SPECTRAL FACTORIZATION OF THE KRYLOV MATRIX AND CONVERGENCE OF GMRES *

ILYA ZAVORIN†

Abstract. Is it possible to use eigenvalues and eigenvectors to establish accurate results on GMRES performance? Existing convergence bounds, that are extensions of analysis of Hermitian solvers like CG and MINRES, provide no useful information when the coefficient matrix is almost defective. In this paper we propose a new framework for using spectral information for convergence analysis. It is based on what we call the spectral factorization of the Krylov matrix. Using the new apparatus, we prove that two related matrices are equivalent in terms of GMRES convergence, and derive necessary conditions for the worst-case right-hand side vector. We also show that for a specific family of application problems, the worst-case vector has a compact form. In addition, we present numerical data that shows that two matrices that yield the same worst-case GMRES behavior may differ significantly in their average behavior.

Key words. GMRES , Krylov methods, convergence, spectral factorization, iterative methods

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

1. Introduction.

The GMRES method has been used extensively during the last two decades for solving non-Hermitian linear systems. Nevertheless, its convergence properties are still poorly understood. In particular, it is unclear what role eigenvalues and eigenvectors of the coefficient matrix play in convergence of the algorithm or if it is possible to use spectral information to derive accurate convergence results.

These issues has been investigated to some extent in the original work of Saad and Schultz [17]. Suppose we apply GMRES to the linear system

$$Ax = b, \quad A \in \mathcal{C}^{n\times n}, \;\; x, b \in \mathcal{C}^n, \tag{1.1}$$

where A has the spectral decomposition

$$A = V\Lambda V^{-1}, \;\; \Lambda = diag(\lambda), \;\; \lambda = [\lambda_1, \; \dots, \; \lambda_n]^T, \;\; \lambda_j \in \mathcal{C} \setminus \{0\}. \tag{1.2}$$

We denote the GMRES iterate at step m by x_m, $0 \le m \le n$, with x_0 being the initial guess. The corresponding residual is defined by $r_m = b - Ax_m$. The first GMRES convergence result, which appeared in [17], was an extension of convergence analysis of methods like CG [9] and MINRES [15] applied to Hermitian systems. It bounds the ratio of the norms of r_m and r_0 as follows,

$$\frac{\|r_m\|}{\|r_0\|} \le \kappa_2(V) \min_{p_m(t)\in\Pi_m} \max_{i=1,\dots,n} |p_m(\lambda_i)|, \tag{1.3}$$

where Π_m is the set of all polynomials of degree m that equal one at zero, $\kappa_2(V) = \|V\|\|V^{-1}\|$ is the condition number of the eigenvector matrix and $\|\cdot\|$ is the vector or matrix Euclidean norm. When A is normal then $\kappa_2(V) = 1$ and the bound (1.3) is sharp [6], i.e. for every A and every m, there exists a right-hand side vector b for which (1.3) becomes an equality. When A is nonnormal, however, this bound becomes much less useful because the right-hand side expression can be made arbitrarily large by taking an almost-singular V. Also, except for some special cases [1, 13, 5], the above minimax expression is hard to compute or even estimate. Alternative bounds have been developed, that are based on other

*This work was partially supported by the National Science Foundation under Grants CCR 95-03126 and CCR-97-32022.

† Applied Mathematics and Scientific Computing Program, University of Maryland, College Park, MD 20742 (iaz@cs.umd.edu)

characteristics of the coefficient matrix such as the field of values [3] and pseudo-spectrum [20]. However, none of these bounds is immune to the problems of inaccuracy and computational complexity.

More recently Greenbaum and her colleagues discovered that eigenvalues alone cannot explain `GMRES` behavior [8, 7]. However, in this paper, as well as in an accompanying manuscript [23], we demonstrate that if we combine information about the eigenvalues and eigenvectors of A, as well as the right-hand side b, via a Krylov matrix, we can derive explicit expressions for `GMRES` convergence measures and obtain accurate results on performance of the algorithm.

The paper consists of two parts. In Section 2, we express `GMRES` convergence at each iteration in terms of eigenvalues λ, eigenvectors V and the right-hand side represented in the column basis of V. Then, in Sections 3 through 7, we apply the developed apparatus to analysis of `GMRES` convergence. To some extent, the work presented in the second part of the paper is a generalization of the results presented in [23], where we apply the new machinery to a rather extreme case of `GMRES` convergence called stagnation, when the method makes no progress during the first several iterations.

Most of the existing literature on convergence of `GMRES` is devoted to derivation of precise upper bounds of the quantity $\|r_m\|/\|r_0\|$. In this paper, we, too, present convergence bounds and discuss their accuracy, but we also go beyond this. For instance, in Section 5, where we present the main result of the paper, we demonstrate that two related matrices yield the same worst-case behavior at every step of `GMRES`, and establish necessary conditions for the worst-case vector b. In Section 6, we show that the worst-case right-hand side can sometimes be expressed in a very compact form in terms of some of the quantities derived in Section 2. We also demonstrate that our framework may be applied indirectly to the case of a defective A, provided this A can be expressed as a limit of a parametrized sequence of diagonalizable matrices. Finally, in Section 7, we present numerical data that suggest that when overall `GMRES` behavior is measured by its average convergence, it may yield results different from those produced by worst-case analysis.

When A is Hermitian, `GMRES` is equivalent to `MINRES`. Therefore all results presented in this paper that apply to `GMRES` for Hermitian A hold for `MINRES` as well.

2. GMRES Convergence Measures. The main purpose of this section is to develop a new approach for analysis of `GMRES` performance based on spectral information of the matrix A. First, we discuss relevant properties of the `GMRES` algorithm in Section 2.1. Then, we devote Section 2.2 to derivation of an explicit expression for a `GMRES` convergence measure based on what we call the spectral factorization of the Krylov matrix associated with application of `GMRES` to the problem (1.1).

2.1. GMRES and Its Basic Properties. Given a linear system (1.1) and an initial guess x_0 with the residual $r_0 = b - Ax_0$, at iteration m, `GMRES` computes an approximation $x_m \in x_0 + \mathcal{K}_m(A, r_0)$ to the true solution $\hat{x} = A^{-1}b$, where $\mathcal{K}_m(A, r_0) = span\{r_0,\ Ar_0,\ \ldots,\ A^{m-1}r_0\}$ is the Krylov subspace of dimension m. Without loss of generality we can assume that $x_0 = 0$ and so $r_0 = b$. The iterate x_m is chosen in such a way as to minimize the Euclidean norm of the residual $r_m = b - Ax_m$, i.e. the `GMRES` residual $r_m(A, b)$ at step m satisfies

$$\|r_m(A,b)\| = \min_{x \in \mathcal{K}_m(A,b)} \|b - Ax\|. \tag{2.1}$$

When there is no ambiguity, we denote $r_m(A, b)$ by r_m. We also denote by `GMRES(`A`,`b`)` application of `GMRES` to the linear system (1.1) with $x_0 = 0$, or by `GMRES(`A`)` when the right-hand side vector is unspecified. We assume infinite precision, so our derivations do not depend on a specific implementation of the method.

Throughout this paper, various quantities associated with `GMRES` iteration m are denoted by letters subscripted by m. The subscript is dropped for the same quantities at step $m = n - 1$.

The norm of r_m is a nonincreasing function of m. Given a matrix A and a vector b, we say that GMRES(A,b) *terminates* in m steps if $r_m = 0$ and $r_{m-1} \neq 0$. A fundamental property of GMRES is that $r_m(A,b) \neq 0$ iff $dim(\mathcal{K}_{m+1}(A,b)) = m+1$. Thus, while analyzing GMRES performance at iteration m, it is sufficient to consider those vectors b that yield the Krylov matrices $K_{m+1}(A,b) = [b \; Ab \; \ldots \; A^m b]$ of rank $m+1$. Another important property is that the matrix $K_k(A,b)$ is rank-deficient for any $b \in \mathcal{C}^n$ if a diagonalizable matrix A has fewer than k distinct eigenvalues. Therefore we assume that A has at least $m+1$ distinct eigenvalues.

For a given $A \in \mathcal{C}^{n \times n}$, we call $b' \in \mathcal{C}^n$ the *worst-case right-hand side* at step m (with respect to the matrix A) if, for any $b \in \mathcal{C}^n$, $\|r_m(A,b')\|/\|b'\| \geq \|r_m(A,b)\|/\|b\|$.

2.2. GMRES Convergence Measures in Terms of Spectral Decomposition of $K_{m+1}(A,b)$. In this section, we demonstrate that, when A is diagonalizable, a GMRES(A,b) convergence measure can be expressed in terms of eigencomponents of A and the right-hand side vector.

DEFINITION 2.1. *The* GMRES(A,b) *performance measure h_m at iteration m is defined by* $h_m \equiv \|r_m\|/\|r_0\| = \|r_m\|/\|b\| \in [0,1]$.

The function h_m expresses a common way of measuring progress of an iterative method during the first m iterations with its small and large values corresponding to good and bad convergence, respectively.

We now state an important result due to Ipsen [11, Theorem 2.1] that represents one of the two main building blocks which allow us to develop the apparatus presented in Section 2 [1]. It is expressed in terms of the Moore-Penrose pseudoinverse of a full-rank matrix $K_{m+1}(A,b)$ which is well-defined and unique, and can be calculated by [19, 12]

$$K_{m+1}^{\dagger} = (K_{m+1}^H K_{m+1})^{-1} K_{m+1}^H \in \mathcal{C}^{(m+1)\times n}.$$

THEOREM 2.2. *Let A be diagonalized by (1.2) and let $b \in \mathcal{C}^n$. Assume that at step m, $rank(K_{m+1}(A,b)) = m+1$. Define*

$$c_m = (K_{m+1}^{\dagger})^H e_1 \in \mathcal{C}^n, \tag{2.2}$$

which, in case $m = n-1$, simplifies to $c = K^{-H} e_1$. Then the residual of GMRES(A,b) *at step m satisfies* $\|r_m\| = \|c_m\|^{-1}$.

Thus we can rewrite the performance measures of GMRES(A,b) in terms of components of the Krylov matrix and its pseudoinverse as

$$h_m = (\|c_m\|\|b\|)^{-1}. \tag{2.3}$$

This implies that progress of GMRES during the first m iterations can be measured by the angle between c_m and b. More specifically,

COROLLARY 2.3. *For given A, b and m such that the matrix $rank(K_{m+1}) = m+1$, the following relationships between b and c_m hold*

1. *The two vectors can be computed from each other as follows,*

$$\begin{aligned} c_m &= (K_{m+1}(K_{m+1}^H K_{m+1})^{-2} K_{m+1}^H) b \\ b &= (K_{m+1} K_{m+1}^H) c_m \end{aligned}$$

2. $c_m^H b = 1$.
3. $h_m = \cos \angle(c_m, b)$.

[1] Ipsen's result is a special case of those presented by Stewart in [18, Sections 3 and 4].

Proof: To prove Item 1 we first observe that

$$b = K_{m+1}e_1. \tag{2.4}$$

Also, $K^\dagger_{m+1}K_{m+1} = I$ and so it follows that $K^\dagger_{m+1}b = (K^\dagger_{m+1}K_{m+1})e_1 = e_1$. We combine this result with the definition (2.2) of c_m and obtain

$$\begin{aligned} c_m &= (K^\dagger_{m+1})^H e_1 = (K^\dagger_{m+1})^H K^\dagger_{m+1} b \\ &= (K_{m+1}(K^H_{m+1}K_{m+1})^{-1})((K^H_{m+1}K_{m+1})^{-1})K^H_{m+1})b \\ &= (K_{m+1}(K^H_{m+1}K_{m+1})^{-2}K^H_{m+1})b. \end{aligned}$$

The formula for b in terms of c_m is derived similarly by observing that $K^H_{m+1}(K^\dagger_{m+1})^H$ equals identity. To establish Item 2 we combine (2.2) with (2.4) and write

$$c^H_m b = (e^H_1(K^H_{m+1}K_{m+1})^{-1}K^H_{m+1})(K_{m+1}e_1) = e^H_1 e_1 = 1.$$

Finally, to obtain Item 3, we expand the Euclidean inner product as follows,

$$\cos\angle(c_m, b) = (c^H_m b)/(\|c_m\|\|b\|) = 1/(\|c_m\|\|b\|) = h_m.$$

□

We now show that the Krylov matrix associated with GMRES(A,b) at step m can be factorized using eigencomponents of A and the right-hand side vector b represented in the eigenvector basis. This factorization, which we call the *spectral factorization of* $K_{m+1}(A, b)$, is the second major building block which allows us to express convergence of the method in terms of eigenvalues, eigenvectors and the right-hand side. Although this factorization has appeared in literature before (e.g. [11, Proof of Theorem 4.1]), to our knowledge, it has never been stated or proved as a separate result.

THEOREM 2.4. *Let the nonsingular matrix $A \in \mathcal{C}^{n\times n}$ be diagonalized by (1.2) and let $b \in \mathcal{C}^n$. Let $y = V^{-1}b$. Then, regardless of its column rank, the $n \times (m+1)$ Krylov matrix K_{m+1} associated with* GMRES(A,b) *at step m can be factored as*

$$K_{m+1} = VYZ_{m+1}, \tag{2.5}$$

where Z_{m+1} is the Vandermonde matrix computed from eigenvalues of A as follows,

$$Z_{m+1} = \begin{pmatrix} 1 & \lambda_1 & \dots & \lambda_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_n & \dots & \lambda_n^m \end{pmatrix} = (\, e \quad \Lambda e \quad \dots \quad \Lambda^m e\,).$$

Conversely, for a diagonalizable matrix A, take $y \in \mathcal{C}^n$ and compute K_{m+1} by (2.5). Then this matrix is the Krylov matrix associated with GMRES(A,Vy) *at step m.*

Proof: See [23]. □

We now combine the spectral factorization (2.5) with equations (2.3), (2.2), and (2.4) and obtain an explicit expression for GMRES(A,b) convergence at step m,

$$h_m(V, \lambda, y) = \left(\|VYZ_{m+1}(Z^H_{m+1}\overline{Y}WYZ_{m+1})^{-1}e_1\| \; \|Vy\|\right)^{-1}, \tag{2.6}$$

where $W = V^HV$ and $e_1 \in \mathcal{C}^{m+1}$. The case $m = n-1$ deserves special attention since then the expression (2.6) significantly simplifies. First, observe that $rank(K) = n$ iff the eigenvalues of A are distinct and all entries of $y = V^{-1}b$ are nonzero. Then $K^\dagger = K^{-1}$ and it follows from (2.2) that

$$c_{n-1} = c = K^{-H}e_1 = (VYZ)^{-H}e_1 = V^{-H}\overline{Y}^{-1}Z^{-H}e_1.$$

Denote the elements of the first column of Z^{-H} by $u_j,\ 1 \leq j \leq n$. From [10, Section 21.1] it follows that they can be explicitly computed from the eigenvalues of A by

$$u_j = (-1)^{n+1} \mathrm{conj}\left(\prod_{\substack{k=1\\k\neq j}}^{n} \frac{\lambda_k}{\lambda_j - \lambda_k} \right). \tag{2.7}$$

Since the first column of Z is e, these elements also satisfy $u_1 + \ldots + u_1 = 1$. Let us denote the mapping from $\mathcal{C}^n$ to $\mathcal{C}^n$, defined elementwise by (2.7), by $G(\lambda)$. Also, let $u = G(\lambda)$, i.e. u represents the conjugate transpose of the first row of Z^{-1}. Then it follows that $c = V^{-H}\overline{Y}^{-1}u$ and

$$h(V, \lambda, y) = (\|V^{-H}\overline{Y}^{-1}u\| \; \|Vy\|)^{-1}. \tag{2.8}$$

We note that $h_m(V, \lambda, y)$ is invariant to the following scalings.

1. Scaling of the vector b. We may, therefore, assume that $\|b\| = 1$.
2. Scaling of eigenvalues λ. Thus we may assume that $\lambda_1 = 1$.
3. Column scaling of V, i.e. $h_m(V, \lambda, y) = h_m(VD, \lambda, D^{-1}y)$ for any nonsingular diagonal $D \in \mathcal{C}^{n\times n}$. Thus we may assume that columns of V have unit length.
4. Pre-multiplication of V by a unitary matrix, i.e. $h_m(V, \lambda, y) = h_m(PV, \lambda, y)$ for any unitary $P \in \mathcal{C}^{n\times n}$. Thus it is sufficient to consider only matrices V with their SVD of the form $V = SQ^H$.

We conclude this section with a statement of a general property of the worst-case right-hand side vector in terms of the convergence measures.

LEMMA 2.5. *The vector $b^* \in \mathcal{C}^n$ is the worst-case right-hand side for* `GMRES(A,b)` *at step m iff the vector $y^* = V^{-1}b^*$ satisfies $h_m(V, \lambda, y^*) \geq h_m(V, \lambda, y)$ for any other $y \in \mathcal{C}^n$. In other words, y^* is the global maximizer of $h_m(V, \lambda, y)$.*

In the remaining sections, we apply the developed apparatus to analysis of `GMRES`.

3. New GMRES Convergence Bounds. In this section we assume that all vectors b have unit length, which implies that the vectors $y = V^{-1}b$ are restricted to the hyper-ellipsoid surface $E_V = \{\ y \in \mathcal{C}^n \mid y^H W y = b^H b = \|b\|^2 = 1\}$. Our goal is to establish accurate upper bounds on the performance measure $h(V, \lambda, y) = \|c\|^{-1} = \|V^{-H}\overline{Y}^{-1}u\|^{-1},\ y \in E_V$, of `GMRES` at step $m = n-1$, as well as to extend these bounds to arbitrary steps.

THEOREM 3.1. *For $y \in E_V$, the following bounds hold,*

$$h(V, \lambda, y) \leq \hat{h}(V, \lambda, y) \equiv \|V\| / \|\overline{Y}^{-1}u\| \tag{3.1}$$

$$\leq \tilde{h}(V, \lambda, y) \equiv \|V\| \|Y\| / \|u\|. \tag{3.2}$$

Proof: To obtain (3.1), we estimate $\|c\|$ from below as follows,

$$c = V^{-H}\overline{Y}^{-1}u \quad \Longleftrightarrow \quad V^H c = \overline{Y}^{-1}u \quad \Rightarrow \quad \|V^H\| \|c\| \geq \|\overline{Y}^{-1}u\|. \tag{3.3}$$

We now let $t = \overline{Y}^{-1}u$, apply a sequence of steps similar to (3.3), and get $\|Y\|\|t\| \geq \|u\|$, which yields (3.2). $\square$

When A is normal, V is unitary, which yields $\|V^H c\| = \|c\|$, and so the bound (3.1) becomes an equality for every $y \in E_V$. Suppose A is non-normal. Let us assume that the eigenvector matrix has the form $V = SQ^H$, where $S = diag(s_1, \ldots, s_n)$ and $s_1 \geq \ldots \geq s_n$. Then the right singular vectors of $V^H = QS$ are $e_1, \ldots, e_n$, and so

$$\|V^H c\| = \|V^H\|\|c\| \quad \Longleftrightarrow \quad c = \alpha\, e_1,$$

where $\alpha \in \mathcal{C}$ is a scaling constant that ensures that the b corresponding to the c is of unit norm. For what right-hand side vector does the equality hold? We expand

$$V^H(\alpha\, e_1) = \alpha Q S e_1 = \alpha s_1 q_1 = \overline{Y}^{-1} u,$$

where $q_1 = [q_{11}, \ldots, q_{n1}]^T$ is the right singular vector of V corresponding to the largest singular value s_1. Let $u = [u_1, \ldots, u_n]^T$ and $y = [y_1, \ldots, y_n]^T$. We conclude that the elements of the vector y for which the bound (3.1) coincides with $h(V, \lambda, y)$, have the form

$$y_j = conj\left(\frac{u_j}{\alpha s_1 q_{j1}}\right), \quad j = 1, \ldots, n,$$

where α is chosen appropriately. We compared (3.1) and (3.2) with the bound (1.3) on a set of low-dimensional nonsymmetric real matrices A with real positive eigenvalues (see [22, Section 5.1]). Tests showed that both (3.1) and (3.2) gave smaller bounds then (1.3). In addition, they have certain theoretical advantages. The bound (3.1) depends on the right-hand side and not just on its norm, while bound (3.2) better separates components that correspond to eigenvalues and eigenvectors of A and the right-hand side b.

Using the same general approach we can obtain a bound for the performance measure $h_m(V, \lambda, y)$ for $m = 1, \ldots, n-2$. Since $c_m = VYZ_{m+1}(Z_{m+1}^H \overline{Y} WYZ_{m+1})^{-1} e_1$, we have

$$V^H c_m = WYZ_{m+1}(Z_{m+1}^H \overline{Y} WYZ_{m+1})^{-1} e_1,$$

and so

$$h_m(V, \lambda, y) \leq \hat{h}_m(V, \lambda, y) \equiv \|V\| / \|WYZ_{m+1}(Z_{m+1}^H \overline{Y} WYZ_{m+1})^{-1} e_1\|. \tag{3.4}$$

Although (3.4) still appears to be tighter than (1.3), it does not really offer any theoretical advantages over the exact expression (2.6). It is obviously less accurate than (2.6) and yet its components are not as well separated as they are in (3.2). Separation is difficult since in general, unlike the regular inverse, the Moore-Penrose pseudoinverse of a matrix product is not a product of pseudoinverses. Thus finding a better estimate for an arbitrary step of GMRES remains an open question.

4. The Worst-Case Right-Hand Side at Step $m = n - 1$ for Real Symmetric A. In this section we assume that A is real symmetric. We prove that the worst-case y at GMRES step $m = n - 1$ can be computed from the vector $u = G(\lambda)$, where λ contains eigenvalues of A. From Lemma 2.5 it follows that this is equivalent to finding a global minimizer of $h(V, \lambda, y)^{-2}$. Rather than looking at this problem as an unconstrained minimization problem, we restrict y to E_V, which, in the case of symmetric A, becomes the unit sphere in $\mathcal{R}^n$. This yields an optimization problem with a nonlinear objective function and one nonlinear equality constraint. The first-order necessary and second-order sufficient conditions for y to be a (local) minimizer are expressed in terms of the gradient and Hessian of h^{-2} (see, e.g. [14, Section 14.5]). We prove by construction that the necessary condition is satisfied and is actually sufficient for the global minimizer.

LEMMA 4.1. *Let $A \in \mathcal{R}^{n\times n}$ be symmetric with distinct eigenvalues. Let $u = G(\lambda)$. Consider real vectors b. Then the worst-case vectors and the worst-case performance of* `GMRES(A)` *at step $m = n-1$ are*

$$
\begin{aligned}
&y_{worst} = \gamma\ [\ \pm\sqrt{|u_1|},\ \ldots,\ \pm\sqrt{|u_n|}\ ]^T, \\
&h_{worst}(V,\lambda,y_{worst}) = \left(\sum_{j=1}^{n}|u_j|\right)^{-1},
\end{aligned}
\tag{4.1}
$$

where $\gamma \in \mathcal{R}$ is any nonzero scaling constant.

Proof: Since V is orthogonal, finding the worst-case behavior of `GMRES(A)` at step $m = n-1$ is equivalent to solving the following constrained minimization problem

$$
\begin{aligned}
\min_{y} &\quad f(y), \\
\text{subject to} &\quad g(y) = 0
\end{aligned}
$$

where

$$
f(y) = \left(\frac{u_1}{y_1}\right)^2 + \ldots + \left(\frac{u_n}{y_n}\right)^2 \quad \text{and} \quad g(y) = y_1^2 + \ldots + y_n^2 - 1.
$$

Note that $f(y) = h(V,\lambda,y)^{-2}$ restricted to the domain E_V by $g(y)$. To establish the first-order condition, we compute the Lagrangian $L(y,\mu) = f(y) + \mu g(y)$ and its gradient with respect to y,

$$
\frac{\partial L(y,\mu)}{\partial y_j} = -2\left(\frac{u_j^2}{y_j^3} - \mu y_j\right), \qquad 1 \le j \le n.
$$

We can assume that $y_j \neq 0,\ 1 \le j \le n$, otherwise $f(y)$ becomes infinitely large. We find zeros of the gradient of the Lagrangian by solving

$$
u_j^2 - \mu y_j^4 = 0 \quad \Longleftrightarrow \quad (y_j^*)^4 = \frac{u_j^2}{\mu} \quad \Longleftrightarrow \quad (y_j^*)^2 = \frac{|u_j|}{\sqrt{\mu}} \quad \Longleftrightarrow \quad y_j^* = \pm\sqrt{\frac{|u_j|}{\sqrt{\mu}}}.
$$

The next step is to determine the value of the Lagrange multiplier μ that would ensure that the solution $y^* = [y_1^*,\ldots,y_n^*]^T$ satisfies the constraint. We solve

$$
0 = g(y^*) = \left(\frac{1}{\sqrt{\mu}}\sum_{j=1}^{n}|u_j|\right) - 1
$$

for μ and obtain $\sqrt{\mu^*} = \sum_{j=1}^{n}|u_j|$ and so $\mu^* = \left(\sum_{j=1}^{n}|u_j|\right)^2$. Therefore all the points y^* where the gradient of the Lagrangian vanishes have the form

$$
y^* = \frac{1}{\sqrt{\sum_{j=1}^{n}|u_j|}}\ [\ \pm\sqrt{|u_1|},\ \ldots,\ \pm\sqrt{|u_n|}\ ]^T \tag{4.2}
$$

We evaluate the objective function $f(y)$ at y^* and obtain

$$f^* = f(y^*) = \sum_{j=1}^{n} \frac{u_j^2}{(y_j^*)^2} = \sum_{j=1}^{n} u_j^2 \frac{\sum_{j=1}^{n} |u_j|}{|u_j|}$$
$$= \left(\sum_{j=1}^{n} |u_j| \right) \left(\sum_{j=1}^{n} \frac{u_j^2}{|u_j|} \right) = \left(\sum_{j=1}^{n} |u_j| \right)^2 .$$

Note that because all variables appear squared in $f(y)$, the value $f(y^*)$ is the same regardless of the sign pattern of y^*.

Now let us consider a certain aspect of the behavior of $h(V, \lambda, y)$ over its respective domain E_V, where V may or may not be unitary. Fix an arbitrary $j = 1, \ldots, n$ and consider the intersection of E_V with the coordinate plane $y_j = 0$. It is a hyper-ellipsoid surface of dimension $n-1$ that splits E_V in half. On one side of this dividing surface, all vectors $y \in E_V$ have $y_j < 0$, while on the other side $y_j > 0$. Along the dividing surface, $h(V, \lambda, y) = 0$. Thus we can always think of E_V as a union of 2^n nonoverlapping patches. Each patch is characterized by the following two properties, (i) along its boundaries, $h(V, \lambda, y) = 0$ and (ii) all points $y \in E_V$ that belong to a given patch have the same sign pattern, and no point outside of it has that pattern. We conclude that along the patch boundaries, h^{-2} is infinitely large. Thus, unless it is identically equal to infinity over a given patch, which is impossible, it must have at least one minimizer inside that patch. This implies that in the symmetric case, when $h(V, \lambda, y)^{-2} = f(y)$, the points y^* defined by (4.2) constitute global minimum points of $f(y)$, since these are the only points with zero gradient and they all produce the same $f(y^*)$.

Finally, we observe that since `GMRES` is invariant to scaling of the b and y, we can rewrite (4.2) as (4.1). $\square$

5. Equivalence of A and A^H.

We start with a definition of equivalence of two matrices.

DEFINITION 5.1. *Let $A, \tilde{A} \in \mathcal{C}^{n\times n}$ and let $b, \tilde{b} \in \mathcal{C}^n$. By $r_m(A, b)$ and $r_m(\tilde{A}, \tilde{b})$ we denote residuals of* `GMRES(`A`,`b`)` *and* `GMRES(`$\tilde{A}$`,`$\tilde{b}$`)` *at step m, respectively. We say that A and $\tilde{A}$ are equivalent at step m in terms of* `GMRES` *convergence if*

$$\max_{b\neq 0} \frac{\|r_m(A, b)\|}{\|b\|} = \max_{\tilde{b}\neq 0} \frac{\|r_m(\tilde{A}, \tilde{b})\|}{\|\tilde{b}\|}.$$

The two matrices are equivalent if they are equivalent at every step m, $1 \le m \le n$.

Note that in general the worst-case right-hand side vector is different for every m. The goal of this section is to show that if A is diagonalizable then it is equivalent to A^H. First, let us define some notation. Columns of V are (right) eigenvectors of A whereas the columns of V^{-H} are its left eigenvectors. On the other hand, since $A^H = V^{-H}\overline{\Lambda}V^H$, the columns of V^{-H} are also right eigenvectors of A^H. We also observe that if Z_{m+1} is the $n \times (m+1)$ Vandermonde matrix computed from eigenvalues of A, then $\bar{Z}_{m+1}$ is the matrix associated with A^H.

Let us denote the right-hand side b associated with A by b_R, and the corresponding vectors c_m and y by c_R and y_R. Similarly, the vectors associated with A^H will be denoted by b_L, c_L, and y_L. Throughout the rest of the paper, we denote by $H_m(V, \lambda, y_R)$ the reciprocal of $h_m(V, \lambda, y_R)$. Also,

$$\begin{array}{rclcl} H_R(y_R) & = & H_m(V, \lambda, y_R) & = & \|c_R\| \, \|b_R\|, \\ H_L(y_L) & = & H_m(V^{-H}, \bar{\lambda}, y_L) & = & \|c_L\| \, \|b_L\|, \end{array}$$

where

$$\begin{array}{rclrcl} c_R & = & VY_R Z_{m+1}(Z_{m+1}^H \overline{Y}_R W Y_R Z_{m+1})^{-1} e_1, & b_R & = & V y_R, \\ c_L & = & V^{-H} Y_L \overline{Z}_{m+1}(Z_{m+1}^T \overline{Y}_L W^{-1} Y_L \overline{Z}_{m+1})^{-1} e_1, & b_L & = & V^{-H} y_L. \end{array} \tag{5.1}$$

We denote matrices $K_{m+1}(A, b_R)$ and $K_{m+1}(A^H, b_L)$ by K_R and K_L, respectively. Finally, if $b_R(b_L)$ is a worst-case right-hand side vector for A (A^H) then this vector, as well as the associated c_R (c_L) will be denoted by b_R^{worst} (b_L^{worst}) and c_R^{worst} (c_L^{worst}), respectively. Before we state the general equivalence result, we state two auxiliary lemmas and prove one of them.

LEMMA 5.2. *Let $\lambda = [\lambda_1, \ldots, \lambda_{m+1}, \lambda_{m+2}, \ldots, \lambda_n]^T \in \mathcal{C}^n$ contain nonzero eigenvalues with $\lambda_1, \ldots, \lambda_{m+1}$ being distinct and let Z_{m+1} be the $n \times m+1$ Vandermonde matrix computed from λ. Let $t \in \mathcal{C}^n$ solve*

$$Z_{m+1}^H t = e_1. \tag{5.2}$$

Then t contains at least $m+1$ nonzero entries corresponding to $\lambda_1, \ldots, \lambda_{m+1}$.

Proof: In order to prove the result, it is sufficient to assume that the vector t has the form $t = [t_1, \ldots, t_{m+1}, 0, \ldots, 0]^T$ and to prove that $t_j \neq 0,\ 1 \leq j \leq m+1$. We let $p = n - m - 1$. We observe that equation (5.2) can be rewritten in the form

$$\tilde{Z}_{m+1}^H t_1 + \tilde{Z}_p^H t_2 = e_1,$$

where

$$\tilde{Z}_{m+1} = \begin{pmatrix} 1 & \lambda_1 & \ldots & \lambda_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{m+1} & \ldots & \lambda_{m+1}^m \end{pmatrix} \in \mathcal{C}^{m+1 \times m+1},$$

$$\tilde{Z}_p = \begin{pmatrix} 1 & \lambda_{m+2} & \ldots & \lambda_{m+2}^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_n & \ldots & \lambda_n^m \end{pmatrix} \in \mathcal{C}^{p \times m+1},$$

and

$$t_1 = [\ t_1,\ \ \ldots,\ \ t_{m+1}\ ]^T \in \mathcal{C}^{m+1}, \quad t_2 = 0 \in \mathcal{C}^p.$$

Since eigenvalues $\lambda_1, \ldots, \lambda_{m+1}$ are distinct, the square matrix $\tilde{Z}_{m+1}$ is invertible and so $t_1 = \tilde{Z}_{m+1}^{-H} e_1 = \tilde{u}$, where $\tilde{u} = G([\lambda_1, \ldots, \lambda_{m+1}]) \in \mathcal{C}^{m+1}$. Since all eigenvalues are nonzero, we conclude from (2.7) that t_1 contains no zeros. This completes the proof. $\square$

LEMMA 5.3. *Let $M \in \mathcal{C}^{k \times n}$, $k \leq n$, $\tilde{c} \in \mathcal{C}^n$ and $d \in \mathcal{C}^k$. Suppose that $M\tilde{c} = d$. Finally, let $rank(M) = k$. Then $\tilde{c}$ can always be written in the form*

$$\tilde{c} = c_0 + c_N, \quad c_0 = M^\dagger d, \quad c_N \in \mathcal{N}(M), \quad c_N \perp c_0, \tag{5.3}$$

where $\mathcal{N}(M)$ is the kernel of M.

Proof: See [22, Section 2.3, Lemma 2]. $\square$

We are now ready to demonstrate that A and A^H are equivalent in terms of the worst-case `GMRES` behavior.

THEOREM 5.4. *Let A be diagonalizable by (1.2) and nonsingular. Let $1 \leq m \leq n-1$ be fixed. Then* `GMRES(`A`)` *and* `GMRES(`A^H`)` *achieve the same worst-case behavior at step m. Furthermore, let $b_R = b_R^{worst}$, the right-hand side vector that yields the worst-case behavior of* `GMRES(`A`)` *at step m. Compute the corresponding c_R^{worst} and set $b_L = c_R^{worst}$. Then $b_L = b_L^{worst}$, i.e. it is the worst-case right-hand side for* `GMRES(`A^H`)`*. Moreover, the resulting c_L^{worst} satisfies $c_L^{worst} = b_R^{worst}$, i.e. the vectors b_R^{worst}, c_R^{worst}, b_L^{worst}, and c_L^{worst} are cross-equal.*

Remark: The converse of the above statement is not true in general. In other words, take an arbitrary b_R, compute the corresponding c_R, set $b_L = c_R$ and compute c_L. Cross-equality of the four vectors, i.e. the relationship $c_L = b_R$, does *not* imply that b_R or b_L is the worst-case right-hand side vector for `GMRES(`A`)` or `GMRES(`A^H`)`, respectively.

Proof of Theorem 5.4: Pick an arbitrary b_R such that the matrix K_R has full rank. This yields a (unique) vector c_R. Define $a_R \equiv (Z_{m+1}^H \overline{Y}_R W Y_R Z_{m+1})^{-1} e_1$. Equations (5.1) imply that

$$V Y_R Z_{m+1} a_R = c_R, \quad Z_{m+1}^H \overline{Y}_R (W Y_R Z_{m+1} a_R) = e_1.$$

We now *set* $b_L = c_R$. Since $b_L = V^{-H} y_L$, we can rewrite the two equations above as follows,

$$W Y_R Z_{m+1} a_R = y_L, \quad Z_{m+1}^H \overline{Y}_R (W Y_R Z_{m+1} a_R) = e_1.$$

We combine the two equations and obtain

$$e_1 = Z_{m+1}^H \overline{Y}_R y_L = Z_{m+1}^T \overline{Y}_L y_R = Z_{m+1}^T \overline{Y}_L V^{-1} V y_R = K_L^H b_R.$$

From Lemma 5.2 is follows that $rank(K_L) = m + 1$. We therefore apply Lemma 5.3 and write $b_R = (K_L^H)^\dagger e_1 + t_L = c_L + t_L$, where $t_L \in \mathcal{N}(K_L^H)$ and $t_L \perp c_L$. By the Pythagorean equation,

$$\begin{array}{rcl} H_R^2 & = & \|b_R\|^2 \|c_R\|^2 \;=\; (\|c_L\|^2 + \|t_L\|^2)\|b_L\|^2 \\ & = & \|c_L\|^2\|b_L\|^2 + \|t_L\|^2\|b_L\|^2 \;=\; H_L^2 + \|t_L\|^2\|b_L\|^2 \\ & \geq & H_L^2, \end{array}$$

with equality holding iff $t_L = 0$.

We now repeat the procedure with A and A^H switched. In other words, we take the b_L from above (which, of course, yields the c_L and H_L from above), and let $\tilde{b}_R = c_L$. This, in turn, yields $\tilde{c}_R$ such that

$$b_L = \tilde{c}_R + t_R, \quad t_R \perp \tilde{c}_R.$$

We conclude that the corresponding $\tilde{H}_R$ satisfies

$$H_R^2 \geq H_L^2 \geq \tilde{H}_R^2. \tag{5.4}$$

Now let $b_R = b_R^{worst}$. Then for any $\tilde{b}_R$,

$$H_R^2 \leq \tilde{H}_R^2. \tag{5.5}$$

We conclude that equations (5.4) and (5.5) both can be true iff

$$t_L = t_R = 0 \qquad \Longleftrightarrow \qquad b_R = c_L = \tilde{b}_R.$$

This proves that $H_R^{worst} \geq H_L^{worst}$. Switching A and A^H and using the same argument, we can show that $H_L^{worst} \geq H_R^{worst}$ which implies that the two quantities are equal. $\square$

We do not know how to distinguish b_R^{worst} from other vectors b_R that yield cross-equality. We do know, however, how to calculate the latter vectors using a simple iterative technique. Let us again examine the double inequality (5.4). It implies that if we start with an *arbitrary* b_R and perform the following sequence of steps,

$$b_R \Rightarrow c_R = b_L \Rightarrow c_L = \tilde{b}_R, \tag{5.6}$$

The CE **Algorithm:**
0. Take any $b_R^{(1)} \in \mathcal{C}^n$ such that $rank(K_R) = m+1$. Set $k=1$.
1. Set $y_R^{(k)} = V^{-1} b_R^{(k)}$.
2. Set $K_R = VY_R^{(k)} Z_{m+1}$ and $c_R^{(k)} = (K_R^H)^\dagger e_1$.
3. Set $b_L^{(k)} = c_R^{(k)}$ and $H_R(k) = \|b_R^{(k)}\|\|c_R^{(k)}\|$.
4. Set $y_L^{(k)} = V^H b_L^{(k)}$.
5. Set $K_L = V^{-H} Y_L^{(k)} \overline{Z}_{m+1}$, $c_L^{(k)} = (K_L^H)^\dagger e_1$, and $H_L(k) = \|b_L^{(k)}\|\|c_L^{(k)}\|$.
6. If $\|c_L^{(k)} - b_R^{(k)}\|$ is sufficiently small, exit.
7. Set $k = k+1$. Set $b_R^{(k)} = c_L^{(k-1)}$. Go to Step 1.

TABLE 5.1
The CE *Algorithm: An Iterative Technique for Finding b_R with Cross-Equality*

and compute H_R and $\tilde{H}_R$ at b_R and $\tilde{b}_R$, respectively, then $H_R \geq \tilde{H}_R$ with equality holding iff b_R is a cross-equality point. If we now complete the loop by setting $b_R = \tilde{b}_R$ and repeat (5.6) recursively, we will obtain a sequence of monotonically decreasing values H_R. In other words, for $k = 1, 2, \ldots$, consider the sequences $\{H_R(k)\}$ and $\{H_L(k)\}$ generated by the iterative algorithm shown in Table 5.1. We call it the CE ("Cross-Equality") algorithm. As $k \to \infty$, $H_R(k)$ monotonically decreases. Since it is also bounded below by H_R^{worst}, it converges to a finite limit. This implies that

$$\lim_{k\to\infty} (H_R(k) - H_R(k+1)) = 0.$$

It follows that in the limit, the above algorithm converges to a cross-equality point b_R for *any* initial guess $b_R^{(1)}$. Clearly, the same applies to $\{H_L(k)\}$ and $b_L^{(k)}$.

Note that the CE algorithm may be used when A is defective. In this case we skip steps 1 and 4 and compute matrices K_R and K_L at steps 2 and 5 directly from matrices A and A^H and vectors b_R and b_L, instead of using their spectral factorizations. Theorem 5.4 was proved only for the diagonalizable case and thus convergence of the CE algorithm is not guaranteed when A is defective. Nevertheless, when we applied it to a few test matrices, like the convection-diffusion matrix with $\alpha = 1$ discussed in the next section, the algorithm always converged.

Experiments suggest that, given a diagonalizable matrix A, at step m, the set of all vectors b_R that give cross-equality is a manifold of dimension $m+1$. It remains an open question as to whether or not it is possible to devise a method similar to that described in Table 5.1, but which is defined on the set of b_R with cross-equality, and which would converge to b_R^{worst}.

When $m = n-1$, $\mathcal{N}(K_R^H) = \mathcal{N}(K_L^H) = \{0\}$. It follows that the CE algorithm always converges in one iteration and *every* $b_R \in \mathcal{C}^n$ yields cross-equality. In fact, in [22] we prove that in this case the associated vectors y_R and y_L satisfy $\overline{Y_R} y_L = G(\lambda)$.

6. A Model Problem: The One-Dimensional Convection-Diffusion Equation. The purpose of this section is to study the worst-case GMRES behavior at step $m = n-1$, when applied to a family of coefficient matrices that arise in discretizations of the one-dimension convection-diffusional equation. Just like in Section 4, we are looking for the vector y that satisfies the first-and second-order conditions in terms of the gradient and Hessian of h_m^{-2} [14, Section 10.2].

6.1. The Worst Case for the Convection-Diffusion Matrix. We consider the one-parameter family of matrices $A = A(\alpha)$ that arises in the discretization of the one-dimensional convection-diffusion equation [4]. Standard discretization schemes like centered differences produce a coefficient matrix of the form [4]

$$A_{CD} = A(\alpha) = tridiag(-1-\alpha, 2, -1+\alpha) \in \mathcal{R}^{n\times n}, \tag{6.1}$$

where $0 \le \alpha \le 1$ for stability reasons [16]. When $\alpha = 0$, which corresponds to the diffusion-dominated case, the matrix is symmetric. In the convection-dominated case $\alpha = 1$, A_{CD} is a single Jordan block, i.e. it is a "maximally defective" matrix with a single eigenvalue 2 repeated n times and a single eigenvector. When $0 < \alpha < 1$, the matrix is nonsymmetric diagonalizable with distinct eigenvalues.

The eigenvalues $\lambda_{CD} = [\lambda_1, \ldots, \lambda_n]$ and eigenvectors V_{CD} of $A(\alpha)$ have the form

$$\lambda_j = \lambda_j(\alpha) = 2\ (1 - \sqrt{1-\alpha^2}\cos(\frac{\pi j}{n+1})), \qquad 1 \le j \le n, \tag{6.2}$$

$$V_{CD} = V(\alpha) = D_{CD} Q_{CD}, \tag{6.3}$$

where $D_{CD} = diag(\delta, \ldots, \delta^n)$, $\delta = \sqrt{(1+\alpha)/(1-\alpha)}$ and $Q_{CD} = [q_{jk}]$ is a symmetric orthogonal matrix computed as follows

$$q_{jk} = \sqrt{\frac{2}{n+1}}\sin(\frac{\pi jk}{n+1}), \qquad 1 \le j,k \le n. \tag{6.4}$$

Unlike Section 4, here we study the worst case as an unconstrained problem, i.e. we do not restrict the vector y to the surface E_V. Thus, in order to establish necessary and sufficient conditions for a minimizer of $h(V,\lambda,y)^{-2}$ we have to compute the gradient and Hessian of the objective function. We do this for the case of arbitrary sets of distinct nonzero eigenvalues λ and eigenvectors V.

THEOREM 6.1. *Let $A \in \mathcal{R}^{n\times n}$ be nonsingular and diagonalizable with distinct real eigenvalues. Define $f(V,\lambda,y) = h(V,\lambda,y)^{-2}$. Also, for a given $y \in \mathcal{R}^n$, define $t = Y^{-1}u$, where $u = G(\lambda)$. This implies that $t_j = u_j/y_j$, $1 \le j \le n$. In addition, define the following scalars and matrices. Let $F_1(y) = (y^T W y) \in \mathcal{R}$ and $F_2(y) = (t^T W^{-1} t) \in \mathcal{R}$. Let $G_1(y) = 2Wy$ and $G_2(y) = -2D_1 W^{-1} t$. Let*

$$D_1 = diag([\frac{t_1^2}{u_1}, \ldots, \frac{t_n^2}{u_n}]), \quad D_2 = diag([\frac{t_1^3}{u_1^2}, \ldots, \frac{t_n^3}{u_n^2}]), \quad D_3 = diag(W^{-1}t),$$

where $W = V^T V$. Then the gradient and Hessian of $f(V,\lambda,y)$ with respect to y can be written as follows,

$$\nabla_y f(V,\lambda,y) = F_1(y)\ G_2(y) + F_2(y)\ G_1(y) \tag{6.5}$$

$$\nabla_y^2 f(V,\lambda,y) = 2F_2(y)\ W + 2F_1(y)\ (D_1 W^{-1} D_1 + 2D_2 D_3) + G_2(y)G_1(y)^T + G_1(y)G_2(y)^T \tag{6.6}$$

Proof: From (2.8) it follows that $f(V,\lambda,y) = F_1(y)F_2(y)$. We observe that vectors $G_1(y)$ and $G_2(y)$ are simply gradients of $F_1(y)$ and $F_2(y)$ with respect to y. Expressions (6.5) and (6.6) are obtained by applying the rule of differentiation of a product to the function $f(V,\lambda,y)$. □

It turns out that in the case of the convection-diffusion matrix, the right-hand side vectors defined by (4.1) set the gradient of $f(V,\lambda,y)$ to zero even when $\alpha > 0$. More precisely,

LEMMA 6.2. *Let V_{CD} and λ_{CD} be defined by (6.2) and (6.3), respectively. Let y be defined by (4.1). Then $\nabla_y f(V_{CD}, \lambda_{CD}, y) = 0$. Also, regardless of the actual sign pattern of y, the corresponding vectors b and c satisfy*

$$\|c\| = \delta^{-(n+1)}\|b\|,$$

and therefore

$$h(V_{CD}, \lambda_{CD}, y) = \delta^{(n+1)} \|b\|^{-2} = \delta^{-(n+1)} \|c\|^{-2}. \tag{6.7}$$

Proof: See [22]. □

Although Lemma 6.2 implies that points y computed by (4.1) satisfy the first-order necessary condition for a minimizer, it does not imply that the Hessian $\nabla_y^2 f(V, \lambda, y)$ is positive-semidefinite at y. In fact, numerical experiments indicate that most of the 2^n points computed by (4.1) are nothing more than saddle points. There is one exception, though. There appears to be one point at which the second-order condition does appear to be satisfied. More precisely, empirical data suggest that for every n and every $0 \le \alpha < 1$, the vector

$$y_{CD} = y(\alpha) = \sqrt{|u_{CD}|} = [\ \sqrt{+u_1},\ \sqrt{-u_2},\ \sqrt{+u_3},\ \sqrt{-u_4},\ \ldots\ ]^T \tag{6.8}$$

is a minimizer of $f(V_{CD}, \lambda_{CD}, y)$. Furthermore, the tests indicate that this is a *global* minimizer. The third important piece of numerical evidence we found was that $h(V_{CD}, \lambda_{CD}, y(\alpha))$ grows monotonically as α goes from zero to one. Since $\kappa_2(V_{CD})$ also grows monotonically with α [4], it appears that the one-dimensional convection-diffusion family of matrices is an example of the negative effect of conditioning of eigenvectors on convergence of `GMRES`.

6.2. The Worst-Case Vector for $\alpha = 1$, an $n = 3$ Example. It is often possible to represent a defective matrix as a limit of a certain parametrized family of diagonalizable matrices as the set of parameters approaches a limit point. The one-dimensional convection-diffusion family of matrices $A(\alpha)$ defined by (6.1) provides one such example with a single parameter α and its limit value of one. Therefore it is logical to expect that analysis of behavior of `GMRES` applied to the defective matrix can be done by considering limits of related quantities corresponding to the diagonalizable matrices. In this section we demonstrate this for the convection-diffusion matrix of size $n = 3$.

The matrix V_{CD} defined by (6.3) has the form

$$V_{CD} = V(\alpha) = \frac{1}{\sqrt{2}} \begin{pmatrix} \delta & \delta & \delta \\ \delta^2 & 0 & -\delta^2 \\ \delta^3 & -\delta^3 & \delta^3 \end{pmatrix}.$$

The worst-case right-hand side vector $b(\alpha)$ equals

$$b(\alpha) = \frac{1}{2\sqrt{2}(1-\alpha)^2} \begin{bmatrix} (1-\alpha)(2\sqrt{1+\alpha^2} + \sqrt{\gamma_2} + \sqrt{\gamma_1}) \\ -\sqrt{\gamma_0}(\sqrt{\gamma_2} - \sqrt{\gamma_1}) \\ (1+\alpha)(-2\sqrt{1+\alpha^2} + \sqrt{\gamma_2} + \sqrt{\gamma_1}) \end{bmatrix}. \tag{6.9}$$

We now use (6.7) to compute $h(\alpha) = h(V(\alpha), \lambda_{CD}, y(\alpha))$ and obtain [2]

$$h(\alpha) = \frac{(1+\alpha^2)^2}{3 + \alpha^4 + \alpha^2(2 + \sqrt{2(1+\alpha^2)}) - 2\alpha\sqrt{1+\alpha^2}(\sqrt{\gamma_2} + \sqrt{\gamma_1})}. \tag{6.10}$$

In order to determine the worst-case behavior of `GMRES(`$A(1)$`)` we compute the limit of $h(\alpha)$ as $\alpha \to 1^-$. We obtain

$$\lim_{\alpha \to 1^-} h(\alpha) = \frac{64}{89} \approx 0.719101.$$

[2] Computations were performed using `Mathematica` version 4 [21].

The components of $b(\alpha)$ given by (6.9) grow infinitely large as α approaches unity. Therefore in order to find the worst-case right-hand side for the defective case, we first scale $b(\alpha)$ by its first component and then compute the limit of the resulting vector $\tilde{b}(\alpha)$ as $\alpha \to 1^-$. We obtain

$$\lim_{\alpha \to 1^-} \tilde{b}(\alpha) = [\ 1,\ \frac{1}{2},\ \frac{3}{8}\ ]^T. \tag{6.11}$$

We now want to verify that the limit we just computed indeed represents the worst-case behavior for GMRES($A(1)$). We assume $b = [1\ \beta_2\ \beta_3]^T$ and obtain

$$K = \begin{pmatrix} 1 & 2 & 4 \\ \beta_2 & -2+2\beta_2 & -8+4\beta_2 \\ \beta_3 & -2\beta_2+2\beta_3 & 4-8\beta_2+4\beta_3 \end{pmatrix},$$

$$K^{-1} = \begin{pmatrix} 1-(1-\beta_2)\beta_2-\beta_3 & 1-\beta_2 & 1 \\ \frac{\beta_2}{2}-\beta_2+\beta_3 & -\frac{1}{2}+\beta_2 & -1 \\ \frac{1}{4}(\beta_2^2-\beta_3) & -\frac{\beta_2}{4} & \frac{1}{4} \end{pmatrix}.$$

We now compute $H(\beta_2, \beta_3)^2 = \|Ke_1\|^2\ \|K^{-T}e_1\|^2 = (1+\beta_2^2+\beta_3^2)(1+(1-\beta_2)^2+(1-(1-\beta_2)\beta_2-\beta_3)^2)$, find its gradient with respect to β_2 and β_3 and compute its zeros. The only real root of the gradient is precisely the point given in (6.11).

What we have demonstrated is that the framework that we have developed for analysis of GMRES applied to diagonalizable matrices A may be applied to defective matrices A as well. If we can express a given defective A_{def} as

$$A_{def} = \lim_{p \to p_0} A(p),$$

where $A(p)$ are diagonalizable, $p \in \mathcal{C}^k$ is a vector of parameters and p_0 is a certain limit value, and if we can derive convergence results for GMRES($A(p)$), we may be able to determine or estimate related quantities for GMRES(A_{def}) by taking limits.

6.3. General Worst-Case Behavior for $\alpha = 1$: Numerical Observations. In this section we present numerical data regarding the worst-case GMRES behavior for the defective convection-diffusion matrix $A_{def} = A(1)$ (see Equation (6.1)) of an arbitrary size. This data suggests that it is possible to use information about the worst-case behavior of the $n \times n$ problem to determine the worst-case behavior of the problem of dimension $n+1$. Although here we do not use the spectral decomposition framework, we can think of the results presented in this section as an extension of what was developed in Section 6.2.

Throughout this section, we use the following notation. Let A_n denote the convection-diffusion matrix $A(1)$ of size n, i.e.

$$A_n = \begin{pmatrix} 2 & & & \\ -2 & 2 & & \\ & \ddots & \ddots & \\ & & -2 & 2 \end{pmatrix} \in \mathcal{R}^{n \times n}.$$

Let $b_n = [1\ \beta_2\ \dots\ \beta_n]^T \in \mathcal{R}^n$ denote the right-hand side vector of size n and let $K_n = K(A_n, b_n)$. Assuming K_n is nonsingular, let $c_n = K_n^{-T}e_1$. Let $h_n(b_n)$ be the GMRES convergence measure at step $n-1$ for the $n-$dimensional problem and let $H_n(b_n)$ be its reciprocal. Then

$$H_n(b_n) = h_n(b_n)^{-1} = \|b_n\|\|c_n\|.$$

Finally, let b_n^{worst}, c_n^{worst}, h_n^{worst}, and H_n^{worst} represent quantities associated with the worst-case behavior of `GMRES(`A_n`)`. Thus in Section 6.2, we have established that

$$b_3^{worst} = [\ 1,\ \frac{1}{2},\ \frac{3}{8}\ ]^T, \quad h_3^{worst} = \frac{64}{89}.$$

We want to determine these quantities for an arbitrary n. To this end, we first look at the structure of K_n. Let us write explicitly its second and third columns,

$$A_n b_n = 2 \begin{bmatrix} 1 \\ -1+\beta_2 \\ -\beta_2+\beta_3 \\ \vdots \\ -\beta_{n-1}+\beta_n \end{bmatrix}, \quad A_n^2 b_n = 4 \begin{bmatrix} 1 \\ -2+\beta_2 \\ 1-2\beta_2+\beta_3 \\ \vdots \\ \beta_{n-2}-2\beta_{n-1}+\beta_n \end{bmatrix}.$$

With a simple induction argument, one can show that the top $n-1$ rows of the matrix K_n do not depend on β_n. This implies that if $b_{n+1} = [b_n^T\ \beta_{n+1}]^T \in \mathcal{R}^{(n+1)}$ then

$$K_{n+1} = \begin{bmatrix} K_n & a_n \\ \tilde{a}_n^T & \alpha_n \end{bmatrix} \in \mathcal{R}^{(n+1)\times(n+1)},$$

where $K_n \in \mathcal{R}^{n\times n}$, $a_n,\ \tilde{a}_n \in \mathcal{R}^n$ and $\alpha_n \in \mathcal{R}$ with K_n and a_n being independent of β_{n+1}. We also observed, although could not prove analytically, that the corresponding vector c_{n+1} is an increment of c_n, i.e. $c_{n+1} = [\gamma_{n+1}\ c_n^T]^T$ where $\gamma_{n+1} \in \mathcal{R}$ depends an all components of b_{n+1}. If this is true in general, and we believe it is, then

$$\begin{aligned} H_{n+1}^2(b_{n+1}) = \|b_{n+1}\|^2\|c_{n+1}\|^2 &= (\|b_n\|^2 + \beta_{n+1}^2)(\|c_n\|^2 + \gamma_{n+1}^2) \\ &= H_n^2(b_n) + \nu_n, \end{aligned} \tag{6.12}$$

where $\nu_n = \|b_n\|^2\gamma_{n+1}^2 + \|c_n\|^2\beta_{n+1}^2 + \gamma_{n+1}^2\beta_{n+1}^2$. We make the following observations. First, Equation (6.12) implies that $h_{n+1}^{worst} \leq h_n^{worst}$. Second, it also suggests that b_n^{worst} and b_{n+1}^{worst} are closely related and one may be computed from the other. We now present experimental results that indicate that this is the case.

For n varying between 4 and 50, we approximately computed b_n^{worst} and h_n^{worst} by evaluating $h_n(b_n)$ over a large mesh of points normally distributed over the unit sphere in $\mathcal{R}^n$. Once a coarse approximation has been computed, we refined it by focusing on the region where $h_n(b_n)$ was the largest. Upon inspection of the results, we conjecture the following. *First*, the vector b_n^{worst} satisfies $1 > \beta_2^{worst} > \ldots > \beta_n^{worst} > 0$, with β_n^{worst} usually being between about 80 and 90 percent of β_{n-1}^{worst}. *Second*, vectors b_n^{worst} and b_n^{worst} are related by

$$b_{n+1}^{worst} = \begin{bmatrix} b_n^{worst} \\ \beta_{n+1}^{worst} \end{bmatrix}. \tag{6.13}$$

Figure 6.1 illustrates our findings. The left subplot shows individual entries of b_{50}^{worst}. In addition, due to the relationship (6.13) it essentially plots the worst-case vectors for all $n < 50$ as well. The right subplot shows the value of h_n^{worst} for $4 \leq n \leq 50$. As predicted by (6.12), it monotonically decreases as n grows.

We now ask the following question: How does performance of `GMRES(`A_n`,`b_n^{worst}`)` compare to that of `GMRES(`A_n`,`b`)` for a random $b \in \mathcal{C}^n$ at intermediate steps of the algorithm? We try to answer this question partially by conducting the following experiment. For various values of n, we generate the defective matrix

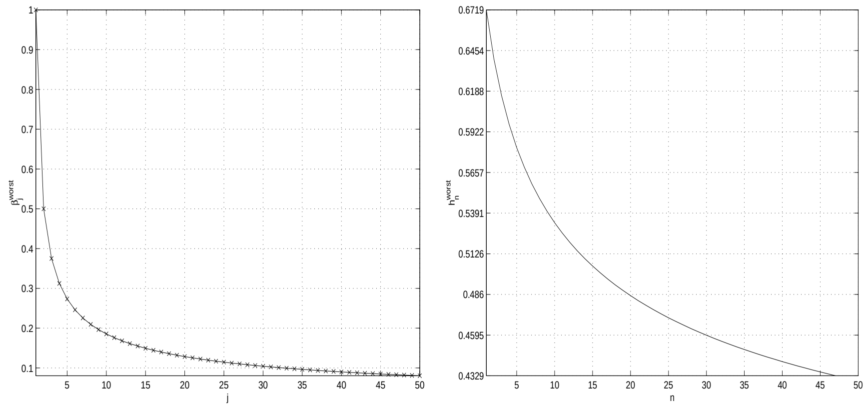

FIG. 6.1. *The vectors b_n^{worst} (left) and the measure h_n^{worst} (right) for the defective convection-diffusion matrix of size $n = 4, \dots, 50$*

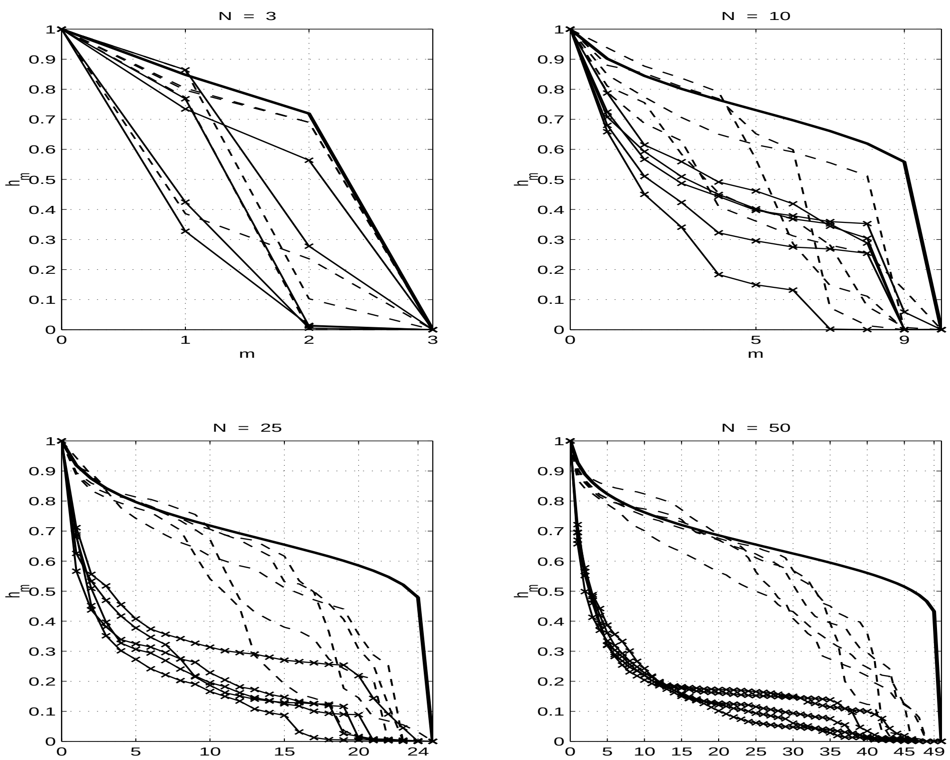

FIG. 6.2. *Performance of* `GMRES(`A_n, b`)` *for different choices of b, for $n = 3,\ 10,\ 25,\ 50$ and $m = 0, \dots, n-1$.*

A_n as well as right-hand side vectors of three types, namely (i) the worst-case vector b_n^{worst}; (ii) M random vectors b with positive entries; and (iii) M random vectors b with arbitrary entries. We then run `GMRES` with each matrix-vector pair and look at the sequence of residual ratios h_m, $m = 0, \dots, n-1$. Results of this test are shown in Figure 6.2.

We tested problems of size $n = 3,\ 10,\ 25$, and 50 and used $M = 5$. We selected such a small value of M to make the plots more readable. We note that the empirical findings we present below were observed for larger values of M as well. On each of the four subplots in Figure 6.2, the solid curve represents the convergence curve of GMRES(A_n,b_n^{worst}). The dashed curves and those labeled with an '$\times$' correspond to positive and mix-sign vectors b, respectively. We make the following observations. *First*, as predicted, at step $m = n-1$, $h(A_n, b_n^{worst})$ is larger than $h(A_n, b)$ for any other b. Moreover, GMRES(A_n,b_n^{worst}) exhibits relatively poor performance at intermediate steps as well, especially at later stages of the algorithms. Nevertheless, b_n^{worst} is not the worst-case vector for $m < n-1$. *Second*, overall, GMRES performs noticeably better when applied to mix-sign vectors b than when positive vectors are used. Also, the performance gap, almost nonexistent for small problems, seems to grow with n.

We obtained similar patterns when we applied the same test to a diagonalizable matrix $A(\alpha)$ for a fixed $\alpha < 1$. We therefore conjecture that at later stages of the algorithm, $h_m(A_n, b_n^{worst})$ may be close the worst-case behavior, while at its early stages, the worst-case b is some other vector with positive components.

7. Can Worst-Case Analysis Be Misleading?. Throughout this document, we have been focusing on the worst-case analysis of GMRES convergence. If GMRES(A) and GMRES(A') achieve the same worst-case performance at step m for the matrices $A = V\Lambda V^{-1}$ and $A' = V'\Lambda'(V')^{-1}$ then clearly measures $h_m(V, \lambda, y)$ and $h_m(V', \lambda', y)$ have the same range of values. However, as we will see in this section, this fact does not necessarily imply that the method has identical overall behavior when applied to the two matrices.

In this section we focus on step $m = n-1$ and present some experimental data that shows that an alternative measure of overall performance, such as the mean of $h(V, \lambda, y)$ over all right-hand side vectors $b = Vy$ of unit length, may be a better indicator of average performance of GMRES.

7.1. Approximate Computation of the Mean. In this section we focus on real matrices A and vectors b and again assume that $\|b\| = 1$. This yields the convergence measure $h(V, \lambda, y) = \|V^{-H}\overline{Y}^{-1}u\|^{-1}$. Let us define the set $R_n^+ = \{b = [\beta_1, \ \ldots, \ \beta_n]^T \in R_n \mid \|b\| = 1, \ \beta_n \geq 0 \ \}$ that constitutes the upper half of the real unit hyper-sphere Without loss of generality we may assume that $h(V, \lambda, V^{-1}b)$ is defined over R_n^+. Thus for given V and λ, $h(V, \lambda, V^{-1}b) : R_n^+ \to [0, 1]$. Overall performance of GMRES(A) can be measured by its mean,

$$\bar{h} = \bar{h}(V, \lambda) = \frac{1}{A(R_n^+)} \int_{R_n^+} h(V, \lambda, V^{-1}b), \tag{7.1}$$

where $A(R_n^+)$ is the total surface area of the half-sphere R_n^+. In other words, $\bar{h}$ is just a scaled surface integral of the measure $h(V, \lambda, V^{-1}b)$.

The formula (7.1) yields a very complicated expression which we have not been able to evaluate exactly. Therefore in our experiments we seek to approximate it. The most straightforward way to do it is to evaluate $h(V, \lambda, b)$ on a discrete mesh over R_n^+ and then to compute the average of all the values at the mesh nodes. Clearly, in order for the approximation to be good, the mesh has to be both fine and uniform. As pointed out in [2], given an integer M, it is possible to obtain a uniform mesh of the sphere by generating M $n-$vectors of normally distributed random numbers with zero mean and unit variance. These vectors can then be scaled to put them on the top half of the unit sphere.

Unfortunately, even in the cases of small n, the mesh has to be rather fine in order to get an accurate picture. For instance, when $n = 3$, the values of M between 10^5 and 10^6 are usually used, and this value grows with n. This often makes computational experiments even with small-dimensional problems expensive in terms of both time and memory.

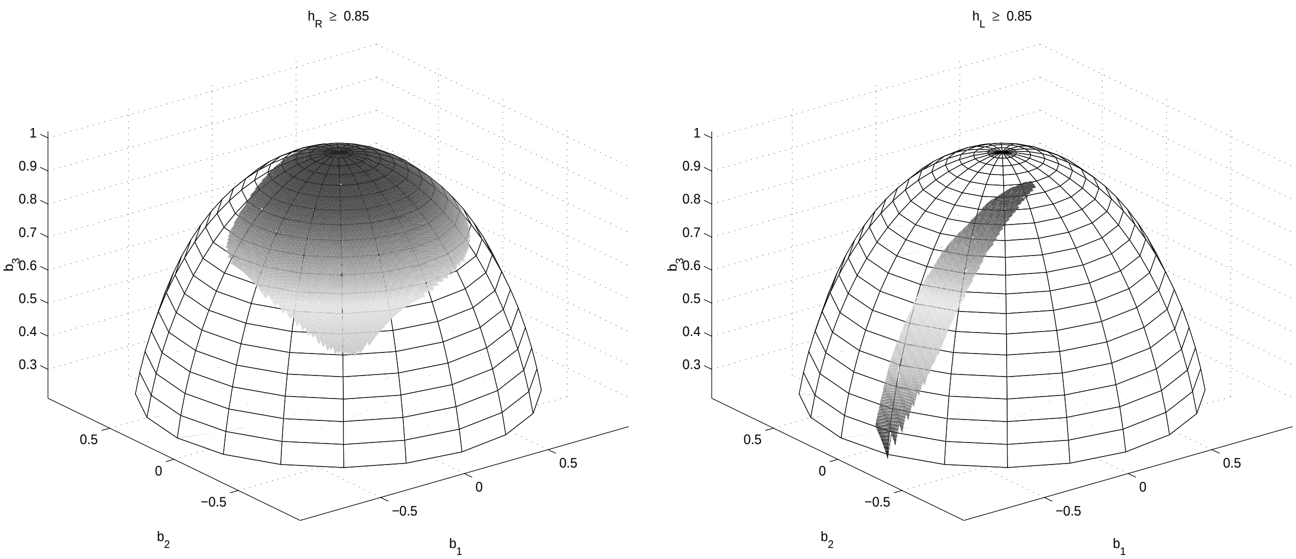

FIG. 7.1. *Different behavior of h_R (left) and h_L (right) in neighborhood of b_{stagn}*

7.2. An Example of Different Behavior of A and A^T. Let us consider the following matrix

$$A = \begin{pmatrix} 3.64347104554523 & -1.30562625697964 & 2.12276233724947 \\ 3.81895186997748 & -0.33626408416579 & 8.43952325416869 \\ 0.12754105943518 & 0.13002776444227 & 2.98820549610000 \end{pmatrix}$$

together with its transpose. This matrix has real spectrum. We compute V_R and $V_L = V_R^{-T}$, which are the eigenvectors of A and A^T, respectively. We consider vectors $b \in R_n^+$ and denote by h_R and h_L the measures $h(V_R, \lambda, V_R^{-1}b)$ and $h(V_L, \lambda, V_L^{-1}b)$, respectively. Thus h_R and h_L are `GMRES` convergence measures at step $m = n-1$ for A and A^T, respectively.

By Theorem 5.4, h_R and h_L attain the same maximums over the unit sphere. In fact, `GMRES(`A`)` and `GMRES(`A^T`)` stagnate [23] at the following two points

$$b_{stagn_1} = \begin{bmatrix} -0.22385545043433 \\ -0.30471918583417 \\ 0.92576182418211 \end{bmatrix}, \quad b_{stagn_2} = \begin{bmatrix} -0.46000942948917 \\ -0.32420970874985 \\ 0.82660715551465 \end{bmatrix},$$

We now generate a mesh of $K = 10^6$ normally distributed points and approximately compute $\bar{h}_R$ and $\bar{h}_L$. The values we obtain are quite different, namely, $\bar{h}_R \approx 0.4512$ and $\bar{h}_L \approx 0.1835$. Closer examination reveals that `GMRES(`A`)` and `GMRES(`A^T`)` behave differently in the neighborhood of the stagnation points.

Let us examine Figure 7.1. The left and right subplots correspond to h_R and h_L, respectively. The shaded areas correspond to the regions where h_R and h_L are larger than 0.85. As expected, the stagnating points $b_{stagn_{1,2}}$ are inside both of these regions. However, the region corresponding to h_R is significantly larger which explains why its mean is larger as well. In other words, h_R in general changes much more slowly in the neighborhood of $b_{stagn_{1,2}}$ than does h_L.

8. Open Questions. As often happens, the development of a new approach to `GMRES` convergence analysis raised more questions than it answered. In addition to various conjectures arising from empirical evidence presented in Sections 6 and 7, there are questions that can be thought of as generalizations of results presented in this paper. Here we mention some of them. In Section 3 we derived bounds for the convergence measure at step $m = n - 1$. Is it possible to obtain an accurate bound for an arbitrary steps using our framework? In Section 6, we studied the matrices arising from the one-dimensional convection-diffusion equations and observed that the form of the worst-case right-hand side for step $m = n - 1$ does not change with α and is computed directly from the vector $u = G(\lambda)$. What about intermediate steps? Also, how does this result generalize to matrices for the two-dimensional convection-diffusion equation like the ones discussed in [4]? Finally, in Section 7 we demonstrated that worst-case-based analysis of `GMRES` performance may be misleading and proposed $mean(h(V, \lambda, b))$ as an alternative overall measure. However, due to the fact that the expression for the mean is extremely complicated we were not able to develop any analytical results. So the question remains whether it is possible to come up with a different means of measuring overall performance of the algorithm that would be simpler than the mean and at the same time would capture the behavior of $h(V, \lambda, b)$ over regions better then does its maximum.

9. Acknowledgments. The author would like to thank Dianne O'Leary and Howard Elman for helpful comments, as well as for guidance and support in author's doctorate research that led to results presented here.

REFERENCES

[1] A. J. Clayton, *Further results on polynomials having least maximum modulus over an ellipse in the complex plane.* UKAEA Memorandum, AEEW, 1963. M 348.

[2] L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, Berlin, New York, 1986.

[3] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.

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

[5] B. Fischer and R. Freund, *Chebyshev polynomials are not always optimal*, J. Approximation Theory, 65 (1991), pp. 261–272.

[6] A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM J. Sci. Stat. Comput., 15 (1994), pp. 348–358.

[7] A. Greenbaum, V. Ptak, and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.

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

[9] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Standards, 49 (1952), pp. 409–436.

[10] N. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[11] I. C. F. Ipsen, *Expressions and bounds for the GMRES residual*, BIT, 40 (2000), pp. 524–533.

[12] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, New York, second ed., 1985.

[13] T. A. Manteuffel, *An Iterative Method for Solving Nonsymmetric Linear Systems with Dynamical Estimation of Parameters*, PhD thesis, University of Illinois at Urbana-Champaign, 1975. Technical Report UIUCDCS-75-758.

[14] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw-Hill, New York, 1996.

[15] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[16] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations – Convection-Diffusion and Flow Problems*, Springer-Verlag, Berlin, New York, 1996.

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

[18] G. W. Stewart, *Collinearity and least squares regression*, Statistical Science, 2 (1987), pp. 68–100. With discussion.

[19] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[20] L. N. Trefethen, *Approximation theory and numerical linear algebra*, in Algorithms for Approximation II, J. Mason and L. Cox, eds., Chapman and Hall, London, 1990.

[21] S. Wolfram, *The Mathematica Book*, Wolfram Media/Cambridge University Press, fourth ed., 1999.

[22] I. Zavorin, *Analysis of GMRES Convergence by Spectral Factorization of the Krylov Matrix*, PhD thesis, University of Maryland, College Park, August 2001.

[23] I. Zavorin, D. P. O'Leary, and H. Elman, *Stagnation of GMRES*, Technical Report CS-TR-4296, Computer Science Department, University of Maryland, College Park, 2001.