SIAM J. MATRIX ANAL. APPL.
Vol. 26, No. 1, pp. 233–251

CONVERGENCE OF GMRES FOR TRIDIAGONAL TOEPLITZ MATRICES*

J. LIESEN† AND Z. STRAKOŠ‡

Abstract. We analyze the residuals of GMRES [Y. Saad and M. H. Schultz, *SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 856–859], when the method is applied to tridiagonal Toeplitz matrices. We first derive formulas for the residuals as well as their norms when GMRES is applied to scaled Jordan blocks. This problem has been studied previously by Ipsen [*BIT*, 40 (2000), pp. 524–535] and Eiermann and Ernst [*Private communication,* 2002], but we formulate and prove our results in a different way. We then extend the (lower) bidiagonal Jordan blocks to tridiagonal Toeplitz matrices and study extensions of our bidiagonal analysis to the tridiagonal case. Intuitively, when a scaled Jordan block is extended to a tridiagonal Toeplitz matrix by a superdiagonal of small modulus (compared to the modulus of the subdiagonal), the GMRES residual norms for both matrices and the same initial residual should be close to each other. We confirm and quantify this intuitive statement. We also demonstrate principal difficulties of any GMRES convergence analysis which is based on eigenvector expansion of the initial residual when the eigenvector matrix is ill-conditioned. Such analyses are complicated by a cancellation of possibly huge components due to close eigenvectors, which can prevent achieving well-justified conclusions.

Key words. Krylov subspace methods, GMRES, minimal residual methods, convergence analysis, Jordan blocks, Toeplitz matrices

AMS subject classifications. 15A09, 65F10, 65F20

DOI. 10.1137/S0895479803424967

1. Introduction.

Consider solving a linear algebraic system $Ax = b$, real or complex, where A is an N by N nonsingular matrix with GMRES [9]. 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 $r_n = b - Ax_n$ 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 and $\|\cdot\|$ denotes the 2-norm. It is easy to see from (1.1) that (in exact arithmetic) the GMRES algorithm terminates, i.e., computes the solution x, in at most N steps. We also wish to point out that, unless there is a well-justified reason for choosing a nonzero initial approximation, one should consider $x_0 = 0$ (see [8]).

Suppose that the vectors $r_0, Ar_0, \ldots, A^n r_0$ generating the $(n+1)$st Krylov subspace $\mathcal{K}_{n+1}(A, r_0) = \mathrm{span}\{r_0, Ar_0, \ldots, A^n r_0\}$ are linearly independent. Then r_n is a nonzero vector and GMRES cannot terminate before the step $n+1$. Denote by K_{n+1} the matrix of the Krylov vectors,

$$K_{n+1} \;=\; [r_0, Ar_0, \ldots, A^n r_0] \;\equiv\; [r_0, W_n R_n]\,, \tag{1.2}$$

*Received by the editors March 17, 2003; accepted for publication (in revised form) by L. Elden January 9, 2004; published electronically September 14, 2004.
http://www.siam.org/journals/simax/26-1/42496.html

†Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft.

‡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, http://www.cs.cas.cz/~strakos). 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.

where W_n has orthonormal columns and R_n is upper triangular.

In [5, Theorem 2.1] Ipsen shows that r_n is determined by the first row of the Moore–Penrose pseudoinverse of K_{n+1},

$$r_n^T \;=\; \|r_n\|^2 \, e_1^T K_{n+1}^+. \tag{1.3}$$

Based on this result she argues that as long as the matrix K_{n+1} is well-conditioned, the decrease of the GMRES residual norms in the steps 1 to n must be slow. Then she applies this relation to analyze the GMRES behavior for scaled Jordan blocks [5, Theorem 3.1].

In [6, pp. 1505–1506], it is shown that

$$r_n^T \;=\; \|r_n\|^2 \, e_1^T [r_0, W_n]^+ \,, \tag{1.4}$$

which refines Ipsen's argument about the relation between ill-conditioning of the Krylov matrix and convergence of the GMRES residual norms. The proofs in [6] are based on the elementary geometrical interpretation of the pseudoinverse (orthogonality relations).

In this paper we study the GMRES residuals for linear systems with tridiagonal Toeplitz matrices T. We start with results analogous to those of Ipsen for scaled Jordan blocks, and then we analyze their extensions. We are particularly interested in the case when the entries on the superdiagonal of T are significantly smaller in modulus (absolute value) than the entries on the subdiagonal. This represents an example of very large eigenvector conditioning (even infinite when the matrix reduces to a scaled Jordan block); i.e., we deal with highly nonnormal matrices. Rather than applying a worst-case analysis based on properties of the matrix T only, we exploit the structure of T and relate the GMRES convergence to the structure and numerical values of the entries of the initial residual r_0. This allows qualitative as well as quantitative statements about the influence of T *as well as* r_0 on the GMRES residuals. In proofs, we follow, as in [6, pp. 1505–1506], the elementary orthogonality idea.

Analytic results for scaled Jordan blocks and general tridiagonal Toeplitz matrices are given in sections 2 and 3, respectively. Section 4 shows numerical experiments, and section 5 contains concluding remarks. In this paper we do not consider rounding errors, i.e., we assume exact arithmetic.

2. Scaled Jordan blocks. For given nonzero parameters γ and λ, consider an N by N scaled Jordan block J,

$$J \;=\; \gamma S + \lambda I \;\equiv\; \gamma(S + \tau I)\,, \qquad \tau \;\equiv\; \frac{\lambda}{\gamma}\,, \tag{2.1}$$

where I is the identity and $S = [e_2, \ldots, e_N, 0]$ is the down shift matrix (e_j denotes the jth vector of the standard Euclidean basis). The scaling does not affect GMRES convergence; it is used for convenience only. The GMRES residual norms for systems with scaled Jordan blocks have been studied in [2] and in [5, section 3]. Here we study the same problem, but we formulate and prove our results differently from [2, 5].

THEOREM 2.1. *Suppose that GMRES is applied to a system with the matrix* $J = \gamma(S + \tau I)$ *and the initial residual* $r_0 = [\rho_1, \ldots, \rho_N]^T$. *Let* ρ_l *be the first nonzero entry of* r_0. *Then for* $n = 0, 1, \ldots, N - l$ *the GMRES residuals satisfy*

$$r_n^T = \|r_n\|^2 \, [1, -\tau, \ldots, (-\tau)^n] \, [r_0, Sr_0, \ldots, S^n r_0]^+ \,, \tag{2.2}$$

$$\|r_n\| \geq \left(\sum_{j=0}^{n} |\tau|^{2j}\right)^{-\frac{1}{2}} \sigma_{\min}\left([r_0, Sr_0, \ldots, S^n r_0]\right), \tag{2.3}$$

and $r_{N-l+1} = 0$*, where* $\sigma_{\min}(X)$ *denotes the minimal singular value of the matrix* X*. Furthermore, for* $n = 0, 1, \ldots, N-l$*,*

$$\|r_n\| \;\leq\; (n+1)^{\frac{1}{2}}\, \|r_0\| \left(\sum_{j=0}^{n} |\tau|^{2j}\right)^{-\frac{1}{2}}. \tag{2.4}$$

Proof. Since $\mathcal{K}_{n+1}(J, r_0) = \mathcal{K}_{n+1}(S, r_0)$ and $\rho_l \neq 0$, it is easy to see that for $n = 0, 1, \ldots, N-l$ the matrices $[r_0, Jr_0, \ldots, J^n r_0]$ have full column rank. Hence, for $n = 0, 1, \ldots, N-l$, (1.3) (see also [6, Theorem 2.1]) shows that

$$r_n^T \;=\; \|r_n\|^2\, e_1^T [r_0, Jr_0, \ldots, J^n r_0]^+ \;\equiv\; \|r_n\|^2\, g_n^T. \tag{2.5}$$

The identity $[r_0, Jr_0, \ldots, J^n r_0]^+ [r_0, Jr_0, \ldots, J^n r_0] \;=\; I$ gives

$$g_n^T [r_0, Jr_0, \ldots, J^n r_0] \;=\; e_1^T.$$

We next prove, by induction,

$$g_n^T [r_0, Sr_0, \ldots, S^n r_0] \;=\; [1, -\tau, \ldots, (-\tau)^n]. \tag{2.6}$$

Clearly,

$$0 = g_n^T J r_0 = \gamma\, g_n^T S r_0 + \lambda\, g_n^T r_0 = \gamma\, g_n^T S r_0 + \lambda, \qquad \text{i.e., } g_n^T S r_0 = -\tau,$$

and the general step,

$$\begin{aligned}
0 = g_n^T J^k r_0 &= g_n^T (\gamma S + \lambda I)^k r_0 \\
&= g_n^T \left(\sum_{j=0}^{k} \binom{k}{j} \gamma^{k-j} \lambda^j S^{k-j}\right) r_0 \\
&= \gamma^k g_n^T S^k r_0 + \sum_{j=1}^{k} \binom{k}{j} \gamma^{k-j} \lambda^j (-\tau)^{k-j} \\
&= \gamma^k g_n^T S^k r_0 - (-\lambda)^k + \sum_{j=0}^{k} \binom{k}{j} (-\lambda)^{k-j} \lambda^j \\
&= \gamma^k g_n^T S^k r_0 - (-\lambda)^k,
\end{aligned}$$

from which $g_n^T S^k r_0 = (-\tau)^k$. Multiplying (2.6) from the right by the pseudoinverse $[r_0, Sr_0, \ldots, S^n r_0]^+$ and using the fact that g_n lies in the range of $[r_0, Sr_0, \ldots, S^n r_0]$ proves (2.2). Then (2.3) follows in an obvious way. To show (2.4), we denote the N by $n+1$ matrix on the left-hand side and the vector on the right-hand side of (2.6) by R and t, respectively. Then, using (2.5),

$$\begin{aligned}
\|r_n\| \;=\; \|g_n\|^{-1} &\leq \|R\|\, \|t\|^{-1} \\
&\leq\; \|R\|_F\, \|t\|^{-1} \\
&\leq (n+1)^{\frac{1}{2}} \|r_0\|\, \|t\|^{-1},
\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. ∎

Writing (2.6) for the maximal $n = N - l$ in a transposed form gives the upper triangular system for the nonzero entries of $g_{N-l} = [0, \ldots, 0, \chi_l, \chi_{l+1}, \ldots, \chi_N]$,

$$\begin{bmatrix} \rho_l & \rho_{l+1} & \cdots & \rho_N \\ & \rho_l & \cdots & \rho_{N-1} \\ & & \ddots & \vdots \\ & & & \rho_l \end{bmatrix} \begin{bmatrix} \chi_l \\ \chi_{l+1} \\ \vdots \\ \chi_N \end{bmatrix} = \begin{bmatrix} 1 \\ -\tau \\ \vdots \\ (-\tau)^{N-l} \end{bmatrix}. \tag{2.7}$$

The identity (2.5) now immediately implies the following.

COROLLARY 2.2. *With the assumptions and notation of Theorem* 2.1*,*

$$r_{N-l} = \|r_{N-l}\|^2 g_{N-l} \quad \textit{and} \quad \|r_{N-l}\| = \|g_{N-l}\|^{-1}, \tag{2.8}$$

where the nonzero entries of g_{N-l} are determined from (2.7) *by back substitution.*

Theorem 2.1 and Corollary 2.2 show how the GMRES residuals depend on J (particularly on the ratio of λ and γ) *and* the structure of r_0. The bound (2.4) is interesting for large values of $|\tau|$ only, i.e., for diagonally dominant matrices J. In the following examples we give explicit formulas for the nth GMRES residual and its norm for some specific initial residuals.

Example 2.3. Suppose that $r_0 = e_l$ is the lth standard basis vector. Then for $n = 0, 1, \ldots, N - l$, $[r_0, Sr_0, \ldots, S^n r_0] = [e_l, e_{l+1}, \ldots, e_{l+n}]$. Hence (2.2) yields

$$r_n^T = \|r_n\|^2 [0, \ldots, 0, 1, -\tau, \ldots, (-\tau)^n, 0, \ldots, 0],$$

where r_n^T has $l - 1$ leading and $N - n - l$ trailing zeros, respectively. Taking norms on both sides shows that

$$\|r_n\| = \left(\sum_{j=0}^{n} |\tau|^{2j} \right)^{-\frac{1}{2}}, \tag{2.9}$$

i.e., that equality holds in (2.3) with $\sigma_{\min}([r_0, Sr_0, \ldots, S^n r_0]) = 1$. We see that for $r_0 = e_l$, the GMRES residual norms suffer from slow convergence until the very last step whenever $|\tau| \leq 1$. In their unpublished note [2], Eiermann and Ernst give a proof of (2.9) as well as a slightly weaker form of (2.4) based on a formula for the GMRES minimizing polynomial. They also point out that (2.9) is equivalent to the identity

$$\min_{p \in \pi_n} \left\{ \sum_{j=0}^{n} \left| \frac{p^{(j)}(\tau)}{j!} \right|^2 \right\} = \left(\sum_{j=0}^{n} |\tau|^{2j} \right)^{-1},$$

where $p^{(j)}(\tau)$ denotes the jth derivative of the polynomial $p(\tau)$. This can be of interest independent of the GMRES context. ∎

Example 2.4. Consider the particular case $r_0 = e \equiv [1, 1, \ldots, 1]^T$. Then for $n = 1, 2, \ldots, N - 1$,

$$[e, Se, \ldots, S^n e]^+ = \left[e_1, -e_1 + e_2, \ldots, -e_{n-1} + e_n, -e_n + \frac{1}{N-n} S^n e \right]^T,$$

which can easily be verified using the four Moore–Penrose conditions; see, e.g., [11, p. 102]. The GMRES residuals are therefore given by

$$\frac{r_n^T}{\|r_n\|^2} = [1, -\tau, \ldots, (-\tau)^n]\,[e, Se, \ldots, S^n e]^+$$
$$= \left[1+\tau, -(\tau+\tau^2), \ldots, (-1)^{n-1}(\tau^{n-1}+\tau^n), \frac{(-\tau)^n}{N-n}, \ldots, \frac{(-\tau)^n}{N-n}\right],$$

and hence

$$\|r_n\| \;=\; \left(|1+\tau|^2 \sum_{k=0}^{n-1} |\tau|^{2k} \;+\; \frac{|\tau|^{2n}}{N-n} \right)^{-\frac{1}{2}}.$$

Similarly to the case $r_0 = e_l$, the GMRES residual norms converge for $r_0 = e$ slowly until the very last step whenever $|\tau| \le 1$. Unlike in the case $r_0 = e_l$, for $r_0 = e$ the GMRES convergence depends on the sign of the real part of τ. In particular,

$$\|r_n\| \;=\; \left(\tfrac{N-n}{4n(N-n)+1}\right)^{\frac{1}{2}} \qquad \text{for } \tau = 1,$$

$$\|r_n\| \;=\; (N-n)^{\frac{1}{2}} \qquad \text{for } \tau = -1.$$

Thus the stagnation is more severe when $\tau = -1$ (recall that $\|r_0\| = N^{\frac{1}{2}}$). ∎

These examples demonstrate that if $|\tau| \le 1$, then slow convergence of the GMRES residual norms can typically be expected.

3. Tridiagonal Toeplitz matrices. Given nonzero parameters γ, λ, and μ, consider an N by N tridiagonal Toeplitz matrix T,

$$T \;=\; \gamma S + \lambda I + \mu S^T \;\equiv\; \gamma\,(S + \tau I + \zeta S^T), \qquad \tau \;\equiv\; \frac{\lambda}{\gamma}, \quad \zeta \;\equiv\; \frac{\mu}{\gamma}. \tag{3.1}$$

Adding a nonzero superdiagonal μS^T to J in (2.1) causes the resulting matrix T to have N distinct eigenvalues,

$$\sigma_k \;=\; \lambda + \mu\zeta^{-\frac{1}{2}}\,\omega_k\,, \qquad \omega_k \;\equiv\; 2\cos\frac{k\pi}{N+1}\,, \qquad k = 1, \ldots, N, \tag{3.2}$$

with the corresponding normalized eigenvectors given by

$$y_k \;=\; \nu_k\,[\Delta\,u_k]\,, \qquad k = 1, \ldots, N, \tag{3.3}$$

where

$$u_k = \left(\frac{2}{N+1}\right)^{\frac{1}{2}} \left[\sin\frac{k\pi}{N+1}, \ldots, \sin\frac{Nk\pi}{N+1}\right]^T,$$
$$\Delta = \operatorname{diag}\left(\zeta^{-\frac{1}{2}}, \zeta^{-1}, \ldots, \zeta^{-\frac{N}{2}}\right),$$
$$\nu_k = \left(\frac{2}{N+1}\sum_{j=1}^{N} \zeta^{-j}\sin^2\frac{jk\pi}{N+1}\right)^{-\frac{1}{2}};$$

see, e.g., [10, pp. 113–115]. Please note that the matrix $U = [u_1, \ldots, u_N]$ represents the real orthonormal and symmetric eigenvector matrix of any N by N *symmetric* (possibly complex) tridiagonal Toeplitz matrix. The eigenvector matrix $Y = [y_1, \ldots, y_N]$

of T is, apart from the normalization, obtained from U by scaling the rows by the powers of $\zeta^{-\frac{1}{2}}$. Hence the condition number of Y equals $\max(|\zeta|^{\frac{1-N}{2}}, |\zeta|^{\frac{N-1}{2}})$.

When $|\gamma| \approx |\mu|$, meaning $|\zeta| \approx 1$, then Y is well-conditioned and one may base the GMRES convergence analysis on the eigenvalues of T *and* the components of r_0 in the direction of the individual eigenvectors of T.

This paper is motivated by the application of GMRES to convection-diffusion problems with dominating convection [7]. Then the interesting case is characterized by $|\gamma| \approx |\lambda| \gg |\mu|$, meaning $|\tau| \approx 1$ and $|\zeta| \ll 1$. The principal question is, To what extent does the behavior of the GMRES residual for T and a given r_0 resemble the behavior of the GMRES residual for the corresponding J and the same r_0? We focus on this question but we also present some general statements valid for arbitrary nonzero values of γ, λ, and μ.

We would like to stress the following subtle point: When $|\zeta|$ is small, the matrix T can be viewed as a small perturbation of the matrix J. It is therefore tempting to conclude that for each given r_0 the Krylov subspaces generated by T and J are in some sense close to each other. This would imply that generally the GMRES residual norms for J and r_0 are close to the GMRES residual norms for T and r_0. However, it is well known that a small perturbation of a general matrix does not ensure a small change of the Krylov subspace, not even when the matrix is symmetric positive definite. (An instructive example is given below.) It is the structure of J and T that makes such arguments applicable and our analysis possible.

3.1. Explicit mapping. The standard approach to GMRES convergence analysis is based on the eigendecomposition $T = YDY^{-1}$, $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_N)$, giving

$$\|r_n\| = \|Yp_n(D)Y^{-1}r_0\| = \min_{p\in\pi_n} \|Yp(D)Y^{-1}r_0\| \tag{3.4}$$

$$\leq \|Y\|\,\|Y^{-1}\|\,\|r_0\| \min_{p\in\pi_n} \max_k |p(\sigma_k)|; \tag{3.5}$$

see [3, Theorem 5.4] and [9, Proposition 4]. The resulting worst-case bound (3.5) frequently is the basis for discussions of GMRES convergence. However, it does not take into account the fact that for some initial residuals GMRES may behave very differently than for others. In practical problems we work with some particular initial residuals and we are rarely interested in the worst-case behavior. Moreover, when the eigenvector matrix Y is ill-conditioned, then some components of the vector $Y^{-1}r_0$ can be very large, potentially much larger than $\|r_0\|$. On the other hand, the norm of the linear combination $Y\,[p_n(D)Y^{-1}r_0]$ in (3.4) is bounded from above by $\|r_0\|$. This linear combination therefore can contain a significant cancellation, which is not reflected in the minimization problem (3.5). Hence the principal weakness of (3.5) in case of ill-conditioned eigenvectors is not the potentially large multiplicative factor $\|Y\|\,\|Y^{-1}\|$, in our case equal to $\max(|\zeta|^{\frac{1-N}{2}}, |\zeta|^{\frac{N-1}{2}})$. The principal weakness is rather the minimization problem itself. In general, any description of GMRES convergence using the possibly large coordinates $Y^{-1}r_0$ of r_0 in the eigenvector basis, and the mapping from $Y^{-1}r_0$ to the nth GMRES residual r_n, should be applied with proper care for the cancellation that might occur in the presence of close eigenvectors. For more discussion on this topic, see [7] and [12]. In the following we will show the difference when the mapping from $Y^{-1}r_0$ to r_n is replaced by the mapping from r_0 to r_n.

Let us examine the identity

$$r_n = p_n(T)\,r_0 = \Delta\,U p_n(D) U \Delta^{-1} r_0\,. \tag{3.6}$$

We interpret $p_n(T)$ as the mapping from r_0 to r_n, and we denote, for simplicity, $p_n(T) = C_n$. The entries $c_n^{(jk)}$ of C_n, $j,k = 1,2,\ldots,N$, are given by

$$c_n^{(jk)} \;=\; e_j^T C_n\, e_k \;=\; e_j^T \Delta\, U p_n(D) U \Delta^{-1} e_k \;=\; \zeta^{\frac{k-j}{2}} u_j^T p_n(D)\, u_k\,. \tag{3.7}$$

The jth entry of r_n can be expressed as

$$e_j^T r_n \;=\; e_j^T C_n r_0 \;=\; \sum_{k=1}^{N} c_n^{(jk)} \rho_k. \tag{3.8}$$

Note that since T is tridiagonal, the matrices T^n and thus the matrices C_n, for $n = 0,1,\ldots,N-1$, in general have exactly n nonzero subdiagonals and n nonzero superdiagonals. In particular, $c_n^{(jk)} = 0$ for $|j-k| > n$.

THEOREM 3.1. *For each n until GMRES terminates the mapping C_n from r_0 to r_n represents a banded matrix with $2n+1$ nonzero diagonals. We denote the column vectors formed by the entries of each diagonal (ordered from the most outer subdiagonal to the most outer superdiagonal) by*

$$c_n^{(-n)}, c_n^{(-n+1)}, \ldots, c_n^{(0)}, \ldots, c_n^{(n-1)}, c_n^{(n)}.$$

Then the subdiagonals and superdiagonals are related by

$$c_n^{(d)} = \zeta^d\, c_n^{(-d)}\,, \tag{3.9}$$

and the nth GMRES residual can therefore be written in the form

$$r_n \;=\; C_n r_0 \;=\; \sum_{d=0}^{n} [S^d r_0] \odot \begin{bmatrix} 0_d \\ c_n^{(-d)} \end{bmatrix} \;+\; \zeta \sum_{d=1}^{n} \zeta^{d-1} [(S^T)^d r_0] \odot \begin{bmatrix} c_n^{(-d)} \\ 0_d \end{bmatrix}, \tag{3.10}$$

where $a \odot b$ denotes the element-by-element multiple (Hadamard product) of the vectors a and b, and 0_d denotes the zero vector of length d.

Proof. For a given n, and d fixed between 1 and n, the vector $c_n^{(-d)}$ representing the dth subdiagonal consists of the entries $c_n^{(j,j-d)}$, $j = d+1,\ldots,N$. The manipulations

$$\begin{aligned} c_n^{(j,j-d)} &= \zeta^{-\frac{d}{2}}\, u_j^T p_n(D)\, u_{j-d} \\ &= \zeta^{-d}\, (\, \zeta^{\frac{d}{2}}\, u_{j-d}^T\, p_n(D)\, u_j\,) \\ &= \zeta^{-d} c_n^{(j-d,j)} \end{aligned}$$

finish the proof of (3.9). Relation (3.10) is an obvious consequence of (3.9). ∎

When $|\zeta| \ll 1$, the strictly upper triangular part of the mapping C_n is much less significant than its lower triangular part (including the main diagonal). The significance of the superdiagonals is exponentially decreasing with the distance from the main diagonal. Since the proof of Theorem 3.1 does not use that p_n is the GMRES polynomial, the statement can be reformulated for any matrix polynomial $p(T)$, where T is a tridiagonal Toeplitz matrix.

Using (3.8),

$$\|r_n\|^2 \;=\; \sum_{j=1}^{N} |e_j^T r_n|^2 \;=\; \sum_{j=1}^{N} \left| \sum_{k=1}^{N} c_n^{(jk)} \rho_k \right|^2 . \tag{3.11}$$

Since $C_0 = I$, this formula for $n = 0$ reduces to

$$\|r_0\|^2 \;=\; \sum_{j=1}^{N} \left| \sum_{k=1}^{N} c_0^{(jk)} \rho_k \right|^2 \;=\; \sum_{k=1}^{N} |\rho_k|^2 \,. \tag{3.12}$$

A comparison of (3.11) and (3.12) shows that the decrease of the GMRES residual norms is controlled by the behavior of the individual entries $c_n^{(jk)}$ defined in (3.7). Moreover,

$$\|r_n\| \;\le\; \|r_0\| \, \|C_n\| \;\le\; \|r_0\| \, \|C_n\|_F \,. \tag{3.13}$$

These bounds are different from the usual worst-case convergence bounds in that C_n is determined by p_n, which depends on the particular r_0.

The individual entries of the matrices C_n do not decrease monotonically, but their behavior is typically very different from the behavior of the entries of the mapping $Y p_n(D)$ from $Y^{-1} r_0$ to r_n in (3.4). We do not quantify this in a statement but instead present a qualitative argument and experiments. The only term that can seemingly make $c_n^{(jk)}$ large is $\zeta^{\frac{k-j}{2}}$. When, e.g., $|\zeta| \ll 1$, then for $j > k$ this factor becomes large. However, $c_n^{(jk)}$ are the entries of the matrix $C_n = p_n(T)$. Therefore we may expect that the individual nonzero $c_n^{(jk)}$ are of moderate size, and mostly decreasing (although possibly very slowly) with n, which makes the inequalities (3.13) reasonable. The fact that each iteration step n introduces a new nonzero subdiagonal in the mapping from r_0 to r_n hints that when $|\gamma| \approx |\lambda| \gg |\mu|$, i.e., $|\tau| \approx 1 \gg |\zeta|$, the GMRES convergence may be slow.

We emphasize that these considerations about C_n and convergence of GMRES are based on the particular tridiagonal Toeplitz structure of T. On the other hand, when the components of $Y^{-1} r_0$ are large, any approach based on $Y p_n(D)$ can hardly lead to a well-justified insight, even when the special structure of T is exploited.

In Figure 3.1 we plot the values $\log_{10}(|c_n^{(jk)}|)$, $j, k = 1, \ldots, 15$, for $n = 2$, 6, 10, 14, computed when GMRES is applied to the 15 by 15 matrix $T_1 = S + I + 0.01\, S^T$ and the initial residual $r_0 = e$. Corresponding results for T_1 and $r_0 = \texttt{rand}(15, 1)$ are shown in Figure 3.2 (`rand` is the pseudorandom number generator in MATLAB), and Figure 3.3 shows results for the diagonally dominant matrix $T_2 = S + 2I + 0.01\, S^T$ and $r_0 = e$. In Figure 3.4 we plot the respective GMRES residual norms and in Figure 3.5 the values $\|C_n\|_F$, representing an upper bound on $\|r_n\|/\|r_0\|$; cf. (3.13).

In Figures 3.1 to 3.3 we see a decrease of $|c_n^{(jk)}|$ on the superdiagonals of C_n that is exponential in the distance from the main diagonal. Hence in the individual sums $|\sum_{k=1}^{N} c_n^{(jk)} \rho_k|^2$, $j = 1, \ldots, N$, on the right-hand side of (3.11) only the terms for $j \ge k$ play a significant role.

For T_1 and $r_0 = e$ as well as $r_0 = \texttt{rand}(15, 1)$, the significant entries $c_n^{(jk)}$ maintain approximately the same orders of magnitude throughout the GMRES iteration. Correspondingly, the residual norms (solid and dash-dot curves in Figure 3.4) decrease very slowly until the very last step. The initial residual $r_0 = e$ presents a case that yields almost a perfect plateau of significant $|c_n^{(jk)}|$ in every step. The variation of the entries in $r_0 = \texttt{rand}(15, 1)$ causes a larger variation among the absolute values of the significant entries of C_n. For T_2 and $r_0 = e$, GMRES converges faster (cf. the dashed curve in Figure 3.4) since all significant entries of C_n decrease noticeably in magnitude in every step. A comparison of Figure 3.4 and Figure 3.5 illustrates that

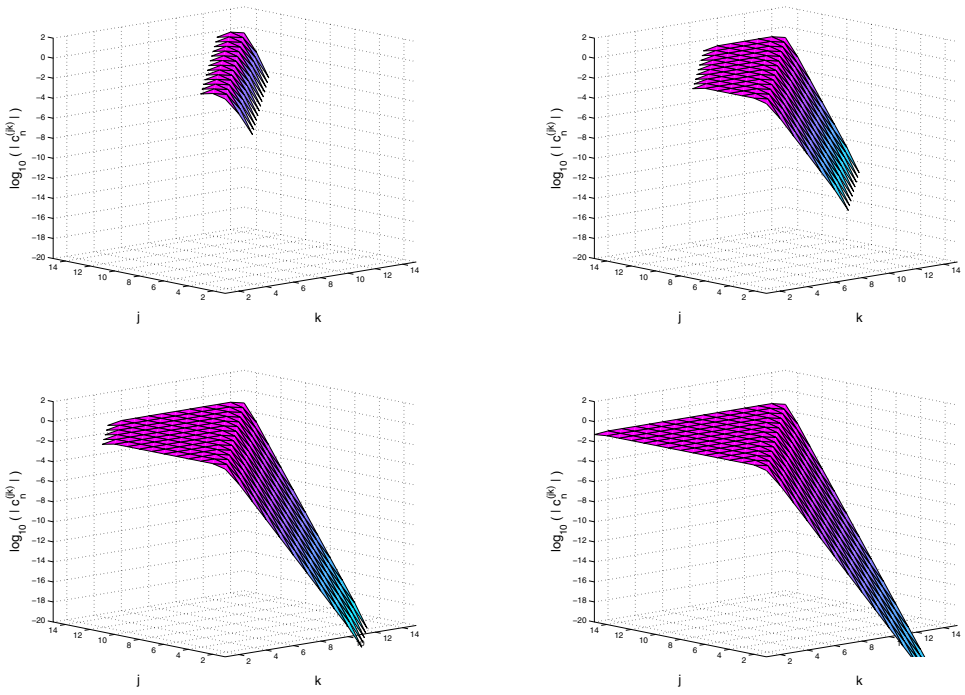

FIG. 3.1. *The values* $\log_{10}(|c_n^{(jk)}|)$ *for* $j,k = 1,\dots,N$ *and* $n = 2$ *(top left),* 6 *(top right),* 10 *(bottom left),* 14 *(bottom right), computed when GMRES is applied to the* 15 *by* 15 *matrix* $T_1 = S + I + 0.01\,S^T$ *and* $r_0 = e$.

the inequalities (3.13) are for our data quite sharp, and, consequently, that there is no significant cancellation among the individual terms in (3.11).

In the following subsection we develop an analogue of Theorem 2.1 for tridiagonal Toeplitz matrices.

3.2. Extension of the bidiagonal analysis. For each scaled (lower bidiagonal) Jordan block J and each r_0 it is easy to see when GMRES terminates: if ρ_l is the first nonzero entry of r_0, then GMRES applied to J and r_0 terminates in exactly $N-l+1$ steps, giving $r_{N-l} \neq 0$ and $r_{N-l+1} = 0$. For a tridiagonal Toeplitz matrix T with nonzero sub- and superdiagonal, the situation is more complicated. Here the total number of GMRES steps for a given nonzero pattern of r_0 can depend on the actual numerical values of its nonzero entries. However, since we are not interested in conditions for termination of GMRES in a given number of steps, we will not specify this number and merely assume that it is greater than $N-l$.

THEOREM 3.2. *Suppose that GMRES is applied to a system with the matrix* $T = \gamma(S + \tau I + \zeta S^T)$ *and the initial residual* $r_0 = [\rho_1,\dots,\rho_N]^T$. *Let* ρ_l *be the first nonzero entry of* r_0. *Moreover, suppose that* r_0 *has at least* $N-l$ *nonzero components in the directions of the individual eigenvectors of the matrix* T *(GMRES does not terminate in the first* $N-l$ *steps). Then for* $n = 0,1,\dots,N-l$ *the GMRES residuals satisfy*

$$r_n^T = \|r_n\|^2\,[1,-\tau,\dots,(-\tau)^n]\,[r_0,(S+\zeta S^T)r_0,\dots,(S+\zeta S^T)^n r_0]^+\,, \tag{3.14}$$

$$\|r_n\| \geq \left(\sum_{j=0}^{n} |\tau|^{2j}\right)^{-\frac{1}{2}} \sigma_{\min}\left([r_0,(S+\zeta S^T)r_0,\dots,(S+\zeta S^T)^n r_0]\right). \tag{3.15}$$

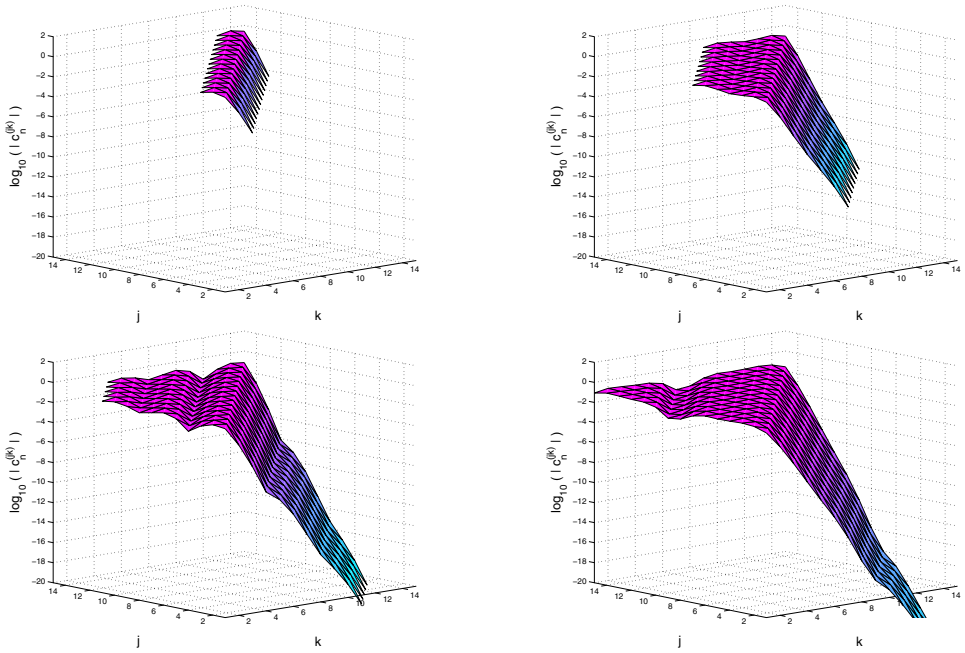

FIG. 3.2. *The values* $\log_{10}(|c_n^{(jk)}|)$ *for* $j,k = 1,\ldots,N$ *and* $n = 2$ *(top left),* 6 *(top right),* 10 *(bottom left),* 14 *(bottom right), computed when GMRES is applied to the* 15 *by* 15 *matrix* $T_1 = S + I + 0.01\,S^T$ *and* $r_0 = \texttt{rand}(15,1)$.

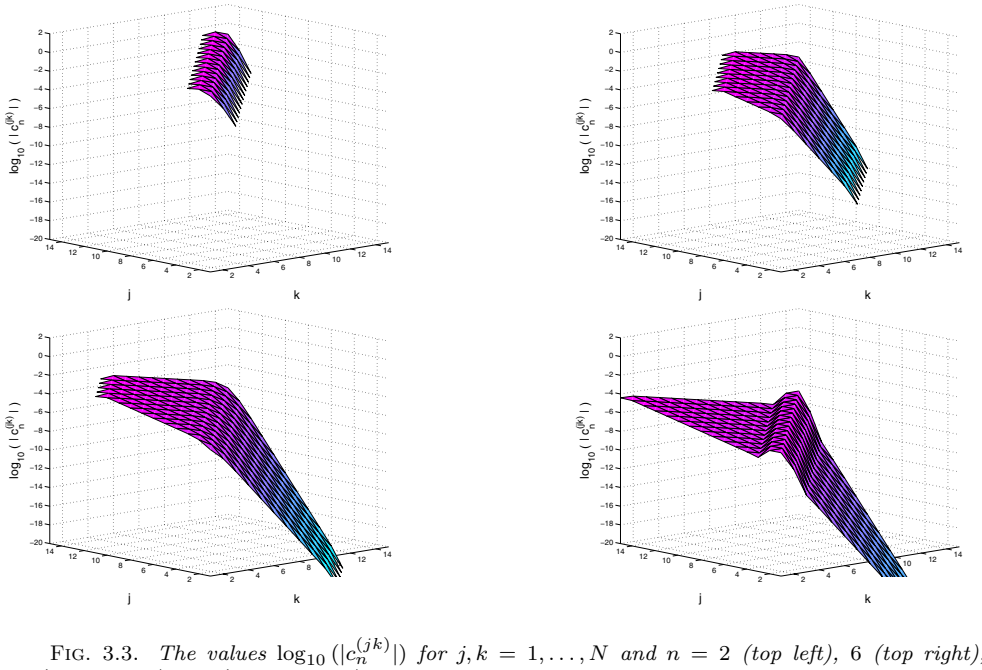

FIG. 3.3. *The values* $\log_{10}(|c_n^{(jk)}|)$ *for* $j,k = 1,\ldots,N$ *and* $n = 2$ *(top left),* 6 *(top right),* 10 *(bottom left),* 14 *(bottom right), computed when GMRES is applied to the* 15 *by* 15 *matrix* $T_2 = S + 2I + 0.01\,S^T$ *and* $r_0 = e$.

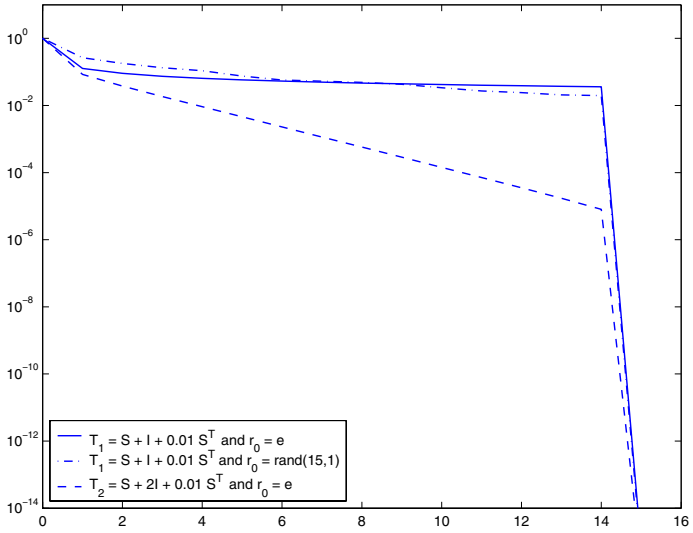

FIG. 3.4. *Residual norms* $\|r_n\|/\|r_0\|$ *of GMRES applied to* $T_1 = S + I + 0.01\,S^T$ *and* $r_0 = e$ *(solid),* T_1 *and* $r_0 = \mathtt{rand}(15,1)$ *(dash-dot),* $T_2 = S + 2I + 0.01\,S^T$ *and* $r_0 = e$ *(dashed).*

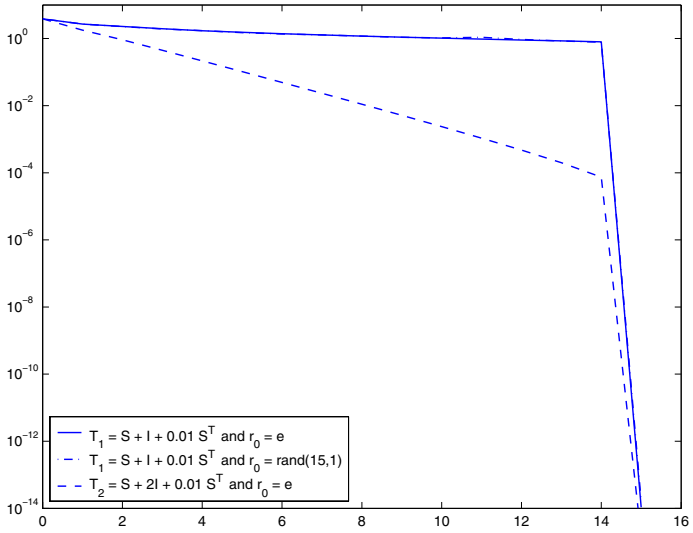

FIG. 3.5. *The values* $\|C_n\|_F$ *for* $T_1 = S+I+0.01\,S^T$ *and* $r_0 = e$ *(solid),* T_1 *and* $r_0 = \mathtt{rand}(15,1)$ *(dash-dot),* $T_2 = S + 2I + 0.01\,S^T$ *and* $r_0 = e$ *(dashed).*

Proof. For $n = 0, 1, \ldots, N-l$, the matrix $[r_0, Tr_0, \ldots, T^n r_0]$ has full column rank. The rest is similar to the proof of Theorem 2.1, with T and $S + \zeta S^T$ taking over the roles of J and S, respectively. Indeed,

$$r_n^T \;=\; \|r_n\|^2\, e_1^T\, [r_0, Tr_0, \ldots, T^n r_0]^+ \;\equiv\; \|r_n\|^2\, g_n^T\,,$$

from which we receive $g_n^T\,[r_0, Tr_0, \ldots, T^n r_0] = e_1^T$. Then

$$0 \;=\; g_n^T T r_0 \;=\; \gamma g_n^T\,(S + \zeta S^T)\, r_0 \;+\; \lambda g_n^T r_0, \qquad \text{i.e., } g_n^T\,(S + \zeta S^T)\, r_0 \;=\; -\tau,$$

and an induction shows that in fact $g_n^T(S+\zeta S^T)^k r_0 = (-\tau)^k$ for $k = 1, 2, \ldots$. Hence

$$g_n^T \left[r_0, (S+\zeta S^T) r_0, \ldots, (S+\zeta S^T)^n r_0 \right] \;=\; \left[1, -\tau, \ldots, (-\tau)^n \right]. \tag{3.16}$$

Now note that

$$g_n \in \mathrm{span}\{r_0, Tr_0, \ldots, T^n r_0\} = \mathrm{span}\{r_0, (S+\zeta S^T) r_0, \ldots, (S+\zeta S^T)^n r_0\}.$$

A multiplication of (3.16) from the right with

$$[r_0, (S+\zeta S^T) r_0, \ldots, (S+\zeta S^T)^n r_0]^+$$

yields (3.14). The lower bound (3.15) is a direct consequence. ∎

Consider, for simplicity, the iteration step $n = N - l$. The principal difference between the cases with J and T is in the form of (2.7) and (3.16). The system of equations (3.16) is for $l \neq 1$ underdetermined, and its system matrix is constructed from r_0 in a much more complicated way than in (2.7). However, the system matrix in (3.16) can be written in the form

$$\begin{aligned} &[r_0, Sr_0, \ldots, S^{N-l} r_0]^T + \zeta \left[0, S^T r_0, \ldots, \zeta^{-1} \left\{ (S+\zeta S^T)^{N-l} - S^{N-l} \right\} r_0 \right]^T \\ &\qquad\qquad \equiv [O, R] + \zeta P, \end{aligned} \tag{3.17}$$

where O denotes the $N-l+1$ by $l-1$ zero matrix, and R denotes the upper triangular matrix described in (2.7). The columns of P^T are given by

$$p_j \;=\; \zeta^{-1} \left\{ (S+\zeta S^T)^j - S^j \right\} r_0 \qquad \text{for } j = 0, 1, \ldots, N-l. \tag{3.18}$$

Since S and S^T do not commute, $(S+\zeta S^T)^j$ cannot be evaluated by the binomial theorem. However, for $j = 1, \ldots, N-l$, this expression can be formally written as

$$(S+\zeta S^T)^j \;=\; \Sigma_{j,0} + \zeta \Sigma_{j,1} + \cdots + \zeta^{j-1} \Sigma_{j,j-1} + \zeta^j \Sigma_{j,j}.$$

Here $\Sigma_{j,k}$ denotes the sum of all possible matrix products involving $j-k$ times the matrix S and k times the matrix S^T. In particular, $\Sigma_{j,0} = S^j$ and $\Sigma_{j,j} = (S^T)^j$. Consequently, for $j = 1, \ldots, N-l$,

$$p_j \;=\; \left(\Sigma_{j,1} + \zeta \Sigma_{j,2} + \cdots + \zeta^{j-2} \Sigma_{j,j-1} + \zeta^{j-1} \Sigma_{j,j} \right) r_0 \,. \tag{3.19}$$

Note that the matrix $\Sigma_{j,k}$ is, for $1 \le j \le N-l$ and $1 \le k \le j$, the sum of $\binom{j}{k}$ products of shift matrices and that $\|\Sigma_{j,k}\| \;\le \binom{j}{k}$. Therefore, assuming $|\zeta| \ll (j-1)^{-1}$,

$$\|p_j\| \;\le\; \|r_0\| \sum_{k=1}^{j} |\zeta|^{k-1} \begin{pmatrix} j \\ k \end{pmatrix} \;=\; j\, \|r_0\| \, (1 + \mathcal{O}(|\zeta|\, j)),$$

where $\mathcal{O}(z)$ is bounded from above by z multiplied by a constant (here close to one). When $|\zeta| \;\ll\; (N-l)^{-\frac{3}{2}}$,

$$\|P\| \;\le\; (N-l)^{\frac{1}{2}} \max_j \|p_j\| \;\le\; (N-l)^{\frac{3}{2}} \, \|r_0\| \left(1 + \mathcal{O}\left((N-l)^{-\frac{1}{2}} \right) \right).$$

The matrix (3.17) can then be considered a small perturbation of the upper triangular system matrix in (2.7), extended by a zero block.

We will now use this perturbation idea for analyzing when GMRES applied to T and r_0 behaves similarly to GMRES applied to J and r_0. As mentioned above, this phenomenon depends in a complicated way on the initial residual r_0; cf. (3.16) and (3.17). Any general result with a nontrivial quantitative meaning can therefore be expected to reflect this complicated nature. In the following we have chosen to preserve a quantitative character of the bounds at the price of an assumption on $R^{-1}P$.

We will use the following notation. The residual for GMRES applied to J with r_0 and the auxiliary vector obtained as a solution of (2.7) will be denoted by $r_n^{(J)}$ and $g_n^{(J)}$, respectively. Analogously, $r_n^{(T)}$, respectively, $g_n^{(T)}$, will denote the residual for GMRES applied to T with r_0, respectively, the minimum norm solution of (3.16). As above, let $r_0 = [\rho_1, \ldots, \rho_N]^T$ with ρ_l being its first nonzero entry. As in Theorem 3.2 we will assume that GMRES applied to T with r_0 does not terminate in the first $N-l$ steps. Then from (3.16),

$$\begin{aligned} g_{N-l}^{(T)} &= ([O,R] + \zeta P)^+ \, [1, -\tau, \ldots, (-\tau)^{N-l}]^T \\ &= \left([O,I] + \zeta R^{-1}P\right)^+ R^{-1}[1, -\tau, \ldots, (-\tau)^{N-l}]^T \\ &= \left([O,I] + \zeta R^{-1}P\right)^+ g_{N-l}^{(J)} . \end{aligned}$$

Taking norms,

$$\|[O,I] + \zeta R^{-1}P\|^{-1} \, \|g_{N-l}^{(J)}\| \;\le\; \|g_{N-l}^{(T)}\| \;\le\; \|([O,I] + \zeta R^{-1}P)^+\| \, \|g_{N-l}^{(J)}\| . \tag{3.20}$$

Assuming that $|\zeta| \, \|R^{-1}P\| < 1$,

$$\|([O,I] + \zeta R^{-1}P)^+\| \;\le\; (1 - |\zeta| \, \|R^{-1}P\|)^{-1} .$$

Considering that $\|r_{N-l}^{(T)}\| = 1/\|g_{N-l}^{(T)}\|$ and $\|r_{N-l}^{(J)}\| = 1/\|g_{N-l}^{(J)}\|$, we proved the following theorem.

THEOREM 3.3. *Using the previous notation and the assumptions of Theorem* 3.2, *let* $|\zeta| \, \|R^{-1}P\| < 1$. *Then the GMRES residuals* $r_{N-l}^{(T)}$ *and* $r_{N-l}^{(J)}$ *satisfy the inequalities*

$$\left(1 + |\zeta| \, \|R^{-1}P\|\right) \|r_{N-l}^{(J)}\| \;\ge\; \|r_{N-l}^{(T)}\| \;\ge\; \left(1 - |\zeta| \, \|R^{-1}P\|\right) \|r_{N-l}^{(J)}\| , \tag{3.21}$$

where R *represents the matrix formed by the last* $N - l + 1$ *columns of the matrix* $[r_0, Sr_0, \ldots, S^{N-l}r_0]^T$ *and* $P = [0, S^T r_0, \ldots, \zeta^{-1}\{(S + \zeta S^T)^{N-l} - S^{N-l}\} \, r_0]^T$.

The main point can be summarized in the following way. Suppose that a scaled Jordan block J is extended to a tridiagonal Toeplitz matrix T by a superdiagonal of sufficiently small modulus (compared to the modulus of the subdiagonal). Assume that GMRES for T and r_0 terminates no earlier than GMRES for J and r_0. Then the convergence of GMRES for T and r_0 will be comparable to the convergence of GMRES for J and r_0. We next consider two examples illustrating our results.

Example 3.4. Suppose that $r_0 = e_1$. Then for J as well as for T the GMRES algorithm terminates in step N. Thus, whenever $|\zeta| \, \|R^{-1}P\| < 1$, the inequalities (3.21) hold with $l = 1$. Note that for $r_0 = e_1$ we have $R = I$ and $\|r_0\| = 1$, so that

$$\begin{aligned} (1 + |\zeta| \, \|P\|) \, \|r_{N-1}^{(J)}\| \;\ge\; \|r_{N-1}^{(T)}\| &\ge (1 - |\zeta| \, \|P\|) \, \|r_{N-1}^{(J)}\| \\ &\ge (1 - |\zeta|(N-1)^{\frac{3}{2}}(1 + \mathcal{O}((N-1)^{-\frac{1}{2}}))) \, \|r_{N-1}^{(J)}\| , \end{aligned}$$

when $|\zeta| \;\ll\; (N-1)^{-\frac{3}{2}}$. ∎

Example 3.5. For $r_0 = e \equiv [1, 1, \ldots, 1]^T$ we can see one of the main differences between the application of GMRES to linear systems with J and with a general extension of J to the tridiagonal Toeplitz matrix T: for any nonzero γ and λ, $\dim \mathcal{K}_N(J, e) = N$, and hence GMRES with J and $r_0 = e$ terminates in step N. For certain nonzero values of λ, γ, and μ, however, $\dim \mathcal{K}_N(T, e) < N$, and hence for certain matrices T and $r_0 = e$ the GMRES algorithm terminates earlier than in step N.

The prime example for the latter case is given by a symmetric T, i.e., $\gamma = \mu$. The normalized eigenvectors of each such matrix are given in (3.3) with $\Delta = I$. These vectors represent discrete sine functions and thus they satisfy certain symmetries. In particular, simple technical manipulations show that

$$\begin{aligned} u_k^T\, e &= \left(\frac{2}{N+1}\right)^{\frac{1}{2}} \sum_{j=1}^{N} \sin\left(\frac{jk\pi}{N+1}\right) \\ &= \left(\frac{2}{N+1}\right)^{\frac{1}{2}} \frac{\cos\left(\frac{k\pi}{2(N+1)}\right) - \cos\left(\frac{(2N+1)k\pi}{2(N+1)}\right)}{2\sin\left(\frac{k\pi}{2(N+1)}\right)} \\ &= \left(\frac{2}{N+1}\right)^{\frac{1}{2}} \frac{\cos\left(\frac{k\pi}{2(N+1)}\right)}{2\sin\left(\frac{k\pi}{2(N+1)}\right)} \left(1 - (-1)^k\right) \\ &= 0 \qquad \text{if } k \text{ is even.} \end{aligned}$$

When $u_k^T r_0 = 0$, the initial residual r_0 has no component in the direction of the eigenvector u_k of T. For a symmetric T and $r_0 = e$, GMRES will therefore terminate in step $N/2$ or $(N+1)/2$ when N is even or odd, respectively. A similar result holds for $\gamma = -\mu$.

In general, however, the normalized eigenvectors of a tridiagonal Toeplitz matrix T are given by $\nu_k[\Delta u_k]$. The components of $r_0 = e$ in the direction of the individual eigenvectors of T are generally given by

$$\nu_k^{-1}\,(u_k^T\,\Delta^{-1}\,e) \;=\; \nu_k^{-1}\,\left(\frac{2}{N+1}\right)^{\frac{1}{2}} \sum_{j=1}^{N} \zeta^{\frac{j}{2}}\,\sin\left(\frac{jk\pi}{N+1}\right)$$

for $k = 1, \ldots, N$. If $|\zeta| \neq 1$, then the initial residual $r_0 = e$ usually has a nonzero component in the direction of *each* of the individual eigenvectors of T. This implies that a very small additive perturbation of a symmetric, even positive definite, tridiagonal Toeplitz matrix by ϵS (or by ϵS^T) may cause GMRES (with $r_0 = e$) to iterate twice as long until it terminates.

Here we are mainly interested in the case $|\zeta| \ll 1$. Then GMRES for J and $r_0 = e$, and usually also for T and $r_0 = e$, terminates in step N. If $|\zeta|\,\|R^{-1}P\| < 1$, then (3.21) holds with $l = 1$. Since $R^{-1} = I - S^T$, we get $\|R^{-1}\| \leq 2$, and since $\|r_0\| = N^{\frac{1}{2}}$, the lower bound in (3.21) yields

$$\begin{aligned} \|r_{N-1}^{(T)}\| &\geq (1 - |\zeta|\,\|(I - S^T)P\|)\,\|r_{N-1}^{(J)}\| \\ &\geq (1 - 2|\zeta|N^2(1 + \mathcal{O}((N-1)^{-\frac{1}{2}})))\,\|r_{N-1}^{(J)}\|\,, \end{aligned}$$

when $|\zeta| \ll (N-1)^{-\frac{3}{2}}$. Numerical examples for this bound are given in section 4. ∎

4. Numerical experiments. The numerical experiments in this section illustrate main points presented and discussed above.

Experiment 4.1. We use the 15 by 15 matrices

$$\begin{aligned} J &= S + I, \\ T_1 &= S + I + 0.01\, S^T, \\ T_2 &= S + I + 0.03\, S^T, \\ T_3 &= S + I + 0.05\, S^T, \\ T_4 &= S + I + 0.999\, S^T, \end{aligned} \tag{4.1}$$

and $r_0 = e$. Since $\dim \mathcal{K}_N(J, e) = N$, and $\dim \mathcal{K}_N(T_j, e) = N$ for all j, GMRES with each of the five matrices and $r_0 = e$ terminates in step N. The relevant values for the application of the bound (3.21) are given in the following table:

j	ζ_j	$\|R_j^{-1}P_j\|$	$\zeta_j \|R_j^{-1}P_j\|$	$1-\zeta_j \|R_j^{-1}P_j\|$
1	0.01	25.58	0.256	0.744
2	0.03	29.50	0.885	0.115
3	0.05	34.33	1.716	$\star$
4	0.999	5.1e+04	5.1e+04	$\star$

For $j = 1, 2$, we have $\zeta_j \|R_j^{-1}P_j\| < 1$, so that the bounds (3.21) are applicable with $l = 1$. The $\star$ for $j = 3, 4$ indicates that since $\zeta_j \|R_j^{-1}P_j\| > 1$, the lower bound in (3.21) is not applicable.

Figures 4.1 and 4.2 show the GMRES residual norms. Since $\tau = 1$, GMRES converges slowly when applied to J (solid). For T_1 (dash-dot) and T_2 (dotted), the GMRES residual norms are very close to the ones for J. The correspondence between $\|r_{14}^{(J)}\|$ and $\|r_{14}^{(T_j)}\|$, $j = 1, 2$, is even closer than predicted by the bounds (3.21). It is also noteworthy that although this bound is not applicable for T_3, the residual norms in this case (dots) are very close to the ones for J as well. The results for T_4 (dashed) show that for a larger perturbation (here $\zeta_4 = 0.999$) the $(N-1)$st GMRES residual norm for a tridiagonal Toeplitz matrix can differ significantly from the corresponding one for the Jordan block.

Experiment 4.2. In Figure 4.3 we used the 15 by 15 matrices

$$\begin{aligned} J &= S + I, \\ T_4 &= S + I + 0.999\, S^T \qquad \text{(as in Experiment 4.1)}, \\ T_5 &= S + I + S^T, \end{aligned}$$

and $r_0 = e$. This experiment demonstrates the difference in the GMRES residual norm curves for T_4 (dash-dot) and T_5 (dotted), despite the fact that $T_4 = T_5 - 0.001\, S^T$ is only a small perturbation of the symmetric matrix T_5. It is interesting to observe that until termination of GMRES for T_5 the convergence curves are very close to each other.

Experiment 4.3. Our last experiment comes from the streamline upwind Petrov–Galerkin (SUPG) discretization of a convection-diffusion model problem with dominating convection. This model problem with rectangular domain, regular grid, and a constant grid aligned convection motivated our work, leading to the results presented in this paper. Here we use it for a short illustration.

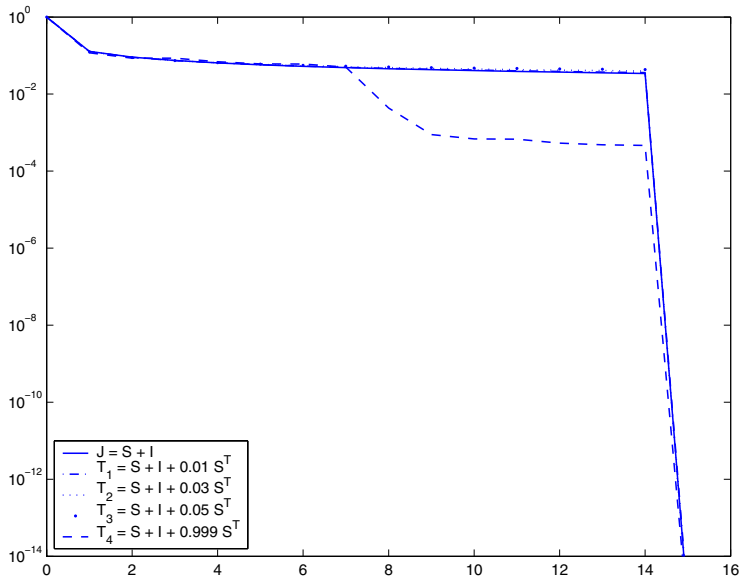

FIG. 4.1. *Residual norms* $\|r_n\|/\|r_0\|$ *of GMRES applied to the five different* 15 *by* 15 *matrices given in* (4.1) *and the initial residual* $r_0 = e$.

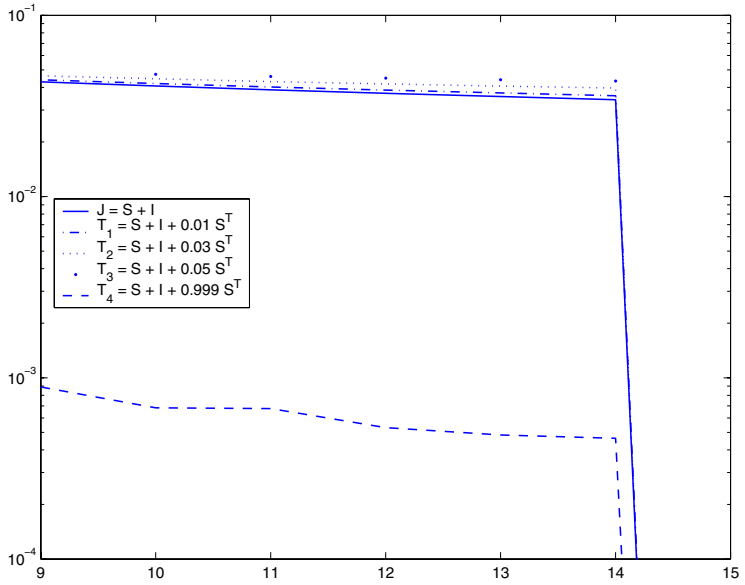

FIG. 4.2. *Close-up of Figure* 4.1.

As explained in [1, 2] and [7], the SUPG discretized model operator can be written as an N^2 by N^2 block-diagonal matrix with N by N nonsymmetric tridiagonal Toeplitz blocks $T_j = \gamma_j(S + \tau_j I + \zeta_j S^T)$, $j = 1, \ldots, N$, on its diagonal. Example values for $|\tau_j|$ and $|\zeta_j|$, as well as the corresponding quantities related to (3.21) with $N = 15$ and $r_0 = e_1$, are given in Table 4.1.

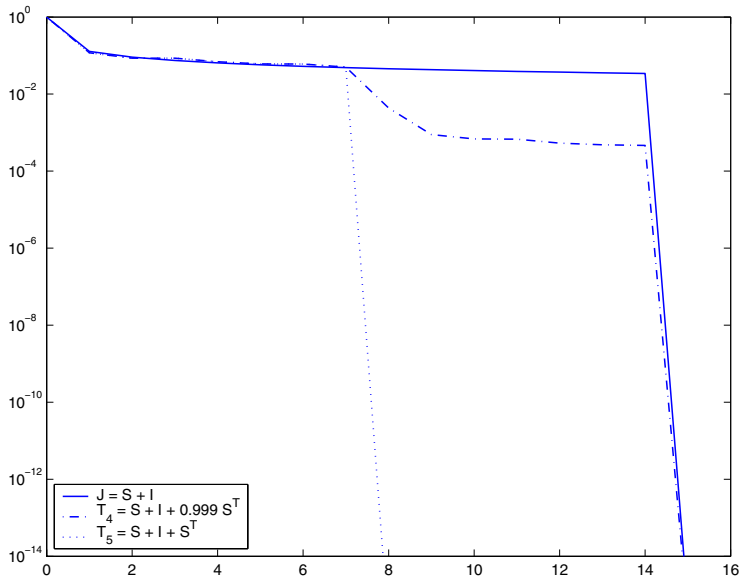

FIG. 4.3. *Residual norms* $\|r_n\|/\|r_0\|$ *of GMRES applied to* 15 *by* 15 *matrices* $J = S + I$ *(solid),* $T_5 = S + I + S^T$ *(dotted),* $T_4 = T_5 - 0.001\, S^T$ *(dash-dot), and the initial residual* $r_0 = e$.

TABLE 4.1
Example values derived from the SUPG discretized convection-diffusion model operator.

j	$\lvert\tau_j\rvert$	$\lvert\zeta_j\rvert$	$\|R_j^{-1}P_j\|$	$\lvert\zeta_j\rvert\,\|R_j^{-1}P_j\|$
1	1.0052	0.0010	13.0002	0.0134
2	1.0209	0.0042	13.0040	0.0544
3	1.0481	0.0096	13.0211	0.1252
4	1.0881	0.0176	13.0708	0.2303
5	1.1431	0.0286	13.1874	0.3774
6	1.2162	0.0432	13.4295	0.5808
7	1.3116	0.0623	13.8989	0.8663
8	1.4348	0.0870	14.7740	1.2847
9	1.5925	0.1185	16.3739	1.9402
10	1.7923	0.1585	19.2798	3.0551
11	2.0409	0.2082	24.5496	5.1108
12	2.3392	0.2678	34.0035	9.1077
13	2.6735	0.3347	50.1498	16.7855
14	3.0033	0.4007	74.1263	29.6989
15	3.2564	0.4513	99.9102	45.0870

Figure 4.4 shows the GMRES residual norm curves for the matrices T_j, $j = 1, \ldots, 15$, and $r_0 = e_1$. For small j we have $|\tau_j| \approx 1$, which leads to very slow convergence of GMRES for the corresponding scaled Jordan blocks and $r_0 = e_1$. Simultaneously there holds $|\zeta_j| \ll 1$, so that the convergence for the respective tridiagonal Toeplitz matrices T_j with the same r_0 is comparably slow. With increasing j, both $|\tau_j|$ and $|\zeta_j|$ increase, and the speed of convergence of GMRES for T_j (as well as for the corresponding Jordan blocks) and $r_0 = e_1$ increases significantly. The slow convergence of GMRES for the matrices T_j with small indices j translates into an initial phase of slow convergence of GMRES for the SUPG discretized model operator. The detailed exposition is beyond the scope of this paper, and we refer an interested reader to [7].

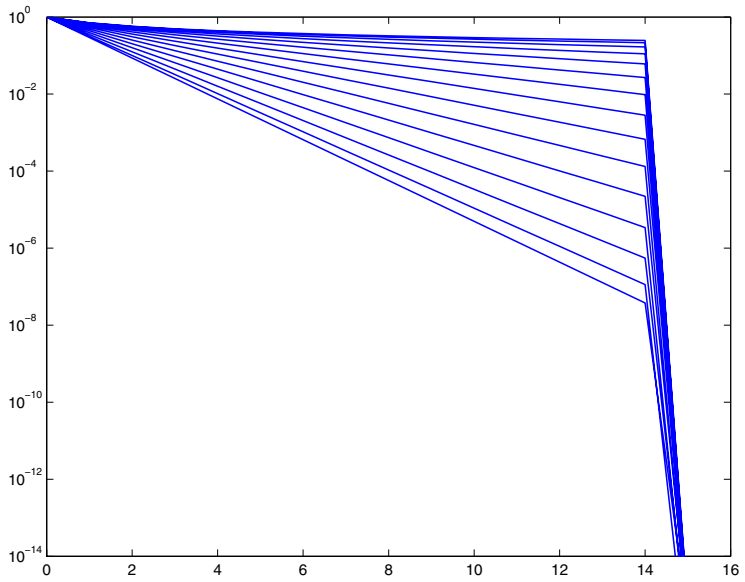

FIG. 4.4. *Residual norms* $\|r_n\|/\|r_0\|$ *of GMRES applied to* 15 *by* 15 *matrices* T_j, $j = 1, \ldots, 15$, *representing the tridiagonal Toeplitz blocks on the diagonal of a SUPG discretized convection-diffusion model operator (see* [7]*) with the initial residuals* $r_0 = e_1$.

5. Conclusions and outlook. Consider GMRES convergence for a matrix A and a given initial residual r_0. Let B be a small perturbation of A. Does the assumption that B is *sufficiently close* to A guarantee that the GMRES residuals for A and r_0 are at every iteration step close to the GMRES residuals for B and r_0? A related question, although in a different context and without the dependence on the initial residual, which we consider vital, was recently also considered by Huhtanen and Nevanlinna [4]. Motivated by applications in convection-diffusion problems [7], our paper studies this question for $A \equiv J = \gamma S + \lambda I$ and $B \equiv T = J + \mu S^T$, and for this particular matrix A and its particular perturbation B it gives an affirmative answer. In general, however, the answer is complicated, which is documented by a nonsymmetric perturbation of a symmetric tridiagonal Toeplitz matrix. To what extent our results can be applied to GMRES convergence analysis of more general problems, e.g., when there exists a well-conditioned transformation of the system matrix into a block diagonal form with tridiagonal blocks, remains the subject of further work.

Acknowledgments. We thank Michael Eiermann and Oliver Ernst for sharing their unpublished notes [2] and for very stimulating discussions and advice about the subject matter of this paper. We also thank the anonymous referee for several suggestions that helped to improve the presentation of the paper. All numerical experiments in this paper were performed using MATLAB [13].

REFERENCES

[1] M. EIERMANN, *Semiiterative Verfahren für nichtsymmetrische lineare Gleichungssysteme*, Habilitationsschrift, Universität Karlsruhe, Karlsruhe, 1989.
[2] M. EIERMANN AND O. ERNST, *GMRES and Jordan blocks,* private communication, 2002.
[3] H. C. ELMAN, *Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations*, Ph.D. thesis, Yale University, New Haven, CT, 1982.

[4] M. Huhtanen and O. Nevanlinna, *Minimal decompositions and iterative methods*, Numer. Math., 86 (2000), pp. 257–281.
[5] I. C. F. Ipsen, *Expressions and bounds for the GMRES residual*, BIT, 40 (2000), pp. 524–535.
[6] J. Liesen, M. Rozložník, and Z. Strakoš, *Least squares residuals and minimal residual methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1503–1525.
[7] J. Liesen and Z. Strakoš, *GMRES convergence analysis for a convection-diffusion model problem*, SIAM J. Sci. Comput., submitted.
[8] C. C. Paige and Z. Strakoš, *Residual and backward error bounds in minimum residual Krylov subspace methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1898–1923.
[9] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[10] G. D. Smith, *Numerical solution of partial differential equations*, 2nd ed., Clarendon Press, Oxford, UK, 1978.
[11] G. W. Stewart and J. G. Sun, *Matrix perturbation theory*, Academic Press, Boston, 1990.
[12] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.
[13] The MathWorks, Inc., *MATLAB* 6.5*, Release* 13, Natick, MA, USA, 2002.