SIAM J. MATRIX ANAL. APPL.
Vol. 22, No. 3, pp. 726–751
© 2000 Society for Industrial and Applied Mathematics

DIFFERENCES IN THE EFFECTS OF ROUNDING ERRORS IN KRYLOV SOLVERS FOR SYMMETRIC INDEFINITE LINEAR SYSTEMS*

GERARD L. G. SLEIJPEN†, HENK A. VAN DER VORST†, AND JAN MODERSITZKI‡

Abstract. The three-term Lanczos process for a symmetric matrix leads to bases for Krylov subspaces of increasing dimension. The Lanczos basis, together with the recurrence coefficients, can be used for the solution of symmetric indefinite linear systems, by solving a reduced system in one way or another. This leads to well-known methods: MINRES (minimal residual), GMRES (generalized minimal residual), and SYMMLQ (symmetric LQ). We will discuss in what way and to what extent these approaches differ in their sensitivity to rounding errors.

In our analysis we will assume that the Lanczos basis is generated in exactly the same way for the different methods, and we will not consider the errors in the Lanczos process itself. We will show that the method of solution may lead, under certain circumstances, to large additional errors, which are not corrected by continuing the iteration process.

Our findings are supported and illustrated by numerical examples.

Key words. linear systems, iterative methods, MINRES, GMRES, SYMMLQ, stability

AMS subject classifications. 65F10, 65N12

PII. S0895479897323087

1. Introduction.

We consider iterative methods for the construction of approximations to the solution of a linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is supposed to be a real symmetric n by n matrix. Without loss of generality, we assume $\mathbf{x}_0 = 0$. Let $\mathbf{r}_k = \mathbf{b} - \mathbf{Ax}_k$ (in particular, $\mathbf{r}_0 = \mathbf{b}$) and

$$\mathcal{K}_k(\mathbf{A}; \mathbf{b}) \equiv \text{Span}\{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\},$$

the k-dimensional Krylov subspace. The methods to be analyzed build the iterates $\mathbf{x}_k$ such that

1. $\mathbf{x}_k \in \mathcal{K}_k(\mathbf{A}; \mathbf{b})$ and $\|\mathbf{b} - \mathbf{Ax}_k\|_2 = \min$ (GMRES, MINRES),
2. $\mathbf{x}_k \in \mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{b})$ and $\|\mathbf{A}^{-1}\mathbf{b} - \mathbf{x}_k\|_2 = \min$ (SYMMLQ).

With the standard three-term Lanczos process, we generate an orthonormal basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$ for $\mathcal{K}_k(\mathbf{A}; \mathbf{b})$, with $\mathbf{v}_1 \equiv \mathbf{b}/\|\mathbf{b}\|_2$. The three-term Lanczos process can be recast in matrix formulation as

$$\mathbf{AV}_k = \mathbf{V}_{k+1}\underline{T}_k, \tag{1}$$

in which $\mathbf{V}_j$ is defined as the n by j matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_j$, and $\underline{T}_k$ is a $k+1$ by k tridiagonal matrix.

Paige [9] has shown that in finite precision arithmetic, the Lanczos process can be implemented so that the *computed* $\mathbf{V}_{k+1}$ and $\underline{T}_k$ satisfy

$$\mathbf{AV}_k = \mathbf{V}_{k+1}\underline{T}_k + \mathbf{F}_k, \tag{2}$$

*Received by the editors June 17, 1997; accepted for publication (in revised form) by Z. Strakoš March 28, 2000; published electronically October 25, 2000.

http://www.siam.org/journals/simax/22-3/32308.html

†Mathematical Institute, Utrecht University, P.O. Box 80.010, 3508 TA Utrecht, The Netherlands (sleijpen@math.uu.nl, vorst@math.uu.nl).

‡Institute of Mathematics, Medical University of Lübeck, Wallstraße 40, 23560 Lübeck, Germany (modersitzki@math.mu-luebeck.de).

with, under mild conditions for k,

$$\|\mathbf{F}_k\|_2 \leq 2\sqrt{k}\,(7\|\mathbf{A}\|_2 + m_1\|\,|\mathbf{A}|\,\|_2)\,\mathbf{u}$$

($\mathbf{u}$ is the machine precision, and m_1 denotes the maximum number of nonzeros in any row of $\mathbf{A}$). Since $\|\,|\mathbf{A}|\,\|_2 \leq \sqrt{m_1}\,\|\mathbf{A}\|_2$ (see Lemma A.1), we obtain the convenient expression

$$\|\mathbf{F}_k\|_2 \leq 2\sqrt{k}\,(7 + m_1\sqrt{m_1}\,)\,\|\mathbf{A}\|_2\,\mathbf{u}. \tag{3}$$

Popular Krylov subspace methods for symmetric linear systems can be derived with formula (1) as a starting point: MINRES, GMRES (adapted to symmetric matrices; see below), and SYMMLQ. The matrix $\underline{T}_k$ can be interpreted as the restriction of $\mathbf{A}$ with respect to the Krylov subspace, and the main idea behind these Krylov solution methods is that the given system $\mathbf{Ax} = \mathbf{b}$ is replaced by a smaller system with $\underline{T}_k$ over the Krylov subspace. This reduced system is solved—implicitly or explicitly—in a convenient way and the solution is transformed with $\mathbf{V}_k$ to a solution in the original n-dimensional space. The main computational differences between the methods are due to a different way of solution of the reduced system and to differences in the back transformation to an approximate solution of the original system. We will describe these differences in relevant detail in coming sections.

Of course, these methods have been derived assuming exact arithmetic; for instance, the generating formulas are all based on an exact orthogonal basis for the Krylov subspace. In numerical reality, however, we have to compute this basis, as well as all other quantities in the methods, and then it is of importance to know how the generating formulas behave in finite precision arithmetic. The errors in the underlying Lanczos process have been analyzed by Paige [9, 10]. It has been proven by Greenbaum and Strakoš [7] that rounding errors in the Lanczos process may have a delaying effect on the convergence of iterative solvers but do not prevent eventual convergence in general. Usually, a rigorous error analysis is on a worst case scenario, and as a consequence, the error bounds cannot very well be used to explain differences between these methods, as observed in practical situations.

In this paper, we propose a different way of analyzing these methods, different in the way that we do not attempt to derive sharper upper bounds, but that we try to derive upper bounds for relevant differences between these processes in finite precision arithmetic. This will not help us to understand why any of these methods converges in finite precision, but it will give us some insight in answering practical questions such as the following.

• When and why is MINRES less accurate than SYMMLQ? This question was already posed in the original publication [11], but the answer in [11, p. 625] is largely speculative.

• Is MINRES suspect for ill-conditioned systems, because of the minimal residual approach (see [11, p. 619])? Hints are given for the explanation of the observation that MINRES may be more inaccurate than SYMMLQ [11, p. 625]. We will further substantiate this. In [2, p. 43] an explicit relation is suggested between MINRES and working with $\mathbf{A}^2$, and it is argued that its sensitivity to rounding errors of the solution depends on $\kappa_2(\mathbf{A})^2$. (It is even stated: 'the squared condition number of $\mathbf{A}^2$', implying $\kappa_2(\mathbf{A}^2)^2 = \kappa_2(\mathbf{A})^4$, which seems to be an unlucky formulation.)

• Why and when does SYMMLQ converge slower than, for instance, MINRES or GMRES?

```
Choose x_0
x = x_0,  r = b − Ax,  ρ = ||r||,  v = r/ρ
β = 0,  β̃ = 0,  c = −1,  s = 0
v_old = 0,  w = 0,  w̃̃ = v
while |ρ| > tol do
   ṽ ← A v − β v_old
   α ← v* ṽ,  ṽ ← ṽ − α v
   β ← ||ṽ||,  v_old ← v,  v ← ṽ/β
   ℓ_1 ← s α − c β̃,  ℓ_2 ← s β
   α̃ ← −s β̃ − c α,  β̃ ← c β
   ℓ_0 ← sqrt(α̃^2 + β^2),  c ← α̃/ℓ_0,  s ← β/ℓ_0
   w̃ ← w̃̃ − ℓ_1 w,  w̃̃ ← v − ℓ_2 w
   w ← w̃/ℓ_0
   x ← x + (ρ c) w,  ρ ← s ρ
end while
```

FIG. 1. *The* MINRES *algorithm.*

• Why does MINRES sometimes lead to rather large residuals, whereas the error in the approximation is significantly smaller? See, for instance, observations on this made in [11, p. 626]. Most important, understanding the differences between these methods will help us in making a choice.

We will now briefly characterize the different methods in our investigation.

1. **MINRES** (see [11]): Determine $\mathbf{x}_k = \mathbf{V}_k y_k$, $y_k \in \mathbb{R}^k$, such that $\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2$ is minimal. This minimization leads to a small system with $\underline{T}_k$, and the tridiagonal structure of $\underline{T}_k$ is exploited to get a short recurrence relation for $\mathbf{x}_k$. The advantage of this is that only three vectors from the Krylov subspace have to be saved (in fact, MINRES works with transformed basis vectors; this will be explained in section 2.3). For the implementation of MINRES that we have used, see Figure 1.
2. **GMRES** (see [13]): This method also minimizes, for $y_k \in \mathbb{R}^k$, the residual $\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\|_2$. GMRES was designed for unsymmetric matrices for which the orthogonalization of the Krylov basis is done with Arnoldi's method. This leads to a small upper Hessenberg system that has to be solved. However, when $\mathbf{A}$ is symmetric, then, in exact arithmetic, the Arnoldi method is equivalent to the Lanczos method (see also [6, p. 41]). Although GMRES is commonly presented with an Arnoldi basis, there are various implementations of it that differ in finite precision, for instance, with modified Gram–Schmidt, classical Gram–Schmidt, Householder, and other variants. We view Lanczos as one way to obtain an orthogonal basis, and therefore, we stick to the name GMRES. However, in order to stress the fact that our version of GMRES relies on Lanczos, we will use the notation GMRES*.

   Due to the way of solution in GMRES* (and in GMRES), all the basis

Choose $\mathbf{x}_0$

$\mathbf{x} = \mathbf{x}_0,\ \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x},\ \rho = \|\mathbf{r}\|,\ \mathbf{v} = \mathbf{r}/\rho$

$\beta = 0,\ \widetilde{\beta} = 0,\ c = -1,\ s = 0$

$\mathbf{v}_{\text{old}} = \mathbf{0},\ \mathbf{V} = [\,],\ z = [\,],\ k = 0$

while $\rho > tol$ **do**

  $\mathbf{V} \leftarrow [\mathbf{V}, \mathbf{v}],\ k \leftarrow k + 1$

  $\widetilde{\mathbf{v}} \leftarrow \mathbf{A}\mathbf{v} - \beta\,\mathbf{v}_{\text{old}}$

  $\alpha \leftarrow \mathbf{v}^*\widetilde{\mathbf{v}},\ \widetilde{\mathbf{v}} \leftarrow \widetilde{\mathbf{v}} - \alpha\,\mathbf{v}$

  $\beta \leftarrow \|\widetilde{\mathbf{v}}\|,\ \mathbf{v}_{\text{old}} \leftarrow \mathbf{v},\ \mathbf{v} \leftarrow \widetilde{\mathbf{v}}/\beta$

  $\ell_1 \leftarrow s\,\alpha - c\,\widetilde{\beta},\ \ell_2 \leftarrow s\,\beta$

  $\widetilde{\alpha} \leftarrow -s\,\widetilde{\beta} - c\,\alpha,\ \widetilde{\beta} \leftarrow c\,\beta$

  $\ell_0 \leftarrow \sqrt{\widetilde{\alpha}^2 + \beta^2},\ c \leftarrow \widetilde{\alpha}/\ell_0,\ s \leftarrow \beta/\ell_0$

  if $k = 1$

    $\vec{\ell} = [\,],\ R = [\,\ell_0\,]$

  else

    $R \leftarrow \begin{bmatrix} R \\ \vec{0} \end{bmatrix},\ \vec{\ell} \leftarrow [\,\vec{\ell}, \ell_1, \ell_0\,]$

    $R \leftarrow [\,R, \vec{\ell}^{\,T}\,],\ \vec{\ell} \leftarrow [\,\vec{0}, \ell_2\,]$

  end if

  $z \leftarrow [\,z^{\,T}, c\rho\,]^{\,T},\ \rho \leftarrow s\,\rho$

end while

$\mathbf{x} = \mathbf{x} + \mathbf{V}(R^{-1}z)$

FIG. 2. *The* GMRES* *algorithm. The vector* $\vec{0}$ *for the expansion of the upper triangular matrix* R *is a row vector of zeros of appropriate size (different size at different occurrences).*

Choose $\mathbf{x}_0$

$\mathbf{x} = \mathbf{x}_0,\ \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x},\ \rho = \|\mathbf{r}\|,\ \mathbf{v} = \mathbf{r}/\rho$

$\beta = 0,\ \widetilde{\beta} = 0,\ c = -1,\ s = 0,\ \kappa = \rho$

$\mathbf{v}_{\text{old}} = \mathbf{0},\ \mathbf{w} = \mathbf{v},\ g = 0,\ \widetilde{\widetilde{g}} = \rho$

while $\kappa > tol$ **do**

  $\widetilde{\mathbf{v}} \leftarrow \mathbf{A}\mathbf{v} - \beta\,\mathbf{v}_{\text{old}}$

  $\alpha \leftarrow \mathbf{v}^*\widetilde{\mathbf{v}},\ \widetilde{\mathbf{v}} \leftarrow \widetilde{\mathbf{v}} - \alpha\,\mathbf{v}$

  $\beta \leftarrow \|\widetilde{\mathbf{v}}\|,\ \mathbf{v}_{\text{old}} \leftarrow \mathbf{v},\ \mathbf{v} \leftarrow \widetilde{\mathbf{v}}/\beta$

  $\ell_1 \leftarrow s\,\alpha - c\,\widetilde{\beta},\ \ell_2 \leftarrow s\,\beta$

  $\widetilde{\alpha} \leftarrow -s\,\widetilde{\beta} - c\,\alpha,\ \widetilde{\beta} \leftarrow c\,\beta$

  $\ell_0 \leftarrow \sqrt{\widetilde{\alpha}^2 + \beta^2},\ c \leftarrow \widetilde{\alpha}/\ell_0,\ s \leftarrow \beta/\ell_0$

  $\widetilde{g} \leftarrow \widetilde{\widetilde{g}} - \ell_1\,g,\ \widetilde{\widetilde{g}} \leftarrow -\ell_2\,g,\ g \leftarrow \widetilde{g}/\ell_0$

  $\mathbf{x} \leftarrow \mathbf{x} + (g\,c)\,\mathbf{w} + (g\,s)\,\mathbf{v}$

  $\mathbf{w} \leftarrow s\,\mathbf{w} - c\,\mathbf{v},\ \kappa \leftarrow \sqrt{\widetilde{g}^2 + \widetilde{\widetilde{g}}^2}$

end while

FIG. 3. *The* SYMMLQ *algorithm.*

vectors $\mathbf{v}_j$ have to be stored. For our implementation of GMRES*, see Figure 2.

3. **SYMMLQ** (see [11]): Determine $\mathbf{x}_k = \mathbf{A}\mathbf{V}_k y_k$, $y_k \in \mathbb{R}^k$, such that the error $\mathbf{x} - \mathbf{x}_k$ has minimal Euclidean length. It may come as a surprise that $\|\mathbf{x} - \mathbf{x}_k\|_2$ can be minimized without knowing $\mathbf{x}$, but this can be accomplished by restricting the choice of $\mathbf{x}_k$ to $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{b})$. Conjugate gradient approximations can, if they exist, be computed with little effort from the SYMMLQ information. In the SYMMLQ implementation suggested in [11] this is used to terminate iterations either at a SYMMLQ iterate or a conjugate gradient iterate, depending on which one is best. For the implementation of SYMMLQ that we have used, see Figure 3.

Note that these methods can be carried out with exactly the same basis vectors $\mathbf{v}_j$ and tridiagonal matrices $\underline{T}_j$.

Notations. Quantities associated with n dimensional spaces will be represented in boldface, like $\mathbf{A}$ and $\mathbf{v}_j$. Vectors and matrices on low dimensional subspaces are

denoted in normal mode: T, y. Constants will be denoted by lowercase Greek symbols, with the exception that we will use $\mathbf{u}$ to denote the relative machine precision. The absolute value of a matrix refers to elementwise absolute values, that is, $|A| = (|a_{ij}|)$ for $A = (a_{ij})$.

Most of our bounds on perturbations in the solutions at the kth iteration step will be expressed as bounds for corresponding perturbations to the residual in the kth step, relative to the norm of an initial residual. Since all these iteration methods construct their search spaces from residual vector information (that is, they all start with $\mathbf{r}_0 = \mathbf{b}$), and since we make at least errors in the order of $\mathbf{u}\,\|\mathbf{b}\|_2$ in the computation of the residuals, we may not expect perturbations of order less than $\mathbf{u}\,\kappa_2(\mathbf{A})\|\mathbf{b}\|_2$ in the iteratively computed solutions. So, our bounds can only be expected to show up in the computed residuals, if the errors are larger than the error induced by the computation of the residuals itself.

2. Differences in round-off errors for MINRES and GMRES*.

2.1. The basic formulas for GMRES* and MINRES in exact arithmetic.

We will first describe the generic formulas for the iterative methods MINRES and GMRES*, and we will assume *exact arithmetic* in the derivation of these formulas.

The aim is to minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2$ over the Krylov subspace, and since

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 &= \|\mathbf{b} - \mathbf{A}\mathbf{V}_k\, y_k\|_2 \\ &= \|\mathbf{b} - \mathbf{V}_{k+1}\,\underline{T}_k\, y_k\|_2 \\ &= \|\underline{T}_k\, y_k - \|\mathbf{b}\|_2\, e_1\|_2, \end{aligned} \tag{4}$$

we see that a minimizer y_k must be the linear least squares solution of the $k+1$ by k overdetermined system

$$\underline{T}_k\, y_k = \|\mathbf{b}\|_2 e_1.$$

This system is solved with Givens rotations, which leads to an upper triangular reduction of $\underline{T}_k$,

$$\underline{T}_k = \underline{Q}_k R_k, \tag{5}$$

in which R_k is k by k upper triangular with bandwidth 3 and $\underline{Q}_k$ is a $k+1$ by k matrix with orthonormal columns. Using (5), y_k can be solved from

$$R_k\, y_k = z_k \equiv \|\mathbf{b}\|_2\, \underline{Q}_k^T e_1,$$

and since $\mathbf{x}_k = \mathbf{V}_k y_k$, we obtain

$$\mathbf{x}_k = \mathbf{V}_k R_k^{-1} \underline{Q}_k^T \|\mathbf{b}\|_2 e_1 = \mathbf{V}_k R_k^{-1} z_k. \tag{6}$$

The GMRES method, proposed for unsymmetric $\mathbf{A}$ in [13], can be characterized by the specific order of computation in the above derivation, indicated by adding parentheses:

$$\mathbf{x}_k = \mathbf{V}_k (R_k^{-1} \underline{Q}_k^T \|\mathbf{b}\|_2 e_1) = \mathbf{V}_k (R_k^{-1} z_k). \tag{7}$$

When $\mathbf{A}$ is symmetric, then Arnoldi's method is equivalent to Lanczos's method, so that (7) describes GMRES for symmetric $\mathbf{A}$ (further referred to as GMRES*). The

well-known disadvantage of this approach is that we have to store all columns of $\mathbf{V}_k$ for the computation of $\mathbf{x}_k$.

MINRES follows essentially the same approach as GMRES for the minimization of the residual, but it exploits the banded structure of R_k in order to get short recurrences for $\mathbf{x}_k$ and in order to save on memory storage.

Indeed, the computations in the generating formula (6) can be grouped as

$$\mathbf{x}_k = \left(\mathbf{V}_k R_k^{-1}\right) z_k \equiv \mathbf{W}_k z_k. \tag{8}$$

For the computation of $\mathbf{W}_k = \mathbf{V}_k R_k^{-1}$, it is easy to see that the last column of $\mathbf{W}_k$ is obtained from the last two columns of $\mathbf{W}_{k-1}$ and $\mathbf{v}_k$. This makes it possible to update $\mathbf{x}_{k-1} = \mathbf{W}_{k-1} z_{k-1}$ to $\mathbf{x}_k$ with a short recurrence, since z_k follows from the kth Givens rotation applied to the vector $(z_{k-1}^T, 0)^T$. This interpretation leads to MINRES.

We see that MINRES and GMRES* both use $\mathbf{V}_k$, R_k, $\underline{T}_k$, $\underline{Q}_k$, and z_k for the computation of $\mathbf{x}_k$. Of course, we are not forced to compute these quantities in exactly the same way for the two methods, but there is no reason to compute them differently. Therefore, we will compare implementations of GMRES* and MINRES that are based on *exactly the same* quantities in floating point finite arithmetic.

From now on we will study in what way MINRES and GMRES* differ in finite precision arithmetic, given exactly the same set $\mathbf{V}_k$, R_k, $\underline{T}_k$, $\underline{Q}_k$, and z_k (all computed in finite precision, too) for the two different methods. Hence, the differences in finite precision between GMRES* and MINRES are only caused by a different order of computation of the formula $\mathbf{x}_k = \mathbf{V}_k R_k^{-1} z_k$, namely,

$$\text{for GMRES*:} \quad \mathbf{x}_k = \mathbf{V}_k \left(R_k^{-1} z_k\right), \tag{9}$$

$$\text{for MINRES:} \quad \mathbf{x}_k = \left(\mathbf{V}_k R_k^{-1}\right) z_k. \tag{10}$$

In finite precision, the relation (5) will not be satisfied exactly. Instead, we have that [8, Theorem 18.4]

$$\underline{T}_k = \underline{Q}_k R_k + \underline{G}_k, \quad \text{where} \quad \|\underline{G}_k\|_F \le c\, k^2\, \mathbf{u}\, \|\underline{T}_k\|_F + \mathcal{O}(\mathbf{u}^2), \tag{11}$$

with c a modest constant. The matrix $\underline{Q}_k$ is orthogonal; it is the product of the exact Givens rotations involved in the elimination of subdiagonal elements in the actually computed reductions of $\underline{T}_k$.

2.2. Error analysis for GMRES*. In order to understand the difference between GMRES* and MINRES, we will study in this section the computational errors in $\mathbf{V}_k \left(R_k^{-1} z_k\right)$, with respect to the exactly evaluated $\mathbf{V}_k R_k^{-1} z_k$ (given the computed $\mathbf{V}_k$, R_k, and z_k). We will indicate actual computation in floating point finite precision arithmetic by *fl*, and the result will be denoted by a $\widehat{\ }$. Then, according to [4, p. 89], in floating point arithmetic the computed solution $\widehat{y}_k = fl(R_k^{-1} z_k)$ satisfies

$$(R_k + \Delta_R)\widehat{y}_k = z_k, \quad \text{with} \quad |\Delta_R| \le 3\,\mathbf{u}\,|R_k| + \mathcal{O}(\mathbf{u}^2). \tag{12}$$

This implies that $\widehat{y}_k = (I + R_k^{-1}\Delta_R)^{-1} R_k^{-1} z_k$, so that, apart from second order terms in $\mathbf{u}$, the error Δ_1 in the computation of y_k is

$$\Delta_1 \equiv \widehat{y}_k - y_k = -R_k^{-1}\Delta_R R_k^{-1} z_k.$$

Here $y_k = R_k^{-1} z_k$: y_k is the exact value based on the computed R_k and z_k. Then we also make errors in the computation of $\mathbf{x}_k$, that is, we compute $\widehat{\mathbf{x}}_k = fl(\mathbf{V}_k \widehat{y}_k)$. With the error bounds for the matrix vector product [8, p. 76], we obtain

$$\widehat{\mathbf{x}}_k = \mathbf{V}_k \widehat{y}_k + \Delta_2, \quad \text{with} \quad |\Delta_2| \le k\,\mathbf{u}\,|\mathbf{V}_k|\,|y_k| + \mathcal{O}(\mathbf{u}^2). \tag{13}$$

Hence, the error $\Delta\mathbf{x}_k = \widehat{\mathbf{x}}_k - \mathbf{x}_k$ (where $\mathbf{x}_k = \mathbf{V}_k R_k^{-1} z_k$), which can be attributed to the evaluation of the generating formula (9) for GMRES*, has two components:

$$\Delta\mathbf{x}_k = \mathbf{V}_k \Delta_1 + \Delta_2. \tag{14}$$

This error leads to a contribution $\Delta\mathbf{r}_k$ to the residual, that is, $\Delta\mathbf{r}_k$ is that part of $\mathbf{r}_k$ that can be attributed to errors in the evaluation of (9) (ignoring $\mathcal{O}(\mathbf{u}^2)$ terms):

$$\begin{aligned} \Delta\mathbf{r}_k \equiv \widehat{\mathbf{r}}_k - \mathbf{r}_k &= -\mathbf{A}\Delta\mathbf{x}_k \\ &= -\mathbf{A}\mathbf{V}_k\Delta_1 - \mathbf{A}\Delta_2 r \\ &= \mathbf{A}\mathbf{V}_k R_k^{-1}\Delta_R R_k^{-1} z_k - \mathbf{A}\Delta_2 \\ &= \mathbf{V}_{k+1}\underline{T}_k R_k^{-1}\Delta_R R_k^{-1} z_k - \mathbf{A}\Delta_2 \\ &= \mathbf{V}_{k+1}\underline{Q}_k \Delta_R R_k^{-1} z_k - \mathbf{A}\Delta_2. \end{aligned} \tag{15}$$

Note that in finite precision we have that $\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\underline{T}_k + \mathbf{F}_k$, and that, because of (3), the term $\mathbf{F}_k$ leads to an additional contribution of $\mathcal{O}(\mathbf{u}^2)$ in $\Delta\mathbf{r}_k$. This is also the case in forthcoming situations where we replace $\mathbf{A}\mathbf{V}_k$ by $\mathbf{V}_{k+1}\underline{T}_k$ in the derivation of upper bounds for error contributions. In a similar way, the error term $\underline{G}_k R_k^{-1}$ in the formula for $\underline{T}_k R_k^{-1}$ (see (11)) leads to a $\mathcal{O}(\mathbf{u}^2)$ term.

Using the bound in (12) and the bound for Δ_2, we get (skipping higher order terms in $\mathbf{u}$)

$$\begin{aligned} \|\Delta\mathbf{r}_k\|_2 &\le \|\mathbf{V}_{k+1}\underline{Q}_k\|_2\, 3\,\mathbf{u}\, \| \,|R_k|\,\|_2\, \|R_k^{-1} z_k\|_2 + k\,\mathbf{u}\,\|\mathbf{A}\|_2\, \| \,|\mathbf{V}_k|\,\|_2 \|y_k\|_2 \\ &\le 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_{k+1}\|_2\,\|R_k\|_2\,\|R_k^{-1} z_k\|_2 + k\sqrt{k}\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|R_k^{-1} z_k\|_2 \\ &\le 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_{k+1}\|_2\,\kappa_2(R_k)\,\|\mathbf{b}\|_2 + k\sqrt{k}\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|R_k^{-1}\|_2\,\|\mathbf{b}\|_2. \end{aligned}$$

Here we have used that $\| \,|R_k|\,\|_2 \le \sqrt{3}\,\|R_k\|_2$ (which follows from [15, Theorem 4.2]; see Lemma A.1 for details) and $\| \,|\mathbf{V}_k|\,\|_2 \le \|\mathbf{V}_k\|_F \le \sqrt{k}$. The factor κ_2 denotes the condition number with respect to the Euclidean norm.

Note that we could bound $\|\mathbf{V}_{k+1}\|_2$ by

$$\|\mathbf{V}_{k+1}\|_2 \le \sqrt{k+1},$$

which is, because of the local orthogonality of the $\mathbf{v}_j$, a crude overestimate. According to [12, p. 267 (bottom)], it may be more realistic to replace this factor $\sqrt{k+1}$ by a factor $\sqrt{m}$, where m denotes the maximum number of times that a Ritz value of T_k has converged to any eigenvalue of $\mathbf{A}$. When solving a linear system, this value of m is usually small, e.g., 2 or 3.

We would like to replace R_k in the error bounds by something that can directly be related to $\mathbf{A}$. Therefore, we note that

$$R_k^T R_k = \underline{T}_k^T \underline{T}_k,$$

ignoring errors in the order of $\mathbf{u}$.

It has been shown in [5, 7] that the matrix $\underline{T}_k$ that has been obtained in finite precision arithmetic may be interpreted as the exact Lanczos matrix obtained from a matrix $\widetilde{\mathbf{A}}$ in which eigenvalues of $\mathbf{A}$ are replaced by multiplets. Each multiplet contains eigenvalues that differ by $\mathcal{O}(\mathbf{u}^{\frac{1}{4}})$ from an original eigenvalue of $\mathbf{A}$.[1] With

[1] This order of difference is pessimistic; factors proportional to $\mathbf{u}^{\frac{1}{2}}$, or even $\mathbf{u}$, are more likely but have not been proved [6, section 4.4.2].

$\widetilde{\mathbf{V}}_k$ we denote the orthogonal matrix that generates $\underline{T}_k$, in exact arithmetic, from $\widetilde{\mathbf{A}}$. Hence,

$$\underline{T}_k^T \underline{T}_k = \widetilde{\mathbf{V}}_k^T \widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}} \widetilde{\mathbf{V}}_k,$$

so that

$$\sigma_{\min}(R_k^T R_k) \geq \sigma_{\min}(\widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}) \quad \text{and} \quad \sigma_{\max}(R_k^T R_k) \leq \sigma_{\max}(\widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}),$$

which implies (ignoring errors proportional to mild orders of $\mathbf{u}$)

$$\kappa_2(R_k) \leq \kappa_2(\widetilde{\mathbf{A}}) = \kappa_2(\mathbf{A}). \tag{16}$$

This finally results in an upper bound for the error in the residual for GMRES*, which can be attributed to the evaluation of the generating formula (9):

$$\frac{\|\Delta \mathbf{r}_k\|_2}{\|\mathbf{b}\|_2} \leq (3\sqrt{3}\,\|\mathbf{V}_{k+1}\|_2 + k\sqrt{k})\,\mathbf{u}\,\kappa_2(\mathbf{A}). \tag{17}$$

Note that, even if there were only rounding errors in the representation of $\mathbf{A}$ or $\mathbf{b}$, then we may expect a perturbation $\Delta\mathbf{x}$ to $\mathbf{A}^{-1}\mathbf{b}$ that is (in norm) up to the order of $\mathbf{u}\,\|\mathbf{A}^{-1}\|_2\,\|\mathbf{b}\|_2$. This corresponds to an error $-\mathbf{A}\Delta\mathbf{x}$ in the residual, for which the norm is up to the order of $\mathbf{u}\,\kappa_2(\mathbf{A})\|\mathbf{b}\|_2$. In this sense the stability of GMRES* is optimal.

Our analysis for GMRES* has been restricted to certain parts of the algorithm. For an analysis of all errors in the original GMRES, including those in the Arnoldi process and the Givens rotations, for unsymmetric $\mathbf{A}$, see [3].

2.3. Error analysis for MINRES. For MINRES we have to study the errors in the evaluation in finite precision of $\left(\mathbf{V}_k R_k^{-1}\right) z_k$.

We will first analyze the floating point errors introduced by the computation of the columns of $\mathbf{W}_k = \mathbf{V}_k R_k^{-1}$. The jth row $w_{j,:}$ of $\mathbf{W}_k$ satisfies

$$w_{j,:} R_k = v_{j,:},$$

which means that in floating point finite precision arithmetic we obtain the solution $\widehat{w}_{j,:}$ of a perturbed system:

$$\widehat{w}_{j,:}(R_k + \Delta_{R_j}) = v_{j,:}, \tag{18}$$

with

$$|\Delta_{R_j}| \leq 3\,\mathbf{u}\,|R_k| + \mathcal{O}(\mathbf{u}^2). \tag{19}$$

Note that the perturbation term Δ_{R_j} depends on j. This gives $\widehat{w}_{j,:} R_k = v_{j,:} - \widehat{w}_{j,:}\Delta_{R_j}$, and when we combine the relations for $j = 1, \ldots, k$, we obtain

$$\widehat{\mathbf{W}}_k = (\mathbf{V}_k + \Delta_W) R_k^{-1}, \tag{20}$$

with

$$|\Delta_W| \leq 3\,\mathbf{u}\,\left|\widehat{\mathbf{W}}_k\right|\,|R_k| + \mathcal{O}(\mathbf{u}^2). \tag{21}$$

We may replace $\widehat{\mathbf{W}}_k$ by $\mathbf{W}_k = \mathbf{V}_k R_k^{-1}$ in (21), because this leads only to $\mathcal{O}(\mathbf{u}^2)$ errors.

We also expect errors in the evaluation of $\widehat{\mathbf{x}_k} = fl((\mathbf{V}_k R_k^{-1}) z_k)$ because of finite precision errors in the multiplication of $\widehat{\mathbf{W}}_k$ with z_k:

$$\widehat{\mathbf{x}}_k = \widehat{\mathbf{W}}_k z_k + \Delta_3, \quad \text{with} \quad |\Delta_3| \leq k\,\mathbf{u}\,|\mathbf{W}_k|\,|z_k| + \mathcal{O}(\mathbf{u}^2). \tag{22}$$

The errors in $\widehat{\mathbf{W}}_k$ and the error term Δ_3 describe the errors that are due to the evaluation of the generating formula for MINRES. Added together, they lead to $\Delta\mathbf{x}_k \equiv \widehat{\mathbf{x}}_k - \mathbf{x}_k$ (with $\mathbf{x}_k = \mathbf{V}_k R_k^{-1} z_k$) related to MINRES

$$\Delta\mathbf{x}_k = \Delta_W R_k^{-1} z_k + \Delta_3, \tag{23}$$

and this leads to the following contribution to the MINRES residual:

$$\Delta\mathbf{r}_k \equiv \widehat{\mathbf{r}}_k - \mathbf{r}_k = -\mathbf{A}\Delta\mathbf{x}_k = -\mathbf{A}\Delta_W R_k^{-1} z_k - \mathbf{A}\Delta_3. \tag{24}$$

If we use the bound (21) for Δ_W, and use for other quantities bounds similar to those for GMRES, then we obtain (again, ignoring $\mathcal{O}(\mathbf{u}^2)$ terms)

$$\begin{aligned}
\|\Delta\mathbf{r}_k\|_2 &\leq 3\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\,|\mathbf{V}_k R_k^{-1}|\,\|_2\,\|\,|R_k|\,\|_2\,\|R_k^{-1} z_k\|_2 + k\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\,|\mathbf{V}_k R_k^{-1}|\,\|_2\,\|z_k\|_2 \\
&\leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\mathbf{V}_k\|_F\,\|R_k^{-1}\|_2\,\|R_k\|_2\,\|R_k^{-1}\|_2\,\|\mathbf{b}\|_2 \\
&\qquad\qquad + k\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\mathbf{V}_k\|_F\,\|R_k^{-1}\|_2\,\|\mathbf{b}\|_2 \\
&\leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_F\,\kappa_2(\mathbf{A})^2\,\|\mathbf{b}\|_2 + k\,\mathbf{u}\,\kappa_2(\mathbf{A})\,\|\mathbf{V}_k\|_F\,\|\mathbf{b}\|_2.
\end{aligned}$$

Here we have also used the fact that

$$\|\,|\mathbf{V}_k R_k^{-1}|\,\|_2 \leq \|\mathbf{V}_k R_k^{-1}\|_F \leq \|\mathbf{V}_k\|_F\,\|R_k^{-1}\|_2, \tag{25}$$

and, with $\|\mathbf{V}_k\|_F \leq \sqrt{k}$, the expression can be further bounded.

This results in the following upper bound for the error contribution in the residual for MINRES, due to the computational errors in the generating formula (10):

$$\frac{\|\Delta\mathbf{r}_k\|_2}{\|\mathbf{b}\|_2} \leq 3\sqrt{3k}\,\mathbf{u}\,\kappa_2(\mathbf{A})^2 + k\sqrt{k}\,\mathbf{u}\,\kappa_2(\mathbf{A}). \tag{26}$$

We see that the generating formula for MINRES leads to an upper bound for the norm of the relative error in the residual that is proportional to the squared condition number of $\mathbf{A}$, whereas for GMRES* this led to an upper bound for the relative error in norm proportional to the condition number only; see (17). This means that if we plot the norms of the residuals for MINRES and GMRES*, then the upper bounds suggest that we may expect to see differences.

More specifically, they suggest that the difference between the norms of the computed residuals for the two methods may be expected to be up to the order of the square of the condition number. As soon as the norm of the computed residual of GMRES* (involving all errors made in the process) gets below $\mathbf{u}\,\kappa_2(\mathbf{A})^2\,\|\mathbf{b}\|_2$, then this difference may be visible. Indeed, our experiments display a clear difference between the residual norms for MINRES and GMRES*, in the order of our upper bounds.

2.4. Discussion. In Figure 4, we have plotted the residuals obtained for GMRES* and MINRES. Our analysis suggests that there may be a difference between both up to the order of the square of the condition number times machine precision relative

to $\|\mathbf{b}\|_2$. Of course, the computed residuals reflect all errors made in both processes, and if all these errors together lead to perturbations in the same order for MINRES and GMRES*, then we will not see much difference in the norms of the residuals. However, as we see, all the errors in GMRES* lead to something proportional to the condition number, and now the effect of the square of the condition number is clearly visible in the error in the residual for MINRES.

Our analysis implies that one has to be careful with MINRES when solving linear systems with an ill-conditioned matrix $\mathbf{A}$, especially when eigenvector components in the solution, corresponding to small eigenvalues, are important.

The residual norm reduction $\|\mathbf{r}_k\|_2/\|\mathbf{b}\|_2$ for the exact (but unknown) MINRES residual can be expressed as the product $\rho_k \equiv |s_1 \cdot \ldots \cdot s_k|$ of the sines s_k of the Givens rotations; see [13, Proposition 1]. (See also (57) and its subsequent discussion). This is the last ($(k+1)$th) coordinate of the vector that is obtained by applying the k Givens rotations (used for the annihilation of the subdiagonal elements of $\underline{T}_k$) to the vector e_1 (of length $k+1$). In GMRES the computed value $\widehat{\rho}_k$, computed with the $\widehat{s}_k$, is often used for monitoring the reduction of the residual norm. In practical computations, a residual norm is not often computed explicitly at each iteration step as $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2$, with $\widehat{\mathbf{x}}_k$ the kth floating point approximate solution, because this would require an extra matrix-vector product.

In Figure 4, we have also plotted the computed residual reduction factors $\widehat{\rho}_k$ for MINRES and GMRES*, as dotted curves. We see that the $\widehat{\rho}_k$ are only close to the actual residual reductions (the drawn curves) until where these stagnate: for MINRES this happens at a level proportional to $\kappa_2(\mathbf{A})^2\mathbf{u}$, and for GMRES* this happens at a level proportional to $\kappa_2(\mathbf{A})\mathbf{u}$.

We do not know whether the $\widehat{\rho}_k$ are always close to the actual residual reduction factors before the latter ones stagnate because of errors due to the evaluation of the generating formulas; this might be not the case if there is a severe loss of orthogonality among the columns of $\mathbf{V}_k$ in an earlier phase of the iteration history.

We have not considered the question of how close to orthogonal $\mathbf{V}_{k+1}$ should be, but we have seen that the generating formula (10) for MINRES may lead to errors that are in norm proportional to $\kappa_2(\mathbf{A})^2\mathbf{u}$. Because the $\widehat{\rho}_k$ cannot reflect computational errors in the solution of the reduced system (in fact, the derivation of the ρ_k assumes exact solution of the reduced system), we should expect at least a deviation by that order of magnitude in $\widehat{\rho}_k$ with respect to $\|\mathbf{A}\widehat{\mathbf{x}}_k - \mathbf{b}\|_2/\|\mathbf{b}\|_2$. This suggests that the computed reduction factor may be very unreliable for ill-conditioned matrices $\mathbf{A}$.

The situation for GMRES* is much better: the errors introduced by the evaluation of the generating formula (9) have the same order of magnitude as the errors that we should expect from a small relative perturbation (of order $\mathcal{O}(\mathbf{u})$) of the given system.

2.5. Diagonal matrices. Numerical analysts often carry out experiments for (unpreconditioned) iterative solvers for symmetric systems with diagonal matrices, because, at least in exact arithmetic, the convergence behavior depends on the distribution of the eigenvalues and the structure of the matrix plays no role in Krylov solvers. However, the behavior of these methods for diagonal systems may be quite different in finite precision, as we will now show, and, in particular for MINRES, experiments with diagonal matrices may give a too optimistic view on the behavior of the method.

Rotating the matrix from diagonal to nondiagonal (i.e., $\mathbf{A} = \mathbf{Q}^T\mathbf{D}\mathbf{Q}$, with $\mathbf{D}$ diagonal and $\mathbf{Q}$ orthogonal, instead of $\mathbf{A} = \mathbf{D}$) has hardly any influence on the errors in the GMRES* residuals (no results shown here). This is not the case for MINRES:

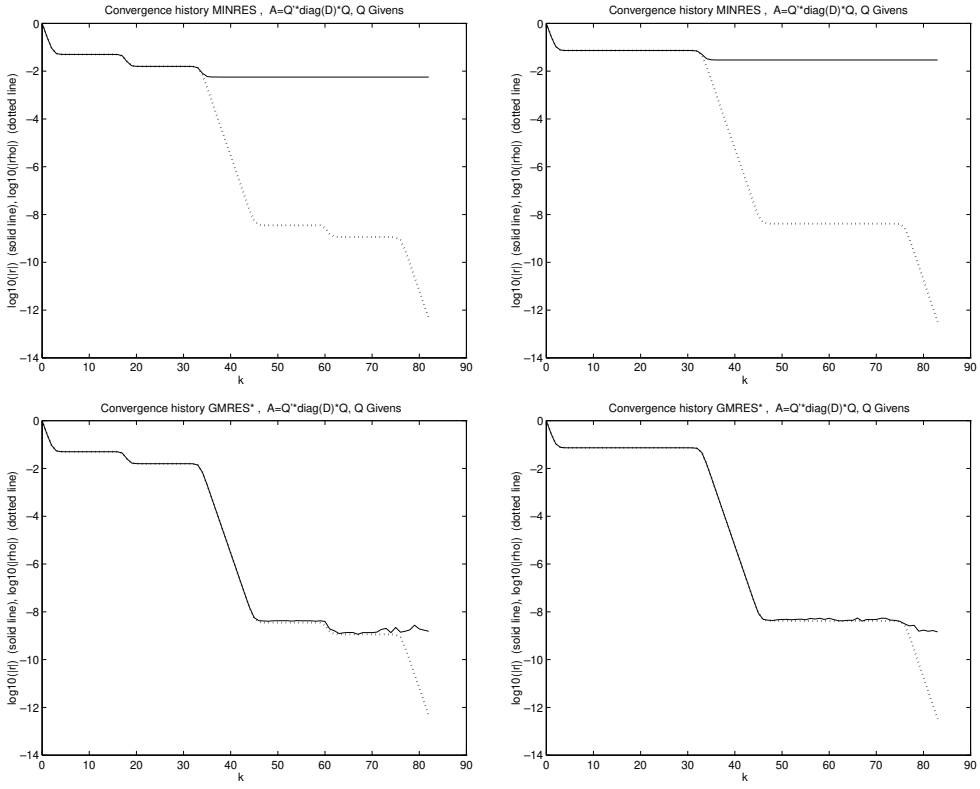

FIG. 4. MINRES (*top*) *and* GMRES* (*bottom*): *solid line* (—) $\log_{10}$ *of* $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2/\|\mathbf{b}\|_2$; *dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* ρ_k. *The pictures show the results for a positive definite system* (*the left pictures*) *and for an indefinite system* (*the right pictures*). *For both examples* $\kappa_2(\mathbf{A}) = 3 \cdot 10^8$. *To be more specific, at the left* $\mathbf{A} = \mathbf{GDG}'$ *with* $\mathbf{D}$ *diagonal,* $\mathbf{D} \equiv \mathrm{diag}(10^{-8}, 2\cdot 10^{-8}, 2 : h : 3)$, $h = 1/789$, *and* $\mathbf{G}$ *the Givens rotation in the* $(1, 30)$*-plane over an angle of* $45°$*; at the right* $\mathbf{A} = \mathbf{GDG}'$ *with* $\mathbf{D}$ *diagonal* $\mathbf{D} \equiv \mathrm{diag}(-10^{-8}, 10^{-8}, 2 : h : 3)$, $h = 1/389$, *and* $\mathbf{G}$ *the same Givens rotation as for the left example; in both examples* (*and others to come*) $\mathbf{b}$ *is the vector with all coordinates equal to 1,* $\mathbf{x}_0 = \mathbf{0}$, *and the relative machine precision* $\mathbf{u} = 1.1 \cdot 10^{-16}$.

experimental results (cf. Figure 5) indicate that the errors in the MINRES residuals for diagonal matrices are of order $\mathbf{u}\,\kappa_2(\mathbf{A})$, similar to GMRES*. This can be understood as follows.

If we neglect $\mathcal{O}(\mathbf{u}^2)$ terms, then, according to (18), the error, due to the inversion of R_k, in the jth coordinate of the MINRES-$\mathbf{x}_k$, due to the evaluation of the generating formula, is given by

$$(\Delta \mathbf{x}_k)_j = (\widehat{w}_{j,:} - w_{j,:})z_k + (\Delta_3)_j = -v_{j,:}\, R_k^{-1} \Delta_{R_j} R_k^{-1} z_k + (\Delta_3)_j,$$

where $(\Delta_3)_j$ is the jth coordinate of Δ_3 (see (22)).

When $\mathbf{A}$ is diagonal with (j, j)-entry λ_j, the error in the jth coordinate of the MINRES residual is equal to (use (1) and (5))

$$\begin{aligned} (\Delta \mathbf{r}_k)_j &= \lambda_j v_{j,:}\, R_k^{-1} \Delta_{R_j} R_k^{-1} z_k - \lambda_j (\Delta_3)_j \\ &= \mathbf{e}_j^T \mathbf{A} \mathbf{V}_k R_k^{-1} \Delta_{R_j} R_k^{-1} z_k - \lambda_j (\Delta_3)_j \\ &= \mathbf{e}_j^T \mathbf{V}_{k+1} \underline{Q}_k \Delta_{R_j} R_k^{-1} z_k - \lambda_j (\Delta_3)_j. \end{aligned} \tag{27}$$

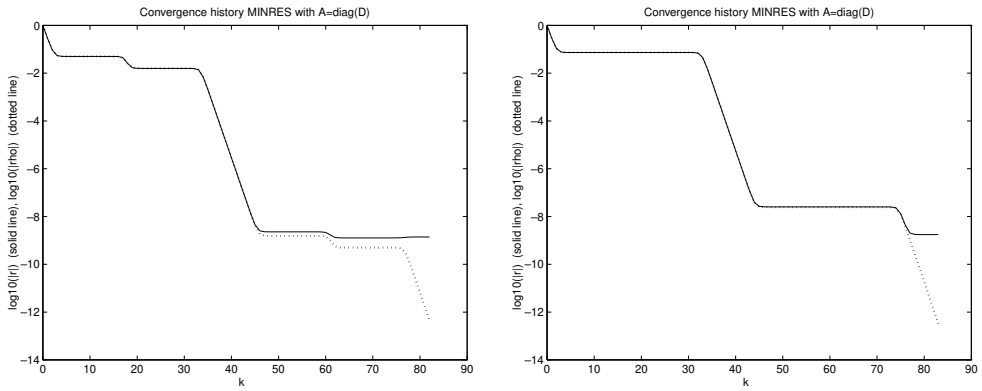

FIG. 5. MINRES: *solid line* (—) $\log_{10}$ *of* $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2/\|\mathbf{b}\|_2$; *dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* $\widehat{\rho}_k$. *The pictures show the results for a positive definite diagonal system* (*the left picture*) *and for an indefinite diagonal system* (*the right picture*). *Except for the Givens rotation, the matrices in these examples are equal to the matrices of the examples in Figure* 4: *here* $\mathbf{G} = \mathbf{I}$.

Therefore, in view of (19), and including the error term for the multiplication with $\widehat{\mathbf{W}}_k$ (cf. (22)), we have for MINRES applied to a diagonal matrix

$$\frac{\|\Delta\mathbf{r}_k\|_2}{\|\mathbf{b}\|_2} \leq (3\sqrt{3}\|\mathbf{V}_{k+1}\|_2 + k\sqrt{k})\,\mathbf{u}\,\kappa_2(\mathbf{A}),$$

which is the same upper bound as for the errors in the GMRES* residuals in (17).

The perturbation matrix Δ_{R_j} depends on the row index j. Since, in general, Δ_{R_j} will be different for each coordinate j, (27) cannot be expected to be correct for non-diagonal matrices. In fact, if $\mathbf{A} = \mathbf{Q}^T\mathrm{diag}(\lambda_j)\mathbf{Q}$, with $\mathbf{Q}$ some orthogonal matrix, then errors of order $\mathbf{u}\,\|R_k^{-1}\|_2\,\kappa_2(R_k)$ in the jth coordinate of $\mathbf{x}_k$ can be transferred by $\mathbf{Q}$ to an mth coordinate and may not be damped by a small value $|\lambda_m|$. More precisely, if Γ is the maximum size of the off-diagonal elements of $\mathbf{A}$ that "couple" small diagonal elements of $\mathbf{A}$ to large ones, then the error in the MINRES residual will be of order $\Gamma\,\mathbf{u}\,\|R_k^{-1}\|_2\,\kappa_2(R_k^{-1}) \leq \Gamma\,\mathbf{u}\,\|\mathbf{A}^{-1}\|_2\,\kappa_2(\mathbf{A})$. If $\Gamma \approx \|\mathbf{A}\|_2$, we recover essentially the bound (26).

2.6. The errors in the approximations. In exact arithmetic we have that $\|\mathbf{x}_k\|_2 = \|\mathbf{V}_k R_k^{-1} z_k\|_2 = \|R_k^{-1} z_k\|_2$. We will in this section assume that, in finite precision, this also gives approximately the right order of magnitude for representations of the solution

$$\|\widehat{\mathbf{x}}_k\|_2 \approx \|\mathbf{x}_k\|_2 = \|y_k\|_2.$$

Then the errors (14) and (23), related to the evaluation of the generating formulas (9) and (10), respectively, can be bounded by essentially the same upper bound:

$$\frac{\|\Delta\mathbf{x}_k\|_2}{\|\widehat{\mathbf{x}}_k\|_2} \lesssim (3\sqrt{3} + k\sqrt{k})\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(R_k) \leq (3\sqrt{3k} + k\sqrt{k})\mathbf{u}\,\kappa_2(\mathbf{A}). \tag{28}$$

This may come as a surprise since the bound for the error contribution to the residual for MINRES is proportional to $\kappa_2(\mathbf{A})^2$.

Based upon our observations in numerical experiments, we think that this can be explained as follows. The error in the GMRES* approximation has its relatively largest components mainly in the direction of the 'small' eigenvectors of $\mathbf{A}$. These components are relatively reduced by the multiplication with $\mathbf{A}$, and then have less effect to the norm of the residual.

On the other hand, the errors in the MINRES approximation are more or less of the same magnitude over the spectrum of eigenvalues of $\mathbf{A}$. Multiplication with $\mathbf{A}$ will make error components associated with larger eigenvalues more effective in the residual.

We will support our viewpoint by a numerical example. The results in Figure 6 are obtained with a positive definite matrix with two tiny eigenvalues. For $\mathbf{b}$ we took a random perturbation of $\mathbf{Ay}$ in the order of 0.01: $\mathbf{b} = \mathbf{Ay} + \mathbf{p}$, $\|\mathbf{p}\|_2 \leq 10^{-2}$. This example mimics the situation where the right-hand-side vector is affected by errors from measurements. The solution $\mathbf{x}$ of the equation $\mathbf{Ax} = \mathbf{b}$ has huge components in the direction of the two eigenvectors with smallest eigenvalue. In the other directions $\mathbf{x}$ is equal to $\mathbf{y}$ plus a perturbation of less than one percent. The coordinates of the vector $\mathbf{y}$ in our example form a parabola, which makes the effects more easily visible.

The convergence histories of GMRES* and of MINRES (not shown here) for this example with $\mathbf{x}_0 = \mathbf{0}$ are comparable to the ones in the left pictures of Figure 4, but, because of a higher condition number, the final stagnation of the residual norm in the present example takes place on a higher level ($\approx 3 \cdot 10^{-8}$ for GMRES* and $\approx 10^0$ for MINRES).

Figure 6 shows the solution $\mathbf{x}_k$ as computed at the 80th step of GMRES (top pictures) and of MINRES (bottom pictures); the right pictures show the component of $\mathbf{x}_k$ orthogonal to the two eigenvectors with smallest eigenvalue, while the left pictures show the complete $\mathbf{x}_k$. Note that $\|\mathbf{x}_k\|_2 \approx 10^7$. The curve of the projected GMRES* solution (top right picture) is a slightly perturbed parabola indeed (the irregularities are due to the perturbation $\mathbf{p}$). The computational errors from the GMRES* process are not visible in this picture: these errors are mainly in the direction of the two 'small' eigenvectors.

In contrast, the irregularities in the MINRES curve (bottom right picture) are almost exclusively the effect of rounding errors in the MINRES process.

3. Error analysis for SYMMLQ. In SYMMLQ we minimize the norm of $\mathbf{x} - \mathbf{x}_k$, for $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{A}\mathbf{V}_k y_k$, which means that y_k is the solution of the normal equations

$$\mathbf{V}_k{}^T \mathbf{A}^T \mathbf{A} \mathbf{V}_k y_k = \mathbf{V}_k{}^T \mathbf{A}^T (\mathbf{x} - \mathbf{x}_0) = \mathbf{V}_k{}^T \mathbf{r}_0 = \|\mathbf{r}_0\|_2 \, e_1.$$

This system can be further simplified by exploiting the Lanczos relations (1):

$$\mathbf{V}_k{}^T \mathbf{A}^T \mathbf{A} \mathbf{V}_k = \underline{T}_k^T \mathbf{V}_{k+1}^T \mathbf{V}_{k+1} \underline{T}_k = \underline{T}_k^T \underline{T}_k.$$

A stable way of solving this set of normal equations is based on an LQ decomposition of $\underline{T}_k^T$, and this is equivalent to the transpose of the QR decomposition of $\underline{T}_k$ (see (5)), which is constructed for GMRES* and MINRES:

$$\underline{T}_k^T = R_k^T \underline{Q}_k^T.$$

This leads to

$$\underline{T}_k^T \underline{T}_k y_k = R_k^T R_k y_k = \|\mathbf{r}_0\|_2 \, e_1,$$

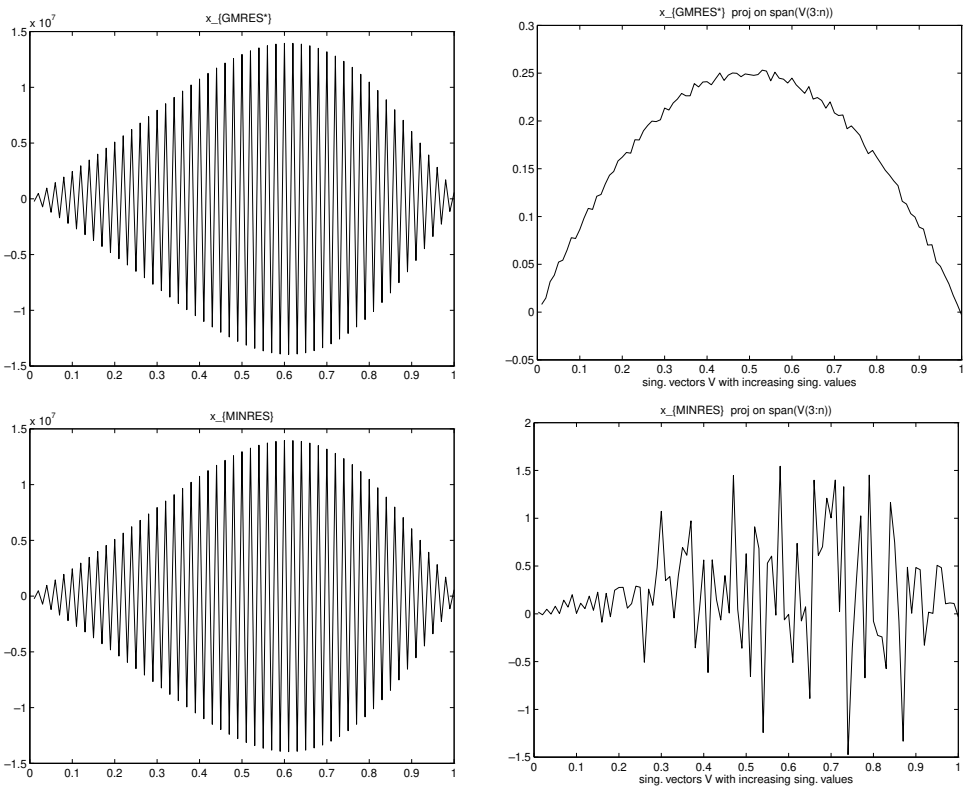

FIG. 6. *The pictures show the solution* $\mathbf{x}$ *of* $\mathbf{Ax} = \mathbf{b}$*, computed with* 80 *steps of* GMRES* (*top pictures*) *and of* MINRES (*bottom pictures*)*. The* i*th coordinate of* $\mathbf{x}_k$ (*along the vertical axis*) *is plotted against* $\frac{i}{n}$ (*along the horizontal axis*)*.* $\mathbf{A} = \mathbf{Q}^*\mathbf{DQ}$ *with* $\mathbf{D} = \mathrm{diag}(10^{-10}, 2 \cdot 10^{-10}, 2 : h : 3)$*,* $h = 1/97$ *and* $\mathbf{Q}$ *unitary,* $\mathbf{Q}_{ij} = \sqrt{\frac{2}{n+1}} \sin \frac{i(n+1-j)}{(n+1)\pi}$*,* $n = 100$*.* $\mathbf{b} = \mathbf{Ay} + \mathbf{p}$ *with* $y_i = \frac{i}{n}(1 - \frac{i}{n})$*, and* $\mathbf{p}$ *random,* $\|\mathbf{p}\|_2 \leq 0.01$*. The right pictures show the component of* $\mathbf{x}_k$ *orthogonal to the two eigenvectors with smallest eigenvalue, while the left pictures show the complete* $\mathbf{x}_k$*.*

from which the basic generating formula for SYMMLQ is obtained:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{x}_0 + \mathbf{A}\mathbf{V}_k R_k^{-1} R_k^{-T} \|\mathbf{r}_0\|_2 \, e_1 \\ &= \mathbf{x}_0 + \mathbf{V}_{k+1} \underline{T}_k R_k^{-1} R_k^{-T} \|\mathbf{r}_0\|_2 \, e_1 \\ &= \mathbf{x}_0 + (\mathbf{V}_{k+1} \underline{Q}_k) \, (L_k^{-1} \|\mathbf{r}_0\|_2 \, e_1), \end{aligned} \tag{29}$$

with $L_k \equiv R_k^T$. We will further assume that $\mathbf{x}_0 = \mathbf{0}$ and hence $\mathbf{r}_0 = \mathbf{b}$. This gives the following generating formula:

$$\mathbf{x}_k = (\mathbf{V}_{k+1} \underline{Q}_k) \, (L_k^{-1} \|\mathbf{b}\|_2 \, e_1). \tag{30}$$

The actual implementation of SYMMLQ [11] is based on an update procedure for $\mathbf{V}_{k+1}\underline{Q}_k$, and on a three-term recurrence relation for $g_k \equiv \|\mathbf{b}\|_2 \, L_k^{-1} e_1$.

The differences in finite precision between MINRES and GMRES* could be analyzed by studying the differences in the evaluation of the generating formula for these methods (see (6)):

$$\mathbf{x}_k = \mathbf{V}_k R_k^{-1} \underline{Q}_k^T \|\mathbf{b}\|_2 e_1. \tag{31}$$

Note that, because of $L_k = R_k^T$, the generating formulas for the three methods contain in principle the same computed ingredients $\mathbf{V}_{k+1}$, $\underline{Q}_k$, R_k, and $\mathbf{b}$. In fact, we see no good reason for using differently computed values for each of the algorithms.

The methods MINRES and GMRES* have been characterized by a different order of evaluation of essentially the same generating formula (see (9) and (10)). For SYMMLQ we have a completely different generating formula which even in exact arithmetic leads to completely different results. Observed differences in the results for SYMMLQ, compared to MINRES and GMRES*, can by no means be attributed to computational errors. However, we have tried to make plausible that eventually the norm of the residual for MINRES may be contaminated by a term proportional to $\|\mathbf{b}\|_2\kappa_2(\mathbf{A})^2\mathbf{u}$, which may lead to a stagnation of the residual norm at a significantly higher level than for GMRES*; see, for instance, Figure 4. Since SYMMLQ may be considered as an alternative for MINRES (one reason is that it avoids storage of the full $\mathbf{V}_{k+1}$), it may be of interest to see whether computational errors in the generating formula may have a similar polluting effect on the residual as for MINRES. Note that even if we can answer this question, then this does not reveal all differences due to rounding errors in MINRES and SYMMLQ. One reason could be that rounding errors in $\mathbf{V}_k$ manifest themselves differently (because of the right multiplication with $\underline{Q}_k$), although this does not seem very likely to us because of the (near) orthogonality of $\underline{Q}_k$.

We postulate that the main factor, for ill-conditioned systems, in the upper bound for the norm of the additional rounding errors in the residual for SYMMLQ, due to the evaluation of the generating formula, comes from solving $L_k g_k = \|\mathbf{b}\|_2 e_1$ for g_k. In order to simplify our rather complicated analysis for SYMMLQ, we have chosen to study only the effect of the errors introduced by this part of the formula.

The resulting error $\Delta\mathbf{x}_k$ is written as

$$\Delta\mathbf{x}_k = \mathbf{V}_{k+1}\,\underline{Q}_k(\widehat{g}_k - g_k) \quad \text{with} \quad L_k g_k = \|\mathbf{b}\|_2\, e_1, \tag{32}$$

where g_k represents the exact solution and $\widehat{g}_k$ is the value obtained in finite precision arithmetic. We write $g_k/\|\mathbf{b}\|_2 = (\gamma_1, \ldots, \gamma_k)^T$, and likewise the coordinates of $\widehat{g}_k/\|\mathbf{b}\|_2$ are denoted by $\widehat{\gamma}_j$. These coordinates can be written as

$$\gamma_k = e_k^T L_k^{-1} e_1, \quad \widehat{\gamma}_k = e_k^T (L_k + \Delta_L)^{-1} e_1, \quad \text{with} \quad |\Delta_L| \le 3\,\mathbf{u}\,|L_k| + \mathcal{O}(\mathbf{u}^2). \tag{33}$$

In order to simplify our formulas, we will omit the $\mathcal{O}(\mathbf{u}^2)$ terms in the further analysis.

For the analysis of the residual, we will be interested in the term $\mathbf{A}\mathbf{V}_{k+1}\,\underline{Q}_k$. Using the relation for the finite precision Lanczos process, we have (cf. (2))

$$\mathbf{A}\mathbf{V}_{k+1}\,\underline{Q}_k = \mathbf{V}_{k+2}\,\underline{T}_{k+1}\,\underline{Q}_k + \mathbf{F}_{k+1}\,\underline{Q}_k.$$

Since T_{k+3} is symmetric, we have for its submatrices that

$$\underline{T}_{k+1} = \underline{T}_{k+2}^T\,\underline{\underline{I}}_{k+1},$$

where $\underline{\underline{I}}_{k+1}$ is the $k+3$ by $k+1$ left block of the $k+3$-dimensional identity matrix. Moreover, for the LQ decomposition in finite precision, we have (cf. (11))

$$\underline{T}_{k+2}^T = L_{k+2}\,\underline{Q}_{k+2}^T + \underline{G}_{k+2}^T.$$

The matrix $\underline{Q}_{k+2}$ is upper Hessenberg. Hence, $\underline{\underline{I}}_{k+1}\,\underline{Q}_k$ consists of the first k columns of $\underline{Q}_{k+2}$ and orthogonality of $\underline{Q}_{k+2}$ implies that

$$\underline{Q}_{k+2}^T\,\underline{\underline{I}}_{k+1}\,\underline{Q}_k = \underline{\underline{I}}_k.$$

Hence, taking into account that $L_{k+2} = (\ell_{i,j})$ is lower tridiagonal ($\ell_{i,j} \neq 0$ only if $i \leq j \leq i+2$),

$$\begin{aligned}\mathbf{A}\mathbf{V}_{k+1}\underline{Q}_k &= \mathbf{V}_{k+2}\underline{T}_{k+1}\underline{Q}_k + \mathbf{F}_{k+1}\underline{Q}_k \\ &= \mathbf{V}_{k+2}L_{k+2}\underline{I}_k + \mathbf{V}_{k+2}\underline{G}_{k+2}^T\underline{I}_{k+1}\underline{Q}_k + \mathbf{F}_{k+1}\underline{Q}_k \\ &= \mathbf{V}_k L_k + [\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} + \mathbf{F}'_{k+1}, \end{aligned} \tag{34}$$

where M_k is the right 2 by 2 lower block of $L_{k+2}\underline{I}_k$,

$$M_k \equiv \begin{bmatrix} \ell_{k+1,k-1} & \ell_{k+1,k} \\ 0 & \ell_{k+2,k} \end{bmatrix},$$

and

$$\mathbf{F}'_{k+1} \equiv \mathbf{V}_{k+2}\underline{G}_{k+2}^T\underline{I}_{k+1}\underline{Q}_k + \mathbf{F}_{k+1}\underline{Q}_k.$$

Note that, on account of (3) and (11),

$$\|\mathbf{F}'_{k+1}\|_2 \leq c'\, k^2\sqrt{k}\,\mathbf{u}\,\|\mathbf{A}\|_2 \tag{35}$$

for some modest constant c'.

We will use that $(L_k + \Delta_L)^{-1} = L_k^{-1} - L_k^{-1}\Delta_L L_k^{-1}$ (neglecting $\mathcal{O}(\mathbf{u}^2)$ terms; cf. (33)). Then, from (34), we find for the residual $\widehat{\mathbf{r}}_k$ corresponding to the computed approximation $\widehat{\mathbf{x}}_k = \mathbf{x}_k + \Delta\mathbf{x}_k$ (see (32)),

$$\begin{aligned}\widehat{\mathbf{r}}_k \equiv \mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k &= \mathbf{b} - \mathbf{A}\mathbf{V}_{k+1}\underline{Q}_k\,(L_k + \Delta_L)^{-1}\|\mathbf{b}\|_2\, e_1 \\ &= \mathbf{b} - \mathbf{V}_k L_k L_k^{-1}\|\mathbf{b}\|_2 e_1 + \mathbf{V}_k L_k L_k^{-1}\Delta_L L_K^{-1}\|\mathbf{b}\|_2 e_1 \\ &\quad - \left([\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} + \mathbf{F}'_{k+1}\right)(L_k + \Delta_L)^{-1}\|\mathbf{b}\|_2\, e_1 \\ &= \mathbf{V}_k\Delta_L\|\mathbf{b}\|_2 L_k^{-1}e_1 - \|\mathbf{b}\|_2\,[\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]\,\widehat{t}_k - \mathbf{F}'_{k+1}\,(L_k + \Delta_L)^{-1}\|\mathbf{b}\|_2\, e_1, \end{aligned} \tag{36}$$

where

$$\widehat{t}_k \equiv M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix}(L_k + \Delta_L)^{-1}e_1. \tag{37}$$

For the process where the system $L_k g_k = \|\mathbf{b}\|_2 e_1$ is solved exactly ($\Delta_L = 0$), we have

$$\mathbf{r}_k \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_k = -\|\mathbf{b}\|_2\,[\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]\,t_k - \mathbf{F}'_{k+1}\,L_k^{-1}\|\mathbf{b}\|_2\, e_1, \tag{38}$$

where

$$t_k \equiv M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1}e_1.$$

Neglecting order $\mathbf{u}^2$ terms (e.g., stemming from $\mathbf{F}'_{k+1}\Delta_L$), we conclude that the error in the SYMMLQ residual $\mathbf{r}_k$, due to the solution of $L_k g_k = \|\mathbf{b}\|_2\, e_1$ in finite precision, can be written as

$$\Delta\mathbf{r}_k \equiv \widehat{\mathbf{r}}_k - \mathbf{r}_k = \|\mathbf{b}\|_2\mathbf{V}_k\Delta_L L_k^{-1}e_1 - \|\mathbf{b}\|_2\,[\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]\,(\widehat{t}_k - t_k). \tag{39}$$

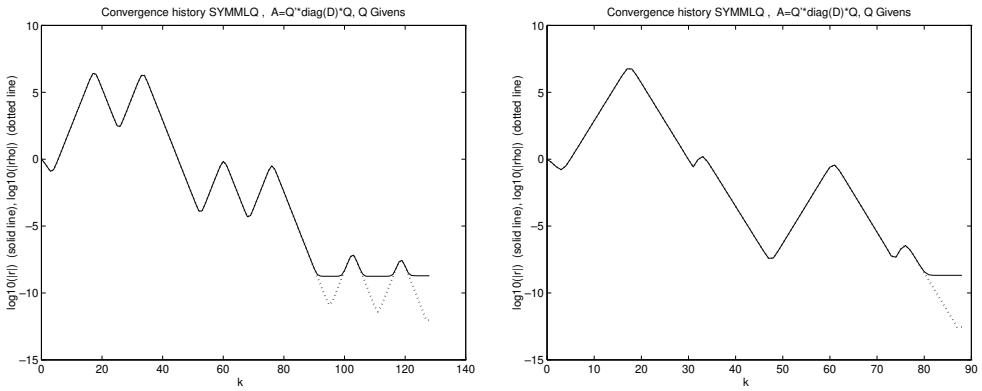

FIG. 7. SYMMLQ: *solid line* (—) $\log_{10}$ *of* $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2/\|\mathbf{b}\|_2$; *dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* $\|\widehat{t}_k\|_2$. *The pictures show the results for the positive definite system* (*the left picture*) *and for the indefinite system* (*the right picture*) *of Figure* 4. *Both systems have condition number* $3 \cdot 10^8$.

To obtain a bound for norm of this error, note that (see (16))

$$\begin{aligned} \|\mathbf{V}_k \Delta_L L_k^{-1} e_1\|_2 &\leq 3\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\|\,|L_k|\,\|_2\,\|L_k\|_2 \leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(L_k) \\ &= 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(R_k) \leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(\mathbf{A}). \end{aligned} \tag{40}$$

Since $\mathbf{v}_{k+1}$ and $\mathbf{v}_{k+2}$ are orthonormal up to machine precision, this leads to

$$\frac{\|\Delta\mathbf{r}_k\|_2}{\|\mathbf{b}\|_2} \leq 3\sqrt{3}\,\|\mathbf{V}_k\|_2\,\mathbf{u}\,\kappa_2(\mathbf{A}) + (1 + c'\mathbf{u})\|\widehat{t}_k - t_k\|_2 \tag{41}$$

for some modest constant c'. A straightforward estimate is

$$\|\widehat{t}_k - t_k\|_2 = \left\| M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} \Delta_L L_k^{-1} e_1 \right\|_2 \leq 3\sqrt{3}\,\mathbf{u}\,\kappa_2(L_k)^2 \leq 3\sqrt{3}\,\mathbf{u}\,\kappa_2(\mathbf{A})^2, \tag{42}$$

which is much larger than the first term in (41). Experiments indicate that $\|\widehat{t}_k - t_k\|_2$ converges towards 0 (even below the value $\mathbf{u}\,\kappa_2(\mathbf{A})$). Below, we will explain why this is to be expected (cf. (60)). Figure 7 illustrates that the upper bound in (41), with $\|\widehat{t}_k - t_k\|_2 \approx 0$, is fairly sharp.

Accuracy. In exact arithmetic (where also $\mathbf{F}_{k+1} = \mathbf{0}$ and $\underline{G}_{k+2} = 0$), the norm $\|\mathbf{r}_k\|_2$ of the SYMMLQ residual is equal to $\|t_k\|_2$ (as can be seen from (38)). Therefore, the computed residual norm reduction $\|\widehat{t}_k\|_2$ is usually used for monitoring the convergence in a stopping criterion. In actual computations with SYMMLQ, no residual vectors are computed. To see how close $\|\widehat{t}_k\|_2$ is to the reduction $\|\widehat{\mathbf{r}}_k\|_2/\|\mathbf{b}\|_2$ of the norm of the actual residual, first note that rounding errors in the multiplication in (37) by M_k and in (36) by $[\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]$ can be bounded by some modest multiple of $\mathbf{u}\,\kappa_2(L_k)$.[2] These bounds will be neglected in the estimates below: since $\kappa_2(L_k) \leq \kappa_2(\mathbf{A})$ (see (16)), they are much smaller than the bound on $\|\mathbf{F}'_{k+1} L_k^{-1} e_1\|_2$ arising from (35). The rounding errors in $\mathbf{v}_{k+1}$ and $\mathbf{v}_{k+2}$ have a similar effect: these vectors are orthonormal up to machine precision.

[2]Note the contrast in the effect of errors in the multiplication by M_k and in the solution of $L_k g_k = e_1$ (cf. (42)).

From (36), (35), and (40), neglecting relatively small terms, it follows that

$$\left| \|\widehat{t}_k\|_2 - \frac{\|\widehat{\mathbf{r}}_k\|_2}{\|\mathbf{b}\|_2} \right| \leq \|\mathbf{V}_k \Delta_L L_k^{-1} e_1\|_2 + \|\mathbf{F}'_{k+1} L_k^{-1} e_1\|_2 \leq c' \, k^{2\frac{1}{2}} \, \mathbf{u} \, \kappa_2(\mathbf{A}). \tag{43}$$

Apparently, SYMMLQ is rather accurate since, for any method, errors in the order $\mathbf{u}\,\kappa_2(\mathbf{A})$ should be expected anyway.

Convergence. It is not clear yet whether the convergence of SYMMLQ is insensitive to rounding errors in the assembly of $\mathbf{x}_k$ (cf. (31)). This would follow from (41) if both t_k and $\widehat{t}_k$ would approach 0. It is unlikely that $\|t_k\|_2$ will be (much) larger than $\|\widehat{t}_k\|_2$, that is, it is unlikely that the inexact process converges faster than the process in exact arithmetic. Therefore, when it is observed that $\|\widehat{t}_k\|_2$ is small (of order $\mathbf{u}\,\kappa_2(\mathbf{A})$), it may be concluded that the speed of convergence has not been affected seriously by rounding errors in the assembly of $\mathbf{x}_k$. In experiments, we see that $\widehat{t}_k$ approaches zero if k increases.

For practical applications, assuming that $\|t_k\|_2 \lesssim \|\widehat{t}_k\|_2$, it is useful to know that the computable value $\|\widehat{t}_k\|_2$ informs us on the accuracy of the computed approximate and on a possible loss of speed of convergence. However, it is of interest to know in advance whether the computed residual reduction will decrease to 0. Moreover, we would like to know whether $\|t_k\|_2 \lesssim \|\widehat{t}_k\|_2$. Of course, it is impossible to prove that SYMMLQ will converge for any symmetric problem: one can easily construct examples for which $\|\mathbf{r}_k\|_2$ will be of order 1 for any $k < n$. But, as we will analyze in the next subsection, the interesting quantities can be bounded in terms of the MINRES residual. That result will be used in order to show that the term $\|\widehat{t}_k - t_k\|_2$ will be relatively unimportant as soon as MINRES has converged to some degree.

3.1. A relation between SYMMLQ and MINRES residual norms. In this subsection we will assume exact arithmetic (in particular, the underlying Lanczos process is assumed to be exact, too). The residuals $\mathbf{r}_k^{\text{MR}}$ and $\mathbf{r}_k^{\text{ME}}$ denote the residuals of MINRES and SYMMLQ, respectively.

The norm of the residual $\mathbf{b} - \mathbf{A}\mathbf{x}^{\text{b}}$, with $\mathbf{x}^{\text{b}}$ the best approximate of $\mathbf{x}$ in $\mathcal{K}_k(\mathbf{A}; \mathbf{b})$, i.e., $\|\mathbf{x} - \mathbf{x}^{\text{b}}\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{y} \in \mathcal{K}_k(\mathbf{A}; \mathbf{b})$, can be bounded in terms of the norm of the MINRES residual $\mathbf{r}_k^{\text{MR}}$:

$$\frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}^{\text{b}}\|_2}{\|\mathbf{r}_k^{\text{MR}}\|_2} \leq \kappa_2(\mathbf{A}). \tag{44}$$

This follows from the observation that $\mathbf{r}_k^{\text{MR}} = \mathbf{b} - \mathbf{A}\mathbf{x}_k^{\text{MR}}$, where $\mathbf{x}_k^{\text{MR}}$ is from the same subspace from which the best approximate $\mathbf{x}^{\text{b}}$ has been selected, and furthermore, that $\|\mathbf{b} - \mathbf{A}\mathbf{x}^{\text{b}}\|_2 \leq \|\mathbf{A}\|_2 \, \|\mathbf{x} - \mathbf{x}^{\text{b}}\|_2$ and $\|\mathbf{x} - \mathbf{x}_k^{\text{MR}}\|_2 \leq \|\mathbf{A}^{-1}\|_2 \, \|\mathbf{r}_k^{\text{MR}}\|_2$. Unfortunately, SYMMLQ selects its approximation $\mathbf{x}_k$ from a different subspace, namely $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{b})$. This makes a comparison less straightforward.

The following lemma will be used for bounding the SYMMLQ error in terms of the MINRES error. Its proof uses the fact that $\mathbf{r}_k^{\text{MR}}$ connects $\mathcal{K}_{k+1}(\mathbf{A}; \mathbf{b})$ and $\mathbf{A}\,\mathcal{K}_k(\mathbf{A}; \mathbf{b})$, that is, $\mathbf{r}_k^{\text{MR}} \in \mathcal{K}_{k+1}(\mathbf{A}; \mathbf{b})$, $\mathbf{r}_k^{\text{MR}} \perp \mathbf{A}\,\mathcal{K}_k(\mathbf{A}; \mathbf{b})$, and hence $\mathcal{K}_{k+1}(\mathbf{A}; \mathbf{b})$ is spanned by $\mathbf{r}_k^{\text{MR}}$ and $\mathbf{A}\,\mathcal{K}_k(\mathbf{A}; \mathbf{b})$.

LEMMA 3.1. *For each* $\mathbf{z} \in \mathcal{K}_{k+1}(\mathbf{A}; \mathbf{b})$, *we have*

$$\|\mathbf{x} - \mathbf{x}_k^{\text{ME}}\|_2^2 \leq \|\mathbf{x} - \mathbf{z}\|_2^2 + |\alpha_k|^2 \, \|\mathbf{r}_k^{\text{MR}}\|_2^2, \quad \textit{where} \quad \alpha_k \equiv \frac{(\mathbf{x}, \mathbf{r}_k^{\text{MR}})}{\|\mathbf{r}_k^{\text{MR}}\|_2^2}. \tag{45}$$

Proof. By construction $\mathbf{x}_k^{\mathrm{ME}}$ minimizes $\|\mathbf{x}-\mathbf{z}\|_2$ over all $\mathbf{z}$ in the space $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{b})$. Hence $\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}} \perp \mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{b})$. Since $\mathbf{r}_k^{\mathrm{MR}} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{b})$, it follows that $(\mathbf{x}_k^{\mathrm{ME}}, \mathbf{r}_k^{\mathrm{MR}}) = 0$, and therefore,

$$\alpha_k = (\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}, \mathbf{r}_k^{\mathrm{MR}})/\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2 \quad \text{and} \quad \mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}} \perp \mathbf{r}_k^{\mathrm{MR}}. \tag{46}$$

Since $\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{b})$ and $\mathbf{r}_k^{\mathrm{MR}} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{b})$, (46) implies that

$$\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}} \perp \mathcal{K}_{k+1}(\mathbf{A};\mathbf{b}).$$

By construction we have that $\mathbf{x}_k^{\mathrm{ME}} - \alpha_k\mathbf{r}_k^{\mathrm{MR}} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{b})$ and, as a consequence,

$$\|\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}}\|_2 \le \|\mathbf{x}-\mathbf{z}\|_2 \quad \text{for all} \quad \mathbf{z} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{b}). \tag{47}$$

From Pythagoras's theorem, with (46), we conclude that

$$\|\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}\|_2^2 = \|\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}}\|_2^2 + |\alpha_k|^2\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2,$$

and (45) follows by combining this result with (47). ∎

Unfortunately, a combination of (45) with $\mathbf{z} = \mathbf{x}_k^{\mathrm{MR}}$ and the obvious estimate $|\alpha_k|\,\|\mathbf{r}_k^{\mathrm{MR}}\|_2 \le \|\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}\|_2$ from (46) does not lead to a useful result. An interesting result follows from an upper bound for $|\alpha_k|$ that can be obtained from a relation between two consecutive MINRES residuals and a Lanczos basis vector. This result is formulated in the next theorem.

THEOREM 3.2.

$$\|\mathbf{r}_k^{\mathrm{ME}}\|_2 \le \nu_{k+1}\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_k^{\mathrm{MR}}\|_2 \quad \textit{with} \quad \nu_k \equiv k + \tfrac{1}{2}\ln(k). \tag{48}$$

Proof. We use the relation

$$\mathbf{r}_k^{\mathrm{MR}} = s^2\mathbf{r}_{k-1}^{\mathrm{MR}} + c^2\mathbf{r}_k^{\mathrm{CG}}, \tag{49}$$

where

$$s \equiv \frac{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2}, \tag{50}$$

and $\mathbf{r}_k^{\mathrm{CG}}$ is the kth conjugate gradient residual. The scalars s and c represent the Givens transformation used in the kth step of MINRES. This relation is a special case of the slightly more general relation between GMRES and FOM residuals, formulated in [1, 16]. For symmetric $\mathbf{A}$, GMRES is equivalent with MINRES, and FOM is equivalent with CG.

Since $\mathbf{r}_k^{\mathrm{CG}} = \|\mathbf{r}_k^{\mathrm{CG}}\|_2\mathbf{v}_{k+1} \perp \mathbf{r}_{k-1}^{\mathrm{MR}} \in \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, it follows that

$$\mathbf{r}_k^{\mathrm{MR}} = s^2\mathbf{r}_{k-1}^{\mathrm{MR}} + \gamma\mathbf{v}_{k+1}, \tag{51}$$

where $\gamma = c^2\|\mathbf{r}_k^{\mathrm{CG}}\|_2$.

Since $\gamma\mathbf{v}_{k+1} \perp \mathbf{r}_{k-1}^{\mathrm{MR}} \in \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, it follows that $\|\gamma\mathbf{v}_{k+1}\|_2 \le \|\mathbf{r}_k^{\mathrm{MR}}\|_2$. Moreover, since $\mathbf{r}_{k-1}^{\mathrm{MR}} \perp \mathbf{A}\,\mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0)$ and $\gamma\mathbf{v}_{k+1} \perp \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, we have that $\mathbf{r}_{k-1}^{\mathrm{MR}} \perp \mathbf{x}_{k-1}^{\mathrm{ME}}$ and $\gamma\mathbf{v}_{k+1} \perp \mathbf{x}_k^{\mathrm{MR}}$. Therefore, with $\mathbf{e}_j^{\mathrm{ME}} \equiv \mathbf{x}-\mathbf{x}_j^{\mathrm{ME}}$, relation (51) implies

$$\begin{aligned}
|\alpha_k|\,\|\mathbf{r}_k^{\mathrm{MR}}\|_2 &= \left|\left(\mathbf{x}, \frac{\mathbf{r}_k^{\mathrm{MR}}}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}\right)\right| \le \frac{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2^2}\left|\left(\mathbf{x}, \frac{\mathbf{r}_{k-1}^{\mathrm{MR}}}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}\right)\right| + \left|\left(\mathbf{x}, \frac{\gamma\mathbf{v}_{k+1}}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}\right)\right| \\
&= \frac{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2^2}\left|\left(\mathbf{x}-\mathbf{x}_{k-1}^{\mathrm{ME}}, \frac{\mathbf{r}_{k-1}^{\mathrm{MR}}}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}\right)\right| + \left|\left(\mathbf{x}-\mathbf{x}_k^{\mathrm{MR}}, \frac{\gamma\mathbf{v}_{k+1}}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}\right)\right|,
\end{aligned}$$

and hence,

$$|\alpha_k| \le \frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{x} - \mathbf{x}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}. \tag{52}$$

A combination of (52) and (45) with $\mathbf{z} = \mathbf{x}_{k+1}^{\mathrm{MR}}$ leads to

$$\frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} \le \frac{\|\mathbf{x} - \mathbf{x}_{k+1}^{\mathrm{MR}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} + \left(\frac{\|\mathbf{e}_{k-1}^{\mathrm{ME}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{x} - \mathbf{x}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \right)^2. \tag{53}$$

With

$$\beta_k \equiv \frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2}{\|\mathbf{A}^{-1}\|_2 \, \|\mathbf{r}_k^{\mathrm{MR}}\|_2},$$

and using the minimal residual property $\|\mathbf{r}_{k+1}^{\mathrm{MR}}\|_2 \le \|\mathbf{r}_k^{\mathrm{MR}}\|_2$, we obtain the following recursive upper bound from (53):

$$\beta_k^2 \le 1 + (\beta_{k-1} + 1)^2.$$

Now, a simple induction argument, using

$$\beta_0 = \frac{1}{\|\mathbf{A}^{-1}\|_2} \frac{\|\mathbf{e}_0^{\mathrm{ME}}\|_2}{\|\mathbf{r}_0^{\mathrm{MR}}\|_2} = \frac{1}{\|\mathbf{A}^{-1}\|_2} \frac{\|\mathbf{x}\|_2}{\|\mathbf{b}\|_2} \le 1,$$

shows that $\beta_k \le \nu_{k+1}$, and the definition of β_k implies

$$\frac{\|\mathbf{r}_k^{\mathrm{ME}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \le \kappa_2(\mathbf{A}) \, \beta_k, \tag{54}$$

which completes the proof. □

For our analysis in section 3.2 of the additional errors in SYMMLQ, we also need a slightly more general result, formulated in the next theorem.

THEOREM 3.3. *Let* $\mathbf{c} = \mathbf{A}\mathbf{y}$ *for some* $\mathbf{y}$. *Consider the best approximation* $\mathbf{y}_k^{\mathrm{ME}}$ *of* $\mathbf{y}$ *in* $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{b})$ *and the* $\mathbf{y}_k^{\mathrm{MR}} \in \mathcal{K}_k(\mathbf{A}; \mathbf{b})$ *for which* $\mathbf{A}\mathbf{y}_k^{\mathrm{MR}}$ *is the best approximation of* $\mathbf{c}$ *in* $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{b})$.

Then, with ν_k *as in* (48), *we have*

$$\frac{\|\mathbf{c} - \mathbf{A}\,\mathbf{y}_k^{\mathrm{ME}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \le \nu_{k+1} \, \kappa_2(\mathbf{A}) \, \mu_k, \quad \textit{where} \quad \mu_k \equiv \sup_{i \le k} \frac{\|\mathbf{c} - \mathbf{A}\,\mathbf{y}_i^{\mathrm{MR}}\|_2}{\|\mathbf{r}_i^{\mathrm{MR}}\|_2}. \tag{55}$$

Proof. The proof comes along the same lines as the proof of Theorem 3.2.

Replace the quantities $\mathbf{x}$ and $\mathbf{x}_k^{\mathrm{MR}}$ by $\mathbf{y}$ and $\mathbf{y}_k^{\mathrm{MR}}$. Since the $\mathbf{y}$ quantities fulfill the same orthogonality relations, (45) is valid also in the $\mathbf{y}$ quantities. This is also the case for the upper bound for $|\alpha_k| \, \|\mathbf{r}_k^{\mathrm{MR}}\|_2 = |(\mathbf{y}, \mathbf{r}_k^{\mathrm{MR}}/\|\mathbf{r}_k^{\mathrm{MR}}\|_2)|$. Hence, with $\mathbf{e}_j^{\mathrm{ME}} \equiv \mathbf{y} - \mathbf{y}_j^{\mathrm{ME}}$, we have

$$\frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} \le \frac{\|\mathbf{y} - \mathbf{y}_{k+1}^{\mathrm{MR}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} + \left(\frac{\|\mathbf{e}_{k-1}^{\mathrm{ME}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{y} - \mathbf{y}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \right)^2. \tag{56}$$

If we define $\widehat{\beta}_k \equiv \beta_k/\mu_k$, we find that

$$\widehat{\beta}_k^2 \le 1 + (\widehat{\beta}_{k-1} + 1)^2 \quad \text{and} \quad \widehat{\beta}_0 = \frac{1}{\mu_0 \|\mathbf{A}^{-1}\|_2} \frac{\|\mathbf{e}_0^{\mathrm{ME}}\|_2}{\|\mathbf{r}_0^{\mathrm{MR}}\|_2} \le 1.$$

Therefore, as in the proof of Theorem 3.2, $\widehat{\beta}_k \leq \nu_{k+1}$, which implies (55). ∎

For the relations between SYMMLQ and MINRES we have assumed exact arithmetic, that is, we have assumed an exact Lanczos process as well as an exact solve of the systems with L_k. However, we can exclude the influence of the Lanczos process by applying Theorem 3.2 right away to a system with a Lanczos matrix T_m and initial residual $\|\mathbf{r}_0\|_2 e_1$. In this setting, we have, for $k < m$, that [13, Proposition 1]

$$\|\mathbf{r}_k^{\mathrm{MR}}\|_2 = \|\mathbf{r}_0\|_2 \, \rho_k, \quad \text{where} \quad \rho_k \equiv |s_1 \cdot \ldots \cdot s_k|, \tag{57}$$

with s_j the sine in the jth Givens rotation for the QR decomposition of $\underline{T}_k$; ρ_k is the estimated reduction of the norms of the MINRES residuals. Note that (57) is also an immediate consequence of (50).

From relation (54) in combination with the fact that $\|\mathbf{r}_k^{\mathrm{ME}}\|_2 = \|\mathbf{r}_0\|_2 \, \|t_k\|_2$ (cf. (38), where, in this setting, $\mathbf{F}'_{k+1} = \mathbf{0}$), we conclude that

$$\|t_k\|_2 \leq \rho_k \, \kappa_2(T_m) \, \nu_{k+1} \quad \text{with} \quad \nu_k = k + \tfrac{1}{2}\ln(k), \tag{58}$$

for all $m > k$.

Note that inequality (58) is correct for any symmetric tridiagonal extension $\widetilde{T}_m$ of T_{k+1}: (58) holds with $\widetilde{T}_m$ instead of T_m. It has been shown in [5] that there is an extension $\widetilde{T}_m$ of which any eigenvalue is in a $\mathcal{O}(\mathbf{u}^{\frac{1}{4}})$-neighborhood of some eigenvalue of $\mathbf{A}$, and therefore, $\kappa_2(\widetilde{T}_m) \approx \kappa_2(\mathbf{A})$ in fairly good precision. This leads to our upper bound

$$\|t_k\|_2 \lesssim \rho_k \, \kappa_2(\mathbf{A}) \, \nu_{k+1} \quad \text{with} \quad \nu_k = k + \tfrac{1}{2}\ln(k). \tag{59}$$

In section 3.2, we will show that

$$\|\widehat{t}_k - t_k\|_2 \lesssim 5 \, \mathbf{u} \, \rho_k \, \kappa_2(\mathbf{A})^2 \, \left(\tfrac{1}{6}k^3 + \mathcal{O}(k^2 \ln k)\right). \tag{60}$$

The upper bound in (60) contains a square of the condition number. However, in the interesting situation where ρ_k decreases towards 0, the effect of the condition number squared will be annihilated eventually.

Remark 3.4. Except for the constants $k+\mathcal{O}(k)$ and $\frac{1}{6}k^3+\mathcal{O}(k^2 \ln k)$, the estimates (59) and (60), respectively, appear to be sharp (see Figure 8).

Although the maximal values of the ratio of $\|\widehat{t}_k - t_k\|_2/\rho_k$ in Figure 8 exhibit slowly growing behavior, the growth is not of order k^3. In the proof of (60) (cf. section 3.2), upper bounds as in (59) are used in a consecutive number of steps. In view of the irregular convergence of SYMMLQ, the upper bound (59) will be sharp for at most a few steps. By exploiting this observation, one can show that a growth of order k^2, or even less, will be more likely.

3.2. SYMMLQ recurrences. In this section we derive the upper bound (60).

Suppose that the jth recurrence for the γ_i's, with γ_i as defined in (33), is perturbed by a relatively small δ and all other recurrence relations are exact:

$$\delta = \ell_{jj}\widetilde{\gamma}_j + \ell_{jj-1}\gamma_{j-1} + \ell_{jj-2}\gamma_{j-2} \quad \text{with} \quad |\delta| \leq \mu \, \mathbf{u} \, |\ell_{jj}| \, |\gamma_j|. \tag{61}$$

The resulting perturbed quantities are labeled as $\widetilde{\ }$.

Then

$$\widetilde{t}_k - t_k = \delta \, M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_j. \tag{62}$$

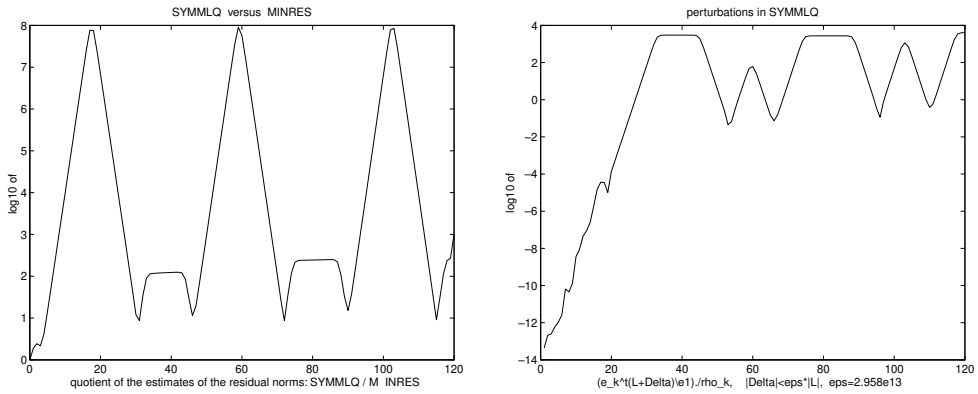

FIG. 8. *Results for the indefinite matrix with condition number* $3 \cdot 10^8$ *(as in the right pictures) of Figure* 4 *and Figure* 7. *The left picture shows* $\log_{10}$ *of the ratio* $\|\widehat{t}_k\|_2/\rho_k$ *of the estimated residual norm reduction* $\|\widehat{t}_k\|_2$ *of* SYMMLQ *and* ρ_k *for* MINRES (*cf.* (59)). *The right picture models* $\|\widehat{t}_k - t_k\|_2/\rho_k$ (*cf.* (60)) *with an artificial random pertubation* $\widetilde{\Delta}_L$, $|\widetilde{\Delta}_L| \gg |\Delta_L|$, *and* Δ_L *as in* (33)*: it shows the* $\log_{10}$ *of* $|e_k^T(L_k + \widetilde{\Delta}_L)^{-1}e_1/\rho_k - e_k^T(L_k + \Delta_L)^{-1}e_1/\rho_k|$, *where* $|\widetilde{\Delta}_L| \leq 3 \cdot 10^{-13}\,|L_k|$.

For $j = 1$, $\widetilde{t}_k - t_k$ is a multiple of the SYMMLQ residual for the T_m-system ($m > k$) and, as in the proof of inequality (59), Theorem 3.2 could be applied for estimating $\|\widetilde{t}_k - t_k\|_2$. For the situation where $j \neq 1$, Theorem 3.3 can be used.

To be more precise, we apply Theorem 3.3 with $\mathbf{v}_i = e_i$, $\mathbf{A} = T_m$, and $\mathbf{c} = e_j$. Then we have (in the notation as indicated in Theorem 3.3),

$$y_k^{\mathrm{ME}} = 0 \quad (k < j), \qquad \|e_j - T_m y_k^{\mathrm{ME}}\|_2 = \left\| M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_j \right\|_2 \quad (k \geq j), \tag{63}$$

and

$$y_k^{\mathrm{MR}} = 0 \quad (k+1 < j), \quad \|e_j - T_m\, y_k^{\mathrm{MR}}\|_2 = c_{j-1}\,\frac{\rho_k}{\rho_{j-1}} \leq \frac{\rho_k}{\rho_{j-1}} \quad (k+1 \geq j), \tag{64}$$

with c_{j-1} the cosine in the $(j-1)$th Givens rotation. Note that $\|e_j - T_m\, y_i^{\mathrm{MR}}\|_2/\rho_i \leq 1/\rho_{j-1}$ for all $i \leq k$. Therefore, by Theorem 3.3,

$$\left\| M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_j \right\|_2 \leq \kappa_2(T_m)\,\nu_{k+1}\,\frac{\rho_k}{\rho_{j-1}}. \tag{65}$$

For this specific situation, where $y_{j-1}^{\mathrm{ME}} = 0$, the estimate for β_k in the proof of Theorem 3.3 can be improved. If we take $\widehat{\beta}_k \equiv \rho_{j-1}\beta_k$, then we now have that $\widehat{\beta}_k^2 \leq 1 + (\widehat{\beta}_{k-1} + 1)^2$ and $\widehat{\beta}_{j-1} \leq 1$. This implies that $\rho_{j-1}\beta_k \leq \nu_{k-j+2}$. Therefore, the ν_{k+1} in (65) can be replaced by ν_{k-j+2}.

A combination of (62) with (65) gives (cf. (58) and following discussion)

$$\|\widetilde{t}_k - t_k\|_2 \leq \frac{|\delta|}{\rho_{j-1}}\,\rho_k\,\kappa_2(T_m)\,\nu_{k-j+2} \lesssim \frac{|\delta|}{\rho_{j-1}}\,\rho_k\,\kappa_2(\mathbf{A})\,\nu_{k-j+2}. \tag{66}$$

Using the definition of M_j and the recurrence relations for the γ_j, we can express t_{j-1} as

$$t_{j-1} = M_{j-1}\begin{bmatrix} \gamma_{j-2} \\ \gamma_{j-1} \end{bmatrix} = \begin{bmatrix} -\ell_{jj}\,\gamma_j \\ \ell_{j+1\,j-1}\,\gamma_{j-1} \end{bmatrix}.$$

Therefore, from (59), we have that

$$|\ell_{jj}|\,\frac{|\gamma_j|}{\rho_{j-1}} \le \frac{\|t_{j-1}\|_2}{\rho_{j-1}} \le \kappa_2(\mathbf{A})\,\nu_j. \tag{67}$$

Hence (cf. (61))

$$\frac{|\delta|}{\rho_{j-1}} \le \mu\,\mathbf{u}\,\kappa_2(\mathbf{A})\,\nu_j,$$

and, with (66), this gives

$$\|\widetilde{t}_k - t_k\|_2 \le \mu\,\mathbf{u}\,\rho_k\,\kappa_2(\mathbf{A})^2\,\nu_j\,\nu_{k-j+2}. \tag{68}$$

Because the recurrences are linear, the effect of a number of perturbations is the cumulation of the effects of single perturbations. If each recurrence relation is perturbed as in (61), then the estimate (60) appears as a cumulation of bounds as in (68). The vector $\widehat{t}_k$ in (60) represents the result of these successive perturbations due to finite precision arithmetic.

Finally, we will explain that the effect of rounding errors in solving $L^{-1}e_1$ can be described as the result of successively perturbed recurrence relations (61), with $\mu = 5$. First we note that the $\widetilde{\gamma}_k$'s resulting from the perturbation

$$\ell_{jj}\widetilde{\gamma}_j + \ell_{jj-1}\gamma_{j-1}(1+\mu\,\xi) + \ell_{jj-2}\gamma_{j-2} = 0 \quad \text{with} \quad |\xi| \le \mathbf{u}$$

are the same as those resulting from the perturbation

$$\ell_{j-1j-1}\widetilde{\gamma}_{j-1}(1+\mu\,\xi) + \ell_{j-1j-2}\gamma_{j-2} + \ell_{j-1j-3}\gamma_{j-3} = 0\,,$$

which means that a perturbation to the second term in the jth recurrence relation can also be interpreted as a similar perturbation to the first term in the $(j-1)$th recurrence relation.

Now we consider perturbations that are introduced in each recurrence relation due to finite precision arithmetic errors. Let $\widehat{\gamma}_j$ represent the actually computed γ_j, then

$$\widehat{\gamma}_j = -\frac{\ell_{jj-1}\widehat{\gamma}_{j-1}(1+\xi') + \ell_{jj-2}\widehat{\gamma}_{j-2}(1+\xi'')}{\ell_{jj}(1+2\,\xi)}, \quad \text{with} \quad |\xi|, |\xi'|, |\xi''| \le \mathbf{u},$$

and this can be rewritten, with different ξ and ξ', as

$$\ell_{jj}\widehat{\gamma}_j(1+3\,\xi) + \ell_{jj-1}\widehat{\gamma}_{j-1}(1+2\,\xi') + \ell_{jj-2}\widehat{\gamma}_{j-2} = 0, \quad \text{with} \quad |\xi|, |\xi'| \le \mathbf{u}.$$

Since the perturbation to the second term in this jth recurrence relation can be interpreted as a similar perturbation to the first term in the $(j-1)$th recurrence relation (which was already perturbed with a factor $(1+3\xi)$), we have that the computed $\widehat{\gamma}_j$ can be interpreted as the result of perturbing each leading term with a factor $(1+5\xi)$.

4. Discussion and conclusions. In Krylov subspace methods there are two main effects of floating point finite precision arithmetic errors. One effect is that the generated basis for the Krylov subspace deviates from the exact one. This may lead to a loss of orthogonality of the Lanczos basis vectors, but the main effect on the

iterative solution process is a delay in convergence rather than misconvergence. In fact, what happens is that we try to find an approximated solution in a subspace that is not as optimal, with respect to its dimension, as it could have been.

The other effect is that the determination of the approximation itself is perturbed with rounding errors, and this is, in our view, a serious point of concern; it has been the main theme of this study. In our study we have restricted ourselves to symmetric indefinite linear systems $\mathbf{Ax} = \mathbf{b}$. Before we review our main results, it should be noted that we should expect upper bounds for relative errors in approximations for $\mathbf{x}$ that contain at least the condition number of $\mathbf{A}$, simply because we can in general not compute $\mathbf{Ax}_k$ exactly. We have studied the effects of perturbations to the computed solution through their effect on the residual, because the residual (or its norm) is often the only information that we get from the process. This residual information is often obtained in a cheap way from some update procedure, and it is not uncommon that the updated residual may take values far smaller than machine precision (relative to the initial residual). Our analysis shows that there are limits on the reduction of the true residual because of errors in the approximated solution. For GMRES, this observation has also been made in [3].

In view of the fact that we may expect at least a linear factor $\kappa_2(\mathbf{A})$, when working with Euclidean norms, GMRES* (section 2.2) and SYMMLQ (section 3) lead to acceptable approximate solutions. When these methods converge, then the relative error in the approximate solution is, apart from modest factors, bounded by $\mathbf{u}\,\kappa_2(\mathbf{A})$. SYMMLQ is attractive since it minimizes the norm of the error, but it does so with respect to $\mathbf{A}$ times the Krylov subspace, which may lead to a delay in convergence with respect to GMRES* (or MINRES), by a number of iterations that is necessary to gain a reduction by $\kappa_2(\mathbf{A})$ in the residual; see Theorem 3.2 (also Figure 8). For ill-conditioned systems, this may be considerable.

As has been pointed out in [11], the conjugate gradient iterates can be constructed with little effort from SYMMLQ information if they exist. For indefinite systems the conjugate gradient iterates are well defined for at least every other iteration step, and they can be used to terminate the iteration if this is advantageous. However, the conjugate gradient process features no minimization property (in contrast to the positive definite case) when the matrix is indefinite, and so there is no guarantee that any of these iterates will be sufficiently close to the desired solution before SYMMLQ converges.

For indefinite symmetric systems we see that MINRES may lead to large perturbation errors: for MINRES the upper bound contains a factor $\kappa_2(\mathbf{A})^2$ (section 2.3). This means that if the condition number is large, then the methods of choice are GMRES or SYMMLQ. Note that for the symmetric case, GMRES can be based on the three-term recurrence relation, which means that the only drawback is the necessity to store all the Lanczos vectors. If storage is at a premium, then SYMMLQ is the method of choice.

If the given system is well conditioned, and if we are not interested in very accurate solutions, then MINRES may be an attractive choice.

Of course, one may combine any of the discussed methods with a variation on iterative refinement: after stopping the iteration at some approximation $\mathbf{x}_k$, we compute the residual $\mathbf{r}(\mathbf{x}_k) = \mathbf{b} - \mathbf{Ax}_k$, if possible in higher precision, and we continue to solve $\mathbf{Az} = \mathbf{r}(\mathbf{x}_k)$. The solution $\mathbf{z}_j$ of this system is used to correct $\mathbf{x}_k$: $\mathbf{x}_{\text{appr}} = \mathbf{x}_k + \mathbf{z}_j$. The procedure could be repeated, and eventually this leads to approximations for $\mathbf{x}$ so that the relative error in the residual is in the order of machine precision (for more details

on this, see [14]). However, if we would use MINRES, then, after restart, we have to carry out at least a number of iterations for the reduction by a factor equal to the condition number, in order to arrive at something of the same quality as GMRES*, which may make the method much less effective than GMRES*. For situations where $\kappa_2(\mathbf{A}) \geq 1/\sqrt{\mathbf{u}}$, MINRES may even be incapable of getting at a sufficient reduction for the iterative refinement procedure to converge.

It is common practice among numerical analysts to test the convergence behavior of Krylov subspace solvers for symmetric systems with well-chosen diagonal matrices. This often gives quite a good impression of what to expect for nondiagonal matrices with the same spectrum. However, as we have shown in our section 2.5, for MINRES this may lead to a too optimistic picture, since floating point error perturbations with MINRES for a diagonal matrix lead to errors in the residual (and the approximated solution) that are a factor $\kappa_2(\mathbf{A})$ smaller than for nondiagonal matrices.

Appendix.

LEMMA A.1. *If, for a matrix* $\mathbf{C}$, $n_C = \min(n_c, n_r)$ *with* n_c *the maximum number of nonzeros per column and* n_r *the maximum number of nonzeros per row, then*

$$\| \, |\mathbf{C}| \, \|_2 \leq \sqrt{n_C} \, \|\mathbf{C}\|_2. \tag{69}$$

Proof. We prove the lemma with respect to columns; the row variant follows from the fact that $\|\mathbf{B}^T\|_2 = \|\mathbf{B}\|_2$ for any matrix $\mathbf{B}$.

Since $\| \, |\mathbf{C}| \, \|_2^2 \leq n_C \max_j \left(\sum_i |c_{ij}|^2 \right)$ (see [15, Theorem 4.2]), we have

$$\| \, |\mathbf{C}| \, \|_2^2 \leq n_C \max_j \|\mathbf{C}e_j\|_2^2 \leq n_C \, \|\mathbf{C}\|_2^2. \qquad \square$$

Acknowledgments. The writing of this paper has been an exercise in modesty. We have to admit that it was only with extensive help of three anonymous referees, who invested embarrassing amounts of time, that the present version of this manuscript could be written. Somehow we seem to have developed a certain blindness for inaccuracies in the often complicated formulas, in the course of expressing our ideas. We are extremely thankful for the patience of the referees and for their detailed advice.

REFERENCES

[1] P. N. BROWN, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 58–78.
[2] A. M. BRUASET, *A Survey of Preconditioned Iterative Methods*, Longman Scientific and Technical, Harlow, UK, 1995.
[3] J. DRKOŠOVÁ, A. GREENBAUM, M. ROZLOŽNÍK, AND Z. STRAKOŠ, *Numerical stability of GMRES*, BIT, 35 (1995), pp. 309–330.
[4] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The John Hopkins University Press, Baltimore, London, 1996.
[5] A. GREENBAUM, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.
[6] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, Frontiers in Applied Mathematics 17, SIAM, Philadelphia, PA, 1997.
[7] A. GREENBAUM AND Z. STRAKOŠ, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.
[8] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[9] C. C. PAIGE, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl., 18 (1976), pp. 341–349.
[10] C. C. PAIGE, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.

[11] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[12] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall Ser. Comput. Math., Prentice-Hall, Englewood Cliffs, NJ, 1980.
[13] 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.
[14] K. Turner and H. F. Walker, *Efficient high accuracy solutions with GMRES(m)*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 815–825.
[15] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969/1970), pp. 14–23.
[16] H. A. van der Vorst and C. Vuik, *The superlinear convergence behaviour of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.