.10}$$

see [15, Lemma 3.2]. In particular, for $\delta = \delta_*$ the equality (2.7) simplifies to

$$\mu_j = \frac{\nu}{3}\,(1 - \cos j\pi h).$$

Hence for small ν the value of μ_j is close to zero, and because of (3.10) the matrix T_j essentially is a (scaled and transposed) Jordan block with an eigenvalue of modulus 1. As explained in [15, Section 3.2], for moderate $P_h > 1$ the relation (3.10) still holds for smaller indices j.

For comparison, the one-dimensional convection-diffusion problem $-\nu u'' + u' = f$, with the finite difference discretization on a regular grid using upwind differences for the first derivative, yields a discretized operator of the form

$$h^{-2}\text{tridiag}(-\nu - h, 2\nu + h, -\nu);$$

see, e.g., [23, Section 4][1]. If $\nu \ll h$, then this operator represents a small perturbation of a (scaled and transposed) Jordan block corresponding to an eigenvalue of modulus 1. A simplified but nevertheless useful interpretation of the model problem (1.1) considers N loosely coupled convection-diffusion problems along the direction of the wind (one for each vertical line of the grid), each of which resembling a discretized one-dimensional convection-diffusion problem. We remark that this interpretation has been used for the analysis of the first phase of the GMRES convergence given in [15, Section 3].

Let us have a closer look at the numbers λ_j. Using (2.5) with $\delta = \delta_*$ yields

$$3(\lambda_j - \lambda_{j+1}) \;=\; (h - 4\nu)\,(c_j - c_{j+1}), \quad j = 1, \ldots, N-1.$$

The factor $c_j - c_{j+1}$ on the right hand side is positive for all $j = 1, \ldots, N-1$. Hence the numbers λ_j are strictly monotonically decreasing whenever $P_h > 2$ (i.e. whenever $h > 4\nu$; cf. (2.2)) and strictly monotonically increasing when $P_h < 2$. For $P_h = 2$ all numbers λ_j coincide. For the maximal relative difference between the largest and the smallest number λ_j a simple computation using (2.5) and $\delta = \delta_*$ gives

$$\frac{\lambda_1 - \lambda_N}{\lambda_1} \approx \frac{2}{3}\left(1 - \frac{2}{P_h}\right), \quad \text{if } P_h > 2, \tag{3.11}$$

$$\frac{\lambda_N - \lambda_1}{\lambda_N} \approx 2\left(\frac{2 - P_h}{4 + P_h}\right), \quad \text{if } 0 < P_h < 2. \tag{3.12}$$

Since we are interested in convection-dominated problems, we will focus on the case $P_h > 2$. For a numerical illustration we consider $N = 150$ and three different values of ν. The corresponding values of $P_h, \lambda_1, \gamma_1, \mu_1$, and the left and right hand sides of (3.11) are shown in the following table, which also illustrates (3.10) for $j = 1$:

ν	P_h	$\lambda_1 \approx -\gamma_1$	μ_1	$\frac{\lambda_1-\lambda_N}{\lambda_1}$	$\frac{2}{3}\left(1 - \frac{2}{P_h}\right)$
0.0015	2.2075	0.0066	1.0821e-7	0.0627	0.0627
0.00015	22.0751	0.0066	1.0821e-8	0.6062	0.6063
0.000015	220.7506	0.0066	1.0821e-9	0.6605	0.6606

In summary:

[1] For an interesting study of one-dimensional convection-diffusion operators we refer to [18].

1. For a large mesh Peclet number $P_h \gg 2$ each diagonal block T_j of the SUPG-discretized operator A_{δ_*} is close to a lower bidiagonal matrix (a scaled and transposed Jordan block) with the diagonal entry λ_j.
2. For a given (fixed) ν, the differences between the diagonal entries $\lambda_j - \lambda_{j+1}, j = 1, \ldots, N-1$ are increasing with P_h, i.e., the diagonal entries become closer with refining the mesh (decreasing the mesh size h). Analogously, for a fixed h, the differences increase with P_h, i.e., they become larger with decreasing ν.

For $P_h = 2$ all λ_j coincide, and for the smaller indices j the diagonal blocks are close to transposed Jordan blocks corresponding to the same eigenvalue. In this case the polynomial $(1 - z/\lambda_1)^N$ almost annihilates these blocks, and convergence after the initial phase of at most $N-1$ steps should be fast. We can expect that the rate of convergence decreases when P_h increases, because a larger P_h means larger differences between the diagonal entries λ_j.

We next apply the general results of Section 3.1 to give a convergence bound for GMRES that captures the described effects. Let us split the SUPG-discretized operator as

$$A_{\delta_*} = B + \tau I, \quad B \equiv \operatorname{diag}(T_1 - \tau I, \ldots, T_N - \tau I). \tag{3.13}$$

The scaled identity matrix τI plays the role of C in (3.1). The upper bound in (3.9) for the GMRES residual norm with the special choice

$$v = e_{k+1}$$

then is

$$\|r_k^{A_{\delta_*}}\| \le \frac{\|[r_0, Br_0, \ldots, B^k r_0]\, e_{k+1}\|}{|\,[1, -\tau, \ldots, (-\tau)^k]\, e_{k+1}\,|} = \frac{\|B^k r_0\|}{|\tau|^k} = \left\| \left(\frac{A_{\delta_*} - \tau I}{-\tau} \right)^k r_0 \right\|. \tag{3.14}$$

In other words, with $v = e_{k+1}$ and the splitting $A_{\delta_*} = B + \tau I$, the developed bound on the kth GMRES residual norm involves the polynomial $(1 - z/\tau)^k$. This simple choice appears to be useful for our application. We will illustrate this numerically.

Consider the model problem (1.1) with $f = 0$ and boundary conditions that are discontinuous on the inflow boundary,

$$u(\eta_1, 0) = u(1, \eta_2) = 1, \qquad \text{for} \quad 1/2 < \eta_1 \le 1 \ \text{ and } \ 0 \le \eta_2 < 1, \tag{3.15}$$

$$u(\eta_1, \eta_2) = 0, \qquad \text{elsewhere on } \partial\Omega; \tag{3.16}$$

see [15, Example 2.1] for more details on these boundary conditions introduced by Raithby [17]. We take $N = 150$ and the three different values of ν as in the table above. For each ν we get a linear algebraic system $A_{\delta_*} x = b_{\delta_*}$. Due to the zero source term, the right hand side b_{δ_*} only depends on the boundary conditions.

In order to motivate the choice of a suitable translation parameter τ, we partition b_{δ_*} into N blocks $b_{\delta_*}^{(1)}, \ldots, b_{\delta_*}^{(N)}$ corresponding to the N blocks $T_1, \ldots, T_N$ on the diagonal of A_{δ_*}. In Figure 3.1 we plot the Euclidean norms of the blocks $b_{\delta_*}^{(j)}$. We see that in each case the first blocks have the largest norm. It is plausible that this "dominance" is affecting the GMRES convergence. Motivated by this reasoning we take

$$\tau_1 \equiv \frac{1}{2}(\lambda_1 + \lambda_2)$$

as the translation parameter.

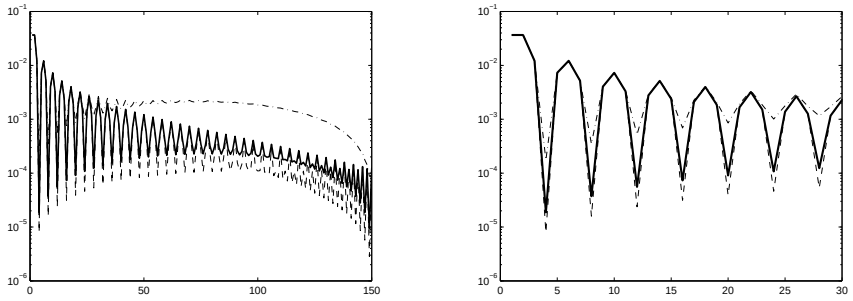

Fig. 3.1 Left: Euclidean norms of the blocks $\|b_{\delta_*}^{(j)}\|, j = 1, \ldots, 150$ for $\nu = 0.0015$ (dash-dotted line), 0.00015 (solid line), and 0.000015 (dashed line). Right: Only for the first 30 blocks $\|b_{\delta_*}^{(j)}\|, j = 1, \ldots, 30$.

The results are shown in the left part of Figure 3.2. The solid lines show the relative GMRES residual norms $\|r_k\|/\|b_{\delta_*}\|$ for the three systems corresponding to the fixed $N = 150$ and varying ν. In each case the initial guess for GMRES is the zero vector. The initial phase lasts 149 steps, which is explained in [15]. The fastest convergence after the initial phase of 149 steps occurs for $\nu = 0.0015$ ($P_h = 2.2075$) followed by 0.00015 ($P_h = 22.0751$), and finally 0.000015 ($P_h = 220.7506$). The behavior of GMRES confirms the reasoning presented above: The rate of convergence is very fast when P_h is close to 2, and it decreases with increasing P_h. The upper bound (3.14) with $\tau_1 = \frac{1}{2}(\lambda_1 + \lambda_2)$ (divided by $\|b_{\delta_*}\|$) is plotted by the dashed lines. The bound is closest for the smallest mesh Peclet number. But even for larger P_h, the bound still captures the convergence acceleration, including the changing slope of the convergence curve in subsequent steps.

The choice of a good translation parameter τ is important for the tightness of the upper bound (3.14). In the right part of Figure 3.2 we show the same GMRES convergence curves as in the left part, while the corresponding bounds are computed with

$$\tau_2 \equiv \frac{1}{2}(\lambda_1 + \lambda_{25}).$$

For $P_h \approx 2$ the difference between the numbers λ_j is small, hence $\tau_2 \approx \tau_1$ and the corresponding bounds are essentially the same. For the larger mesh Peclet numbers the numbers λ_j are farther apart, and the bounds with τ_2 are weaker. In particular, they do not give any useful information in the first N steps. In later iterations, however, the bounds with τ_2 are remarkably close to those with τ_1. A quantitative explanation of this observation needs further work.

4 Convergence acceleration, eigenvectors and eigenvalues

In this section we will address the observation of Fischer et al. [7], that the convergence of GMRES applied to $A_\delta x_\delta = b_\delta$ is faster when δ is closer to the optimal δ_*. We start with a numerical example. Consider the model problem (1.1) with $f = 0$, the boundary conditions (3.15)–(3.16), $\nu = 0.0001$, and $N = 15$. This yields the mesh Peclet number $P_h = 312.5$, and the optimal tuning parameter is $\delta_* = 0.4984$. We will consider this

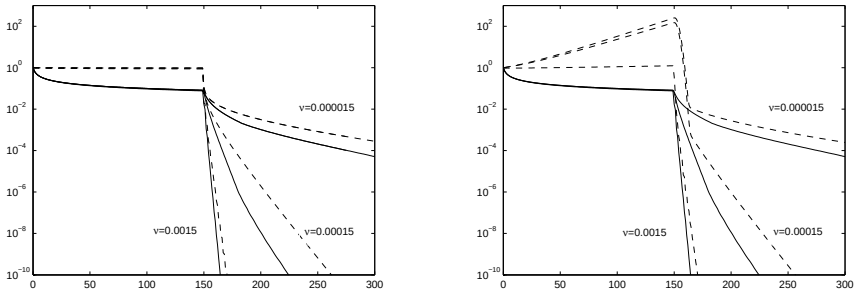

Fig. 3.2 GMRES relative residual norms $\|r_k\|/\|b_{\delta_*}\|$ (solid lines) and the corresponding upper bounds (3.14) divided by $\|b_{\delta_*}\|$ (dashed lines) with $\tau_1 = \frac{1}{2}(\lambda_1+\lambda_2)$ (left) and $\tau_2 = \frac{1}{2}(\lambda_1+\lambda_{25})$ (right) for different values of ν. The solid lines respectively the dashed lines in the left figure coincide for $k = 1, \ldots, 149$.

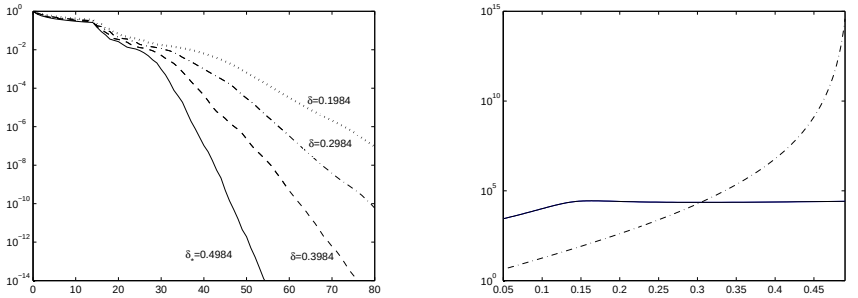

Fig. 4.1 Left: GMRES relative residual norms $\|r_k\|/\|b_{\delta_*}\|$ for $N = 15$, $\nu = 0.0001$ and the tuning parameters $\delta_* = 0.4984$ (solid line), $\delta = 0.3984$ (dashed line), $\delta = 0.2984$ (dashed-dotted line), $\delta = 0.1984$ (dotted line). Right: Comparison of $(\omega_{min}(W_\delta \Theta_\delta))^{-1}$ (solid line) and $\kappa(W_\delta)$ (dash-dotted line) for $N = 15$, $\nu = 0.0001$, and δ ranging from 0.05 to $0.49 \approx \delta_* = 0.4984$.

value, as well as three other tuning parameters, namely $\delta = 0.3984$, $\delta = 0.2984$, and $\delta = 0.1984$. We will focus on tuning parameters $\delta \leq \delta_*$. This has two reasons: First, for $\delta > \delta_*$ the observed effects are much less pronounced. Second, values $\delta > \delta_*$ tend to lead to overly diffused solutions having boundary layers much wider than those of the exact solution; cf. [5, p. 161]. Hence "overstabilization" is not recommended from a practical point of view.

Figure 4.1 shows the GMRES relative residual norms $\|r_k\|/\|b_\delta\|$ with the initial guess equal to zero for the SUPG-discretized systems with the four choices of tuning parameters. We see that the rate of convergence in the second phase (starting with step $N = 15$) is strongly correlated to the distance of the chosen tuning parameter δ to the optimal δ_*; smaller distance means faster convergence. At the same time, however, a smaller distance means *significantly worse conditioned eigenvectors* of the SUPG-

discretized operator:

$$\kappa(W_\delta) \;=\; \begin{cases} 8.05e+34, & \text{if } \delta = \delta_* = 0.4984, \\ 4.93e+06, & \text{if } \delta = 0.3984, \\ 1.67e+04, & \text{if } \delta = 0.2984, \\ 3.82e+02, & \text{if } \delta = 0.1984 \end{cases} \tag{4.1}$$

(these numbers were computed using (2.11)). The condition number $\kappa(W_\delta)$ appears in the standard bound (1.2) for the GMRES residual norms. Hence these observations indicate that this bound should always be carefully examined when applied in the context of highly nonnormal problems. Our goal in the following is to explain why in case of the considered model problem the conditioning of the eigenvectors has little influence on the convergence of GMRES, and to illustrate the role of the eigenvalues.

4.1 A lower bound for the GMRES residual

We first derive a lower bound on the kth GMRES residual norm for a general diagonalizable matrix A. The kth GMRES residual norm is equal to zero when A has fewer than $k+1$ distinct eigenvalues, or when the initial residual has fewer than $k+1$ nonzero components in the invariant subspaces of A corresponding to distinct eigenvalues. This leads to the rather technical but nevertheless natural assumptions in the following theorem.

Theorem 4.1 *Let $A \in \mathbb{C}^{N\times N}$ be diagonalizable with at least $k+1$ distinct eigenvalues $\sigma_1, \ldots, \sigma_{k+1}$, and denote its eigendecomposition by $A = W\mathrm{diag}(\sigma_1, \ldots, \sigma_N)W^{-1}$, where $W \equiv [w_1, \ldots, w_N]$. Suppose that the initial residual is of the form $r_0 = \sum_{i=1}^{\ell} \theta_i w_i$, where $\ell \geq k+1$ and $\theta_1, \ldots, \theta_\ell$ are nonzero scalars. Then the kth GMRES residual norm for A and r_0 satisfies*

$$\|r_k^A\| \geq \frac{\omega_{min}\,(W_\ell \Theta_\ell)}{\|\, e_1^T\, M_{k+1}^{-1}\,\|}\,, \tag{4.2}$$

where $W_\ell \equiv [w_1, \ldots, w_\ell]$, $\Theta_\ell \equiv \mathrm{diag}(\theta_1, \ldots, \theta_\ell)$, and

$$M_{k+1} \equiv \begin{bmatrix} 1 & \sigma_1 & \cdots & \sigma_1^k \\ \vdots & \vdots & & \vdots \\ 1 & \sigma_{k+1} & \cdots & \sigma_{k+1}^k \end{bmatrix} \tag{4.3}$$

is the $(k+1)\times(k+1)$ Vandermonde matrix corresponding to $\sigma_1, \ldots, \sigma_{k+1}$.

Proof Using the eigendecomposition of A and the expansion of r_0 in the eigenvectors of A we can write

$$\begin{aligned} A^m r_0 &= W\,\mathrm{diag}(\sigma_1^m, \ldots, \sigma_N^m)\,[\theta_1, \ldots, \theta_\ell, 0, \ldots, 0]^T \\ &= W\,\mathrm{diag}(\theta_1, \ldots, \theta_\ell, 0, \ldots, 0)\,[\sigma_1^m, \ldots, \sigma_N^m]^T = W_\ell \Theta_\ell\,[\sigma_1^m, \ldots, \sigma_\ell^m]^T\,, \end{aligned}$$

$m \geq 0$, and hence

$$[r_0, Ar_0, \ldots, A^k r_0] \;=\; W_\ell \Theta_\ell \begin{bmatrix} M_{k+1} \\ M_{21} \end{bmatrix} \;=\; W_\ell \Theta_\ell \begin{bmatrix} I_{k+1} \\ M_{21} M_{k+1}^{-1} \end{bmatrix} M_{k+1}\,,$$

where M_{21} is the $(\ell - k - 1) \times (k + 1)$ Vandermonde matrix corresponding to the eigenvalues $\sigma_{k+2}, \ldots, \sigma_\ell$ (note that M_{21} vanishes for $\ell = k+1$). Using (3.3) we obtain[2]

$$\begin{aligned}
\|r_k^A\| &= \left\| e_1^T \left(W_\ell \Theta_\ell \begin{bmatrix} I_{k+1} \\ M_{21} M_{k+1}^{-1} \end{bmatrix} M_{k+1} \right)^+ \right\|^{-1} \\
&= \left\| e_1^T M_{k+1}^{-1} \left(W_\ell \Theta_\ell \begin{bmatrix} I_{k+1} \\ M_{21} M_{k+1}^{-1} \end{bmatrix} \right)^+ \right\|^{-1} \\
&\geq \| e_1^T M_{k+1}^{-1} \|^{-1} \left\| \left(W_\ell \Theta_\ell \begin{bmatrix} I_{k+1} \\ M_{21} M_{k+1}^{-1} \end{bmatrix} \right)^+ \right\|^{-1} \\
&= \| e_1^T M_{k+1}^{-1} \|^{-1} \ \omega_{min} \left(W_\ell \Theta_\ell \begin{bmatrix} I_{k+1} \\ M_{21} M_{k+1}^{-1} \end{bmatrix} \right) \\
&\geq \| e_1^T M_{k+1}^{-1} \|^{-1} \ \omega_{min} \left(W_\ell \Theta_\ell \right) \ \omega_{min} \left(\begin{bmatrix} I_{k+1} \\ M_{21} M_{k+1}^{-1} \end{bmatrix} \right) \\
&= \| e_1^T M_{k+1}^{-1} \|^{-1} \ \omega_{min} \left(W_\ell \Theta_\ell \right) \ \sqrt{1 + \omega_{min}^2 (M_{21} M_{k+1}^{-1})}
\end{aligned}$$

so that (4.2) indeed holds. □

The term $\|e_1^T M_{k+1}^{-1}\|$ depends only on the eigenvalues of A, and the term $\omega_{min}(W_\ell \Theta_\ell)$ depends on the eigenvectors and on the initial residual. Hence in the bound (4.2) the eigenvectors and the initial residual are *not* separated as in the standard bound (1.2). The interplay between the two is taken into account. In particular, the columns of the matrix $W_\ell \Theta_\ell$ represent the components of r_0 in the directions of $w_1, \ldots, w_l$, i.e.,

$$r_0 = (W_\ell \Theta_\ell) e, \qquad e = [1, \ldots, 1]^T.$$

Note that the inverse Vandermonde matrix is explicitly known; see, e.g., [8]. In particular,

$$e_1^T M_{k+1}^{-1} = \left[\prod_{i \neq 1} \frac{\sigma_i}{\sigma_1 - \sigma_i}, \quad \ldots \quad , \prod_{i \neq k+1} \frac{\sigma_i}{\sigma_{k+1} - \sigma_i} \right]. \tag{4.4}$$

Apart from the assumption that the first $k+1$ eigenvalues in Theorem 4.1 are distinct, the ordering of the eigenvalues $\sigma_1, \ldots, \sigma_N$ in this theorem is *not* specified. Hence the lower bound (4.2) holds for *any subset of $k+1$ of the distinct eigenvalues of A.*

4.2 Ill-conditioning of the eigenvectors and its relationship to convergence

We will now examine the numerator of the bound (4.2) in case of the SUPG-discretized model problem. We assume, for simplicity, that A_δ is diagonalizable. This assumption does not represent a major restriction on the applicability of the results, because A_δ is diagonalizable except for a finite number of values δ; cf. Section 2.

[2] We also use that for matrices $X \in \mathbb{C}^{N \times (k+1)}$ and $Y \in \mathbb{C}^{(k+1) \times (k+1)}$, both of full rank $k+1$, the pseudoinverse of the product satisfies $(XY)^+ = Y^{-1} X^+$; see, e.g., [2, Corollary 1.4.1].

The eigenvector matrix of A_δ can be written as $W_\delta = \mathrm{diag}(W_1, \ldots, W_N)$, where

$$W_j = \mathrm{diag}\,(\zeta_j^{-\frac{1}{2}}, \zeta_j^{-1}, \ldots, \zeta_j^{-\frac{N}{2}})V, \quad \zeta_j = \frac{\mu_j}{\gamma_j}, \quad j = 1, \ldots, N,$$

(we use $\mathrm{diag}\,(\cdot)$ for both diagonal and block diagonal matrices) and V has the entries

$$V_{i,k} = \sqrt{\frac{2}{N+1}}\,\sin ikh\pi, \quad i, k = 1, \ldots, N;$$

see (2.9). Note that V is independent of the choice of δ and that $V = V^T = V^{-1}$. For large mesh Peclet numbers the optimal tuning parameter δ_* is close to $\frac{1}{2}$ (cf. (2.4)), giving very small values of ζ_j (see (3.10) in Section 3.2) and in turn severely ill conditioned eigenvectors of A_δ. This has been illustrated numerically at the beginning of Section 4.

We write the initial residual as

$$r_0 = [\rho_1^{(1)}, \ldots, \rho_N^{(1)}, \ldots, \rho_1^{(N)}, \ldots, \rho_N^{(N)}]^T,$$

where the entries $\rho_1^{(j)}, \ldots, \rho_N^{(j)}$ of r_0 correspond to the jth block of the matrix W_δ. We denote by $\Theta_\delta \equiv \mathrm{diag}(\Theta_1, \ldots, \Theta_N)$ the $N^2 \times N^2$ diagonal matrix composed of the diagonal blocks $\Theta_1, \ldots, \Theta_N$,

$$\Theta_j = \mathrm{diag}\left(W_j^{-1}[\rho_1^{(j)}, \ldots, \rho_N^{(j)}]^T\right), \quad j = 1, \ldots, N.$$

For simplicity we assume that the matrix Θ_δ is nonsingular, i.e., r_0 has nonzero components in *all* invariant subspaces of A_δ. Under this assumption the numerator in the lower bound (4.2) for the SUPG-discretized model problem is given by

$$\omega_{min}(W_\delta\Theta_\delta) = \min_{1\le j\le N} \omega_{min}(W_j\Theta_j) = \min_{1\le j\le N} \|(W_j\Theta_j)^{-1}\|^{-1}. \tag{4.5}$$

The following theorem presents some technical details needed for our discussion below.

Theorem 4.2 *With the notation introduced above,*

$$\rho_1^{(j)}\,(W_j\Theta_j)^{-1} = \hat{V}_\delta^{(j)}\mathrm{diag}(1, \zeta_j^{\frac{1}{2}}, \zeta_j, \ldots, \zeta_j^{\frac{N-1}{2}})$$

holds for each $j = 1, \ldots, N$, where the matrix $\hat{V}_\delta^{(j)}$ has the entries

$$(\hat{V}_\delta^{(j)})_{i,k} \equiv \frac{\rho_1^{(j)}\sin(ikh\pi)}{\sum_{\ell=1}^N \zeta_j^{\frac{\ell-1}{2}}\rho_\ell^{(j)}\sin(i\ell h\pi)}, \qquad 1 \le i, k \le N.$$

Proof A simple manipulation shows that

$$\begin{aligned} W_j^{-1}[\rho_1^{(j)}, \ldots, \rho_N^{(j)}]^T &= V\mathrm{diag}\,(\zeta_j^{\frac{1}{2}}, \zeta_j, \ldots, \zeta_j^{\frac{N}{2}})[\rho_1^{(j)}, \ldots, \rho_N^{(j)}]^T \\ &= \sqrt{\frac{2}{N+1}}\sum_{i=1}^N \zeta_j^{\frac{i}{2}}\rho_i^{(j)}\,[\sin ih\pi, \ldots, \sin iNh\pi]^T, \end{aligned}$$

so that

$$\Theta_j^{-1}W_j^{-1} =$$
$$\sqrt{\frac{N+1}{2}}\,\mathrm{diag}\left(\sum_{i=1}^{N}\zeta_j^{\frac{i}{2}}\rho_i^{(j)}\sin ih\pi,\ldots,\sum_{i=1}^{N}\zeta_j^{\frac{i}{2}}\rho_i^{(j)}\sin iNh\pi\right)^{-1}V\mathrm{diag}\,(\zeta_j^{\frac{1}{2}},\ldots,\zeta_j^{\frac{N}{2}}) =$$
$$\frac{1}{\rho_1^{(j)}}\,\hat{V}_\delta^{(j)}\mathrm{diag}\,(1,\zeta_j^{\frac{1}{2}},\zeta_j,\ldots,\zeta_j^{\frac{N-1}{2}}),$$

which finishes the proof. □

The lemma shows that for small ζ_j the matrix $\rho_1^{(j)}(W_j\Theta_j)^{-1}$ is strongly dominated by the first column of $\hat{V}_\delta^{(j)}$. This column has the entries

$$(\hat{V}_\delta^{(j)})_{i,1} = \frac{\rho_1^{(j)}\sin(ih\pi)}{\rho_1^{(j)}\sin(ih\pi)+\sum_{\ell=2}^{N}\zeta_j^{\frac{\ell-1}{2}}\rho_\ell^{(j)}\sin(i\ell h\pi)} = 1+\mathcal{O}(\zeta_j^{\frac{1}{2}}), \qquad 1\le i\le N.$$

Hence for small ζ_j the matrix $\rho_1^{(j)}(W_j\Theta_j)^{-1}$ is close to a rank one matrix,

$$\rho_1^{(j)}(W_j\Theta_j)^{-1} \approx \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}.$$

This means that the numerator in the lower bound (4.2) will have the approximate value

$$\omega_{min}(W_\delta\Theta_\delta) \approx \frac{1}{\sqrt{N}}\min_{1\le j\le N}|\rho_1^{(j)}|; \tag{4.6}$$

see (4.5).

Note that when $P_h \gg 1$ we can neglect the terms of order ν in (2.6)–(2.7), and then

$$\zeta_j = \frac{\mu_j}{\gamma_j} \approx \frac{\delta-\frac{1}{2}}{\delta+\frac{1}{2}} < 1. \tag{4.7}$$

When δ approaches δ_* the numbers ζ_j become very small, and the approximation (4.6) becomes more realistic, independently of the fact that the eigenvector matrix W_δ becomes more ill conditioned. Such an ill conditioning is *not* reflected in (4.6), where the interplay between eigenvectors and initial residual is taken into account. Provided that the lower bound (4.2) is sufficiently tight, this explains the observed insensitivity of GMRES with respect to the conditioning of the eigenvectors. Numerical examples are given in Section 4.4 below. Note that an explanation for the negligible role of the eigenvector conditioning in the case of a single tridiagonal Toeplitz matrix was given in [12,13].

4.3 Eigenvalue distribution and convergence

The numerator $\omega_{min}(W_\delta \Theta_\delta)$ in the lower bound (4.2) is not significantly influenced by the tuning parameter δ. The dependence of the convergence on the value of δ observed in Figure 4.1 must therefore be related to the denominator; see (4.4). In the following we will briefly discuss the situation for mesh Peclet numbers $P_h \gg 1$. Using (2.5)–(2.7),

$$\mu_j \gamma_j = \frac{(\delta h c_j)^2}{9} \left(1 - \frac{1}{4\delta^2} + \mathcal{O}\left(P_h^{-1}\right)\right), \tag{4.8}$$

and, using (2.8), the kth eigenvalue of the diagonal block T_j of A_δ can be expressed as

$$\sigma_{jk} = \frac{2}{3}\delta h c_j \left(1 + \sqrt{1 - \frac{1}{4\delta^2} + \mathcal{O}\left(P_h^{-1}\right)} \cos kh\pi + \mathcal{O}\left(P_h^{-1}\right)\right), \quad 1 \le j,k \le N.$$

This gives the difference between the kth and the ℓth eigenvalue of T_j in the form

$$\sigma_{jk} - \sigma_{j\ell} = \frac{2}{3}\delta h c_j \left(\sqrt{1 - \frac{1}{4\delta^2} + \mathcal{O}\left(P_h^{-1}\right)} \left(\cos kh\pi - \cos \ell h\pi\right) + \mathcal{O}(P_h^{-1})\right).$$

With $P_h \gg 1$ the optimal tuning parameter δ_* is very close to $\frac{1}{2}$, and hence the eigenvalues of the diagonal block T_j are very close, which holds for $j = 1, \ldots, N$.

Each entry of the vector (4.4) consists of a product of ratios, with an eigenvalue in the numerator and a difference between the same eigenvalue and some other eigenvalue in the denominator. When both eigenvalues belong to the same block, say T_j, the ratio takes the form

$$\frac{\sigma_{jk}}{\sigma_{jk} - \sigma_{j\ell}} = \frac{1 + \sqrt{1 - \frac{1}{4\delta^2} + \mathcal{O}\left(P_h^{-1}\right)} \cos kh\pi + \mathcal{O}\left(P_h^{-1}\right)}{\sqrt{1 - \frac{1}{4\delta^2} + \mathcal{O}\left(P_h^{-1}\right)} \left(\cos kh\pi - \cos \ell h\pi\right) + \mathcal{O}(P_h^{-1})}. \tag{4.9}$$

For $P_h \gg 1$ and for optimal stabilization $\delta = \delta_* \approx \frac{1}{2}$ this ratio will be huge, and it will decrease for δ departing from δ_*. Depending on the iteration number and the ordering of the eigenvalues, the entries of the vector (4.4) can be expected to contain ratios of the type (4.9). Hence with δ departing from δ_*, it can be expected that the factor $\|e_1^T M_{k+1}^{-1}\|^{-1}$ in the lower bound (4.2) is increasing. This suggests, in turn, slower convergence.

4.4 Numerical examples

We will now give numerical examples illustrating the bound (4.2) for the SUPG-discretized model problem (1.1). We consider $f = 0$, the boundary conditions (3.15)–(3.16), $N = 15$ and $\nu = 0.0001$ (as in Figure 4.1).

On the right side of Figure 4.1 we plot $(\omega_{min}(W_\delta \Theta_\delta))^{-1}$ (solid line) and $\kappa(W_\delta)$ (dash-dotted line) for δ ranging from 0.05 to $0.49 \approx \delta_* = 0.4984$. We see that $\omega_{min}(W_\delta \Theta_\delta)$ is essentially independent of the choice of δ, as indicated by Theorem 4.2 and by the discussion thereafter. On the other hand, $\kappa(W_\delta)$ increases significantly when δ approaches the optimal δ_*.

On the left sides of Figure 4.2 we show the GMRES behaviour and the lower bound (4.2) for linear algebraic systems generated with the same data as used for the

computation of Figure 4.1 (in all experiments we use $x_0 = 0$). As shown on the right side of Figure 4.1, the numerator of the bound is essentially independent of δ; in each case it is approximately equal to 10^{-5}. To compute the denominator of the bound, we have chosen subsets of the eigenvalues of A_δ using the following algorithm:

1. Let $J = \{1, 2, \ldots, N^2\}$ and let $\sigma_1, \ldots, \sigma_{N^2}$ be the eigenvalues of A_δ
2. For $m = 1, \ldots, N^2$:
   - Choose an index $i \in J$ minimizing $\prod_{\ell \in J, \ell \neq i} |\sigma_i - \sigma_\ell|$.
   - Set $J = J \backslash \{i\}$, $i_m = i$.
3. Reorder the eigenvalues as $\sigma_{i_1}, \ldots, \sigma_{i_{N^2}}$.
4. For $k = 1, 2, \ldots$ use the first $k+1$ of them to compute $\|e_1^T M_{k+1}^{-1}\|$.

We see that although the lower bound (4.2) is not very close to the actual GMRES convergence curve, it captures the slope rather well. This slope is determined by the denominator $\|e_1^T M_{k+1}^{-1}\|$. The bound (4.2) gives the relevant information about the acceleration of GMRES after the initial phase of slow convergence.

Figure 4.2 shows on the right sides the spectra of the corresponding SUPG-discretized operators A_δ. Note that the y-axes in these figures have different scales. The differences between the four spectra on the right part of Figure 4.2 are more apparent in Figure 4.3, in which we plot the convex hulls of the four spectra. When δ decreases from the optimal δ_*, the relative differences between the individual eigenvalues increase significantly. The norm $\|e_1^T M_{k+1}^{-1}\|^{-1}$ (which is independent of eigenvalue scaling) increases, indicating a slower convergence of GMRES.

5 Concluding remarks

For large mesh Peclet numbers and the optimal SUPG stabilization GMRES applied to the discretized model problem (1.1) resembles the behavior of the shifted power monomial applied to the initial residual; see (3.14). With a well chosen shift τ, the monomial captures amazingly well the breaking point when the convergence starts to accelerate rapidly after a period of near stagnation, and, at the same time, it gives a very realistic upper bound for the rate of convergence after the acceleration occurs. With less carefully chosen shifts the breaking point and the period of near stagnation are not captured well, but the slope of the convergence curve after stagnation still is.

The optimal stabilization leads to the most compact spectrum, but very poor conditioning of the eigenvectors of the SUPG discretized operator. When the stabilization parameter decreases from its optimal value (for physical reasons explained above we do not consider overstabilization), the spectrum becomes less compact, but the conditioning of the eigenvector matrix decreases rapidly. This shows that the convergence bounds based on the spectral information separating the role of the eigenvalues (typically using some minmax polynomial approximation) and the eigenvectors (typically using its condition number) do not offer any insight into GMRES convergence when applied to discretized problems like (1.1). The lower bound (4.2) represents a significant underestimation of the true residual norms, however, it gives the correct slope of convergence after the initial phase. It has been shown that in this lower bound the conditioning of the eigenvector matrix plays almost no role. Therefore it indicates that the most compact spectrum (for the optimal $\delta = \delta_*$) corresponds to the fastest convergence, as observed in [7]. It should be noted that for the bound (4.2) the spread of the eigenvalues along the real line is insignificant, because the ratio of the largest and

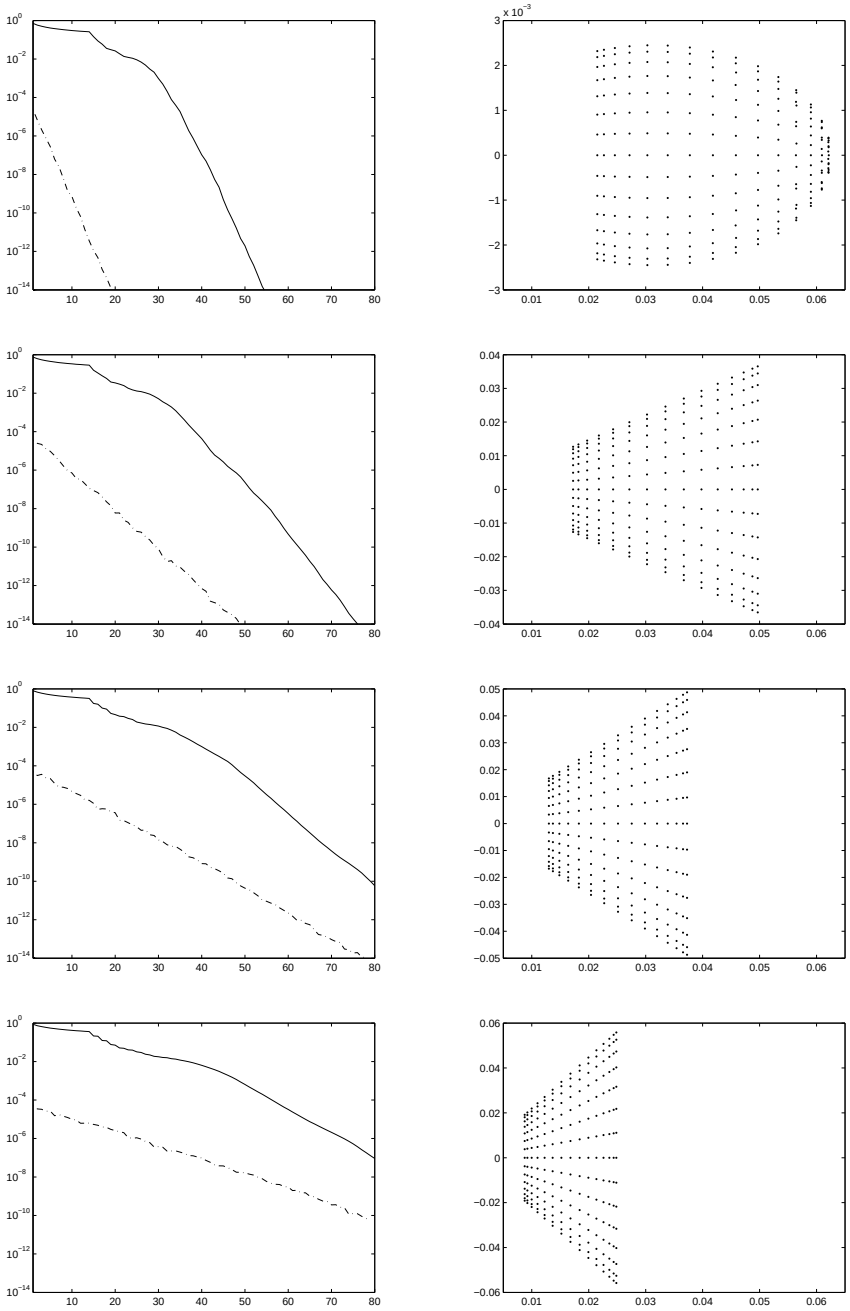

Fig. 4.2 Left: Relative GMRES residual norms (solid line) and the lower bound (4.2) (dashed-dotted line) for, repectively, $\delta_* = 0.4984$, $\delta = 0.3984$, $\delta = 0.2984$ and $\delta = 0.1984$. Right: Eigenvalues of the corresponding SUPG-discretized operator. Note that the y-axes in these figures have different scales.

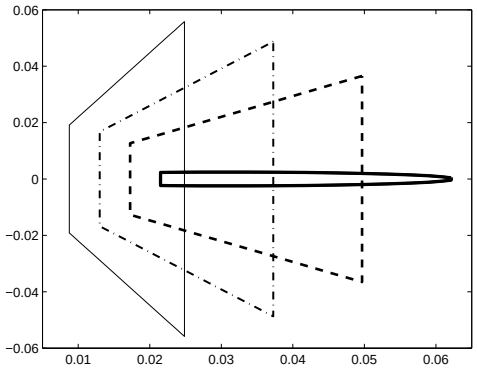

Fig. 4.3 The convex hulls of the four spectra shown on the right of Figure 4.2. The relative distances between the eigenvalues of A_δ increase when δ decreases from δ_*.

smallest real part of the eigenvalues in the spectrum does not depend on the stabilization parameter δ (see (2.5), (2.8) and (4.4)). What matters is the spreading of the eigenvalues in the direction of the imaginary axis; see Figure 4.3.

Apart perhaps from linking the GMRES convergence to shifted power monomials, this paper gives no practically applicable GMRES convergence bounds for the discretized convection-diffusion model problem (1.1). Our goal is different, namely, to get a better understanding of the acceleration of the convergence when GMRES is applied to such a problem. Our analysis is based on some generally applicable results which might find their use elsewhere as well.

References

1. Brooks, A.N., Hughes, T.J.R.: Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. Comput. Methods Appl. Mech. Engrg. **32**(1-3), 199–259 (1982)
2. Campbell, S.L., Meyer Jr., C.D.: Generalized inverses of linear transformations. Dover Publications Inc., New York (1991). Corrected reprint of the 1979 original
3. Elman, H.C., Ramage, A.: An analysis of smoothing effects of upwinding strategies for the convection-diffusion equation. SIAM J. Numer. Anal. **40**(1), 254–281 (2002)
4. Elman, H.C., Ramage, A.: A characterisation of oscillations in the discrete two-dimensional convection-diffusion equation. Math. Comp. **72**(241), 263–288 (2003)
5. Elman, H.C., Silvester, D.J., Wathen, A.J.: Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics. Numerical Mathematics and Scientific Computation. Oxford University Press, New York (2005)
6. Ernst, O.G.: Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations. SIAM J. Matrix Anal. Appl. **21**(4), 1079–1101 (2000)
7. Fischer, B., Ramage, A., Silvester, D.J., Wathen, A.J.: On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems. Comput. Methods Appl. Mech. Engrg. **179**(1-2), 179–195 (1999)
8. Gautschi, W.: On inverses of Vandermonde and confluent Vandermonde matrices. III. Numer. Math. **29**(4), 445–450 (1977/78)
9. Hughes, T.J.R., Brooks, A.: 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*, vol. 34, pp. 19–35. Amer. Soc. Mech. Engrs. (ASME), New York (1979)

10. Ipsen, I.C.F.: Expressions and bounds for the GMRES residual. BIT **40**(3), 524–535 (2000)
11. Liesen, J., Rozložník, M., Strakoš, Z.: Least squares residuals and minimal residual methods. SIAM J. Sci. Comput. **23**(5), 1503–1525 (2002)
12. Li, R.-C., Zhang, W.: The rate of convergence of GMRES on a tridiagonal Toeplitz linear system. Numer. Math. **112**(2), 267–293 (2009)
13. Li, R.-C., Zhang, W.: The rate of convergence of GMRES on a tridiagonal Toeplitz linear system II. Linear Algebra Appl. **431**(12), 2425–2436 (2009)
14. Liesen, J., Strakoš, Z.: Convergence of GMRES for tridiagonal Toeplitz matrices. SIAM J. Matrix Anal. Appl. **26**(1), 233–251 (2004)
15. Liesen, J., Strakoš, Z.: GMRES convergence analysis for a convection-diffusion model problem. SIAM J. Sci. Comput. **26**(6), 1989–2009 (2005)
16. Morton, K.W.: Numerical Solution of Convection-Diffusion Problems, *Applied Mathematics and Mathematical Computation*, vol. 12. Chapman & Hall, London (1996)
17. Raithby, G.D.: Skew upstream differencing schemes for problems involving fluid flow. Comput. Methods Appl. Mech. Engrg. **9**(2), 153–164 (1976)
18. Reddy, S.C., Trefethen, L.N.: Pseudospectra of the convection-diffusion operator. SIAM J. Appl. Math. **54**(6), 1634–1649 (1994)
19. Roos, H.G., Stynes, M., Tobiska, L.: Numerical Methods for Singularly Perturbed Differential Equations, *Springer Series in Computational Mathematics*, vol. 24. Springer-Verlag, Berlin (1996)
20. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Statist. Comput. **7**(3), 856–869 (1986)
21. Sadok, H.: Méthodes de projection pour les systèmes linéaires et non linéaires, *University of Lille 1*, Habilitation Thesis. Lille, France (1994)
22. Stewart, G.W.: Collinearity and least squares regression. Statist. Sci. **2**(1), 68–100 (1987).
23. Stynes, M.: Steady-state convection-diffusion problems. Acta Numer. **14**, 445–508 (2005)