Differences in the effects of rounding errors in Krylov solvers for symmetric indefinite linear systems

Gerard L.G. Sleijpen* Henk A. Van der Vorst* Jan Modersitzki†

April 6, 1999

Abstract

The 3-term Lanczos process leads, for a symmetric matrix, 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 the reduced system in one way or another. This leads to well-known methods: MINRES, GMRES, and SYMMLQ. 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, that are not corrected by continuing the iteration process.
Our findings are supported and illustrated by numerical examples.

1 Introduction

We will consider iterative methods for the construction of approximate solutions, starting with $\mathbf{x}_0 = \mathbf{0}$,[1] for the linear system $\mathbf{Ax} = \mathbf{b}$, with $\mathbf{A}$ an n by n symmetric matrix, in the k-dimensional Krylov subspace

$$\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0) \equiv \{\mathbf{r}_0, \mathbf{Ar}_0, ..., \mathbf{A}^{k-1}\mathbf{r}_0\},$$

with $\mathbf{r}_0 \equiv \mathbf{b} - \mathbf{Ax}_0 = \mathbf{b}$.
With the standard 3-term Lanczos process, we generate an orthonormal basis $\mathbf{v}_1$, ..., $\mathbf{v}_k$, with $\mathbf{v}_1 \equiv \mathbf{r}_0/\|\mathbf{r}_0\|_2$, for $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$. The 3-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 $\mathbf{v}_1$, ..., $\mathbf{v}_j$, and $\underline{T}_k$ is a $k+1$ by k tridiagonal matrix.

*Mathematical Institute, Utrecht University, Budapestlaan 6, Utrecht, the Netherlands.
†Institute of Mathematics, Medical University of Lübeck, Wallstraße 40, 23560 Lübeck, Germany.
E-mail: sleijpen@math.uu.nl, vorst@math.uu.nl, modersitzki@informatik.mu-luebeck.de

[1] This assumption does not mean a loss of generality, since the case $\mathbf{x}_0 \neq \mathbf{0}$ can be reduced to this by a simple shift

Paige [11] 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

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

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, 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 5.1), 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,[2] 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 $\mathbf{A}\mathbf{x} = \mathbf{b}$ 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 differences between the methods are due to a different way of solution of the reduced system and to differences in the backtransformation 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 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 analysed by Paige [11, 12]. It has been proven by Greenbaum and Strakos [8], 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, this type of error analysis is on a worst case scenario, and as a consequence the error bounds are pessimistic. In particular, the error bounds cannot very well be used to explain differences between these methods, so as we observe them in practical situations.

In this paper, we propose a different way of analysing 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:
– When and why is MINRES less accurate than SYMMLQ? This question was already posed in the original publication [14], but the answer in [14, p.625] is largely speculative.
– Is MINRES suspect for ill-conditioned systems, because of the minimal residual approach (see [14, p.619])? Although hints are given for the reasons of inaccuracies in MINRES, for MINRES, it is also stated in [14, p. 625] that it is not as accurate as SYMMLQ for the reason

[2] GMRES has been designed in combination with Arnoldi's method for unsymmetric systems, but for symmetric systems Arnoldi's method and Lanczos' method lead, in exact arithmetic, to the same relation (1)

that the minimal residual method is suspect. In [3, p. 43] an explicit relation is suggested between MINRES and working with $\mathbf{A}^2$, and it is argued that for that reason 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 a mistake).
– Why and when is SYMMLQ slower than for instance MINRES or GMRES?
– 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 [14, 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** [14]: 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 the Appendix.

2. **GMRES** [16]: 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 orthogonalisation 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 [7, 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 stick to the name GMRES rather than to introduce a new and possibly confusing acronym. Due to the way of solution in GMRES, all the basis vectors $\mathbf{v}_j$ have to be stored, also when $\mathbf{A}$ is symmetric.

3. **SYMMLQ** [14]: determine $\mathbf{x}_k = \mathbf{A}\mathbf{V}_k y_k$, 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 $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$. Conjugate Gradient approximations can, if they exist, be computed with little effort from the SYMMLQ information. In the SYMMLQ implementation suggested in [14] 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 the Appendix.

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

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 $\|\mathbf{r}_0\|_2$), 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.

Notations: Quantities associated with n dimensional spaces will be represented in bold face, 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 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 $A = (a_{ij})$.

2 Differences in round-off error behaviour between 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. Without loss of generality, we may assume that $\mathbf{x}_0 = \mathbf{0}$, so that $\mathbf{r}_0 = \mathbf{b}$.

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 for minimizing $\|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2$, y_k must be the linear least squares solution of the $k+1$ by k overdetermined system

$$\underline{T}_k\, y_k = \|\mathbf{b}\|_2 e_1. \tag{5}$$

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

$$\underline{T}_k = \underline{Q}_k R_k, \tag{6}$$

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 (6), y_k can be solved from

$$R_k\, y_k = z_k \equiv \|\mathbf{b}\|_2\, \underline{Q}_k^T e_1, \tag{7}$$

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

$$\begin{aligned} \mathbf{x}_k &= \mathbf{V}_k (R_k^{-1} \underline{Q}_k^T \|\mathbf{b}\|_2 e_1) \\ &= \mathbf{V}_k (R_k^{-1} z_k). \end{aligned} \tag{8}$$

The parentheses have been included in order to indicate the order of computation. In the original publication [16], GMRES was proposed for unsymmetric $\mathbf{A}$, in combination with Arnoldi's method for an orthonormal basis for the Krylov subspace. However, when $\mathbf{A}$ is symmetric then Arnoldi's method is equivalent to Lanczos' method, so that (8) describes GMRES for symmetric $\mathbf{A}$. 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 (8) can be reordered as

$$\begin{aligned} \mathbf{x}_k &= \left(\mathbf{V}_k R_k^{-1}\right) z_k \\ &\equiv \mathbf{W}_k\, z_k. \end{aligned} \tag{9}$$

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 dictated to compute these items 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* items 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 (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

- for GMRES: $\mathbf{x}_k = \mathbf{V}_k \left(R_k^{-1} z_k\right)$,
- for MINRES: $\mathbf{x}_k = \left(\mathbf{V}_k R_k^{-1}\right) z_k$.

Of course, we could have tried to get upper bounds for all errors made in each process, but this would most likely not reveal the differences between the two methods. If we want to study the differences between the two methods then we have to concentrate on the two generating formulas.

2.2 Error analysis for GMRES

In order to understand the difference between GMRES and MINRES, we have to study the computational errors in $\mathbf{V}_k \left(R_k^{-1} z_k\right)$. We will indicate actual computation in floating point finite precision arithmetic by fl, and the result will be denote by a $\widehat{\ }$. Then, according to [5, p. 89], in floating point arithmetic the computed solution $\widehat{y}_k = fl(R_k^{-1} z_k)$ satisfies

$$(R_k + \Delta_R)\widehat{y}_k = z_k, \quad \text{with} \quad |\Delta_R| \leq 3\,\mathbf{u}\,|R_k| + \mathcal{O}(\mathbf{u}^2). \tag{10}$$

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 $\mathbf{u}$

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

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 make also 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 [10, p.76], we obtain

$$\widehat{\mathbf{x}}_k = \mathbf{V}_k \widehat{y}_k + \Delta_2, \tag{11}$$

with $|\Delta_2| \leq k\,\mathbf{u}\,|\mathbf{V}_k|\,|y_k| + \mathcal{O}(\mathbf{u}^2)$. Hence, the error $\Delta\mathbf{x}_k = \widehat{\mathbf{x}}_k - \mathbf{x}_k$, that can be attributed to differences between MINRES and GMRES, has two components

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

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 differences between MINRES and GMRES (ignoring $\mathcal{O}(\mathbf{u}^2)$ terms):

$$\begin{aligned}
\Delta\mathbf{r}_k &= -\mathbf{A}\mathbf{V}_k\Delta_1 - \mathbf{A}\Delta_2 \\
&= \mathbf{A}\mathbf{V}_k R_k^{-1}\Delta_R R_k^{-1} z_k - \mathbf{A}\Delta_2 \\
&= \mathbf{V}_{k+1}\underline{T}_k R_k^{-1}\Delta_R R_k^{-1} z_k - \mathbf{A}\Delta_2 \\
&= \mathbf{V}_{k+1}\underline{Q}_k\Delta_R R_k^{-1} z_k - \mathbf{A}\Delta_2. \qquad (12)
\end{aligned}$$

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 a 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.

Using the bound in (10) and the bound for Δ_2, we get (skipping higher order terms in $\mathbf{u}$)

$$\begin{aligned}
\|\Delta\mathbf{r}_k\|_2 &\leq \|\mathbf{V}_{k+1}\underline{Q}_k\|_2\, 3\,\mathbf{u}\,\|\,|R_k|\,\|_2\,\|R_k^{-1}\|_2\,\|\mathbf{b}\|_2 + k\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\,|\mathbf{V}_k|\,\|_2\|y_k\|_2 \\
&\leq 3\,\mathbf{u}\,\|\mathbf{V}_{k+1}\|_2\,\|\,|R_k|\,\|_2\,\|R_k^{-1}\|_2\,\|\mathbf{b}\|_2 + k\sqrt{k}\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\mathbf{V}_k\|_2\|y_k\|_2 \\
&\leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_{k+1}\|_2\,\kappa_2(R_k)\,\|\mathbf{b}\|_2 + k\sqrt{k}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\|\mathbf{A}\|_2\,\|R_k^{-1}\|_2\,\|\mathbf{b}\|_2. \qquad (13)
\end{aligned}$$

Here we have used that $\|\,|R_k|\,\|_2 \leq \sqrt{3}\,\|R_k\|_2$, and $\|\,|\mathbf{V}_k|\,\|_2 \leq \sqrt{k}\,\|\mathbf{V}_k\|_2$, (which follows from [21, Th. 4.2]; see Lemma 5.1 for details). The factor κ_2 denotes the condition number with respect to the Euclidean norm.[3]

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

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

which is, because of the local orthogonality of the $\mathbf{v}_j$, a crude overestimate. According to [15, p. 267 (bottom)], it may be more realistic to replace this factor $\sqrt{k+1}$ by a factor $\sqrt{m}$, where m denotes the number of times that a Ritz value of T_k has converged to an eigenvalue of $\mathbf{A}$. When solving a linear system, this value of m is usually very modest, 2 or 3 say.

Finally, we note that

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

It has been shown in [6] that the matrix $\underline{T}_k$ that has been obtained in finite precision arithmetic, may 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 $\mathbf{A}$.[4] With $\widetilde{\mathbf{V}}_k$ we denote the orthogonal matrix that generates $\underline{T}_k$, in exact arithmetic, from $\widetilde{\mathbf{A}}$. Hence,

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

[3] We also have used that the computed $\underline{Q}_k$ are orthogonal matrices, with errors in the order of $\mathbf{u}$, i.e., $\|\underline{Q}_k^T\underline{Q}_k - I_k\|_2 = \mathcal{O}(\mathbf{u})$. These $\mathcal{O}(\mathbf{u})$-errors lead to $\mathcal{O}(\mathbf{u})^2$-errors in (13).

[4] 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 [7, Sect.4.4.2].

so that

$$\sigma_{\min}(R_k^T R_k) \geq \sigma_{\min}(\widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}),$$

and

$$\sigma_{\max}(R_k^T R_k) \leq \sigma_{\max}(\widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}),$$

which implies $\kappa_2(R_k) \leq \kappa_2(\widetilde{\mathbf{A}}) = \kappa_2(\mathbf{A})$ (ignoring errors proportional to mild orders of $\mathbf{u}$). This finally results in the upper bound for the error in the residual due to the difference between GMRES and MINRES:

$$\frac{\|\Delta \mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq (3\sqrt{3} + k\sqrt{k})\, \mathbf{u}\, \|\mathbf{V}_{k+1}\|_2\, \kappa_2(\mathbf{A}). \tag{14}$$

Note that, even if there were only rounding errors in the matrix-vector multiplication, then the perturbation $\Delta\mathbf{x}$ to $\mathbf{A}^{-1}\mathbf{b}$ would have been (in norm) in the order of $\mathbf{u}\,\|\mathbf{A}^{-1}\|_2\,\|\mathbf{b}\|_2$. This corresponds to an error $\|\mathbf{A}\Delta\mathbf{x}\|_2 \sim \mathbf{u}\,\kappa_2(\mathbf{A})\|\mathbf{b}\|_2$ in the residual. Therefore, the stability of GMRES cannot essentially be improved.

2.3 Error analysis for MINRES

The differences in finite precision between MINRES and GMRES are reflected by $\left(\mathbf{V}_k R_k^{-1}\right) z_k$. 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:

$$\widehat{w}_{j,:}(R_k + \Delta_{R_j}) = v_{j,:}, \tag{15}$$

with

$$|\Delta_{R_j}| \leq 3\,\mathbf{u}\,|R_k| + \mathcal{O}(\mathbf{u}^2). \tag{16}$$

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 $j = 1, ..., k$, we obtain

$$\widehat{\mathbf{W}}_k = (\mathbf{V}_k + \Delta_W) R_k^{-1}, \tag{17}$$

with

$$|\Delta_W| \leq 3\,\mathbf{u}\,|\widehat{\mathbf{W}}_k|\,|R_k| + \mathcal{O}(\mathbf{u}^2) \tag{18}$$

We may replace $\widehat{\mathbf{W}}_k$ by $\mathbf{W}_k = \mathbf{V}_k R_k^{-1}$ in (18), because this leads only to $\mathcal{O}(\mathbf{u}^2)$ errors.

Finally, we make errors in the computation of $\mathbf{x}_k$ because of finite precision errors in the multiplication of $\widehat{\mathbf{W}}_k$ with z_k:

$$\widehat{\mathbf{x}}_k = \widehat{\mathbf{W}}_k z_k + \Delta_3, \tag{19}$$

with $|\Delta_3| \leq k\,\mathbf{u}\,|\mathbf{W}_k|\,|z_k| + \mathcal{O}(\mathbf{u}^2)$. The errors made in $\widehat{\mathbf{W}}_k$ and the error term Δ_3, are the only errors that can be held responsible for the difference between MINRES and GMRES. Added together, they lead to the $\Delta\mathbf{x}_k$ related to MINRES:

$$\Delta\mathbf{x}_k = -\Delta_W R_k^{-1} z_k + \Delta_3, \tag{20}$$

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

$$\Delta \mathbf{r}_k = \mathbf{A}\Delta_W R_k^{-1} z_k - \mathbf{A}\Delta_3.$$

If we use the bound (18) for Δ_W, and use for other quantities bounds similar as for GMRES, then we obtain

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

Here we have also used the fact that

$$\|\,|\mathbf{V}_k R_k^{-1}|\,\|_2 \leq \|\mathbf{V}_k R_k^{-1}\|_F \leq \|\mathbf{V}_k\|_F\,\|R_k^{-1}\|_2, \tag{21}$$

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

This finally results in the following upper bound for the error contribution in the residual due to the differences in the implementation between MINRES and GMRES:

$$\frac{\|\Delta \mathbf{r}_k\|_2}{\|\mathbf{b}\|_2} \leq 3\sqrt{3k}\,\mathbf{u}\,\kappa_2(\mathbf{A})^2 + k\sqrt{k}\,\mathbf{u}\,\kappa_2(\mathbf{A}). \tag{22}$$

We see that the different implementation for MINRES leads to a relative error in the residual that is proportional to the squared condition number of $\mathbf{A}$, whereas for the GMRES implementation the difference led to a relative error proportional to the condition number only. This means that if we plot the residuals for MINRES and GMRES then we may expect to see differences, more specifically, the difference between the computed residuals for the two methods may be expected to be in the order of the square of the condition number. As soon as the computed residual of GMRES gets below $\mathbf{u}\,\kappa_2(\mathbf{A})^2\,\|\mathbf{b}\|_2$, then this difference may be visible.

2.4 Discussion

In Fig. 1, we have plotted the residuals obtained for GMRES and MINRES. Our analysis suggests that there may be a difference between both in 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. 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 $\mathbf{A}$, specially 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 computed efficiently as a product $\rho_k \equiv |s_1 \cdot \ldots \cdot s_k|$ of the sines s_k of the Givens rotations. In MINRES (as well as GMRES) this value ρ_k is used to measure the reduction of the residual norm: in practical computations, a residual norm is not often computed explicitly

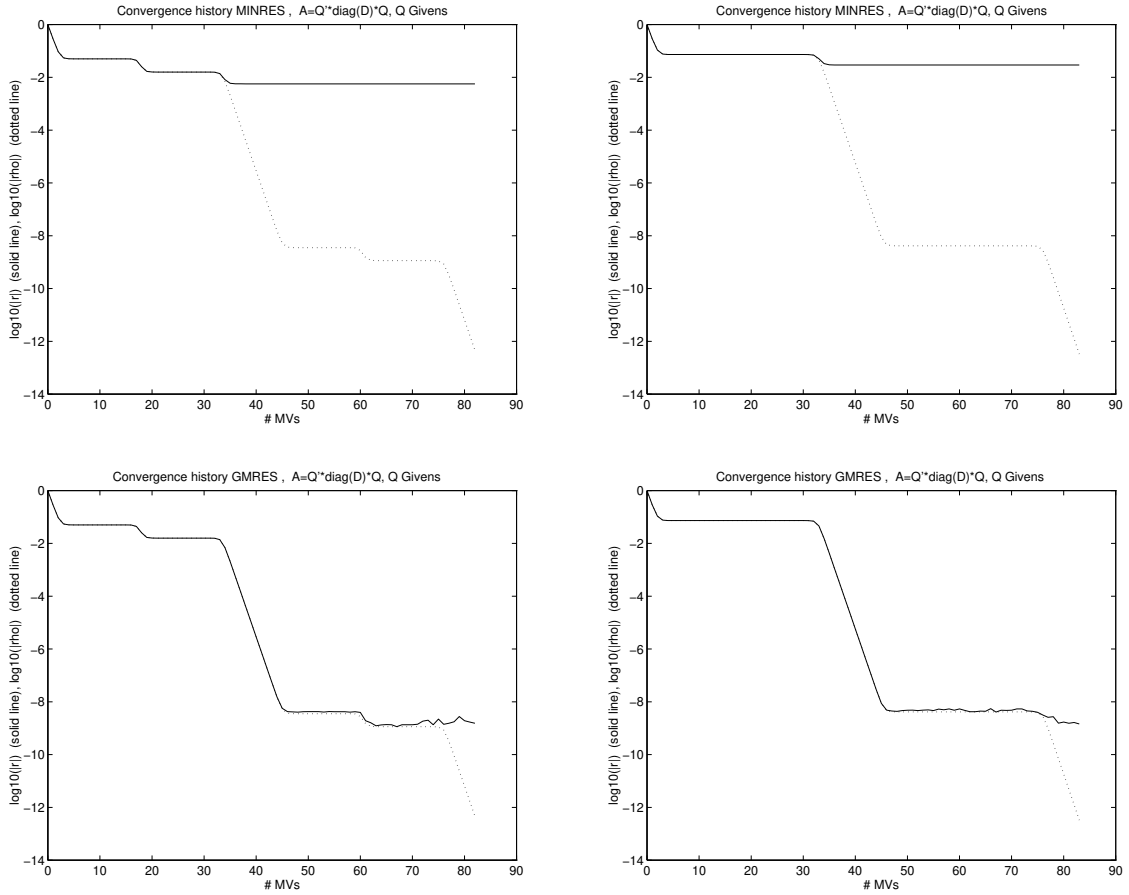

FIGURE 1. MINRES (*top*) *and* GMRES (*bottom*): *solid line* (—) $\log_{10}$ *of* $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2/\|\mathbf{r}_0\|_2$; *dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* ρ_k. *The pictures show the results for a positive definite system* (*the left pictures*) *and for a non-definite system* (*the right pictures*). *For both examples* $\kappa_2(\mathbf{A}) = 3\,10^8$. *To be more specific*: *at the left* $\mathbf{A} = \mathbf{GDG}'$ *with* $\mathbf{D}$ *diagonal,* $\mathbf{D} \equiv diag(10^{-8}, 2\,10^{-8}, 2 : h : 3)$, $h = 1/789$, *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* $\mathbf{D} \equiv diag(-10^{-8}, 10^{-8}, 2 : h : 3)$, $h = 1/389$, *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* $\mathbf{u} = 1.1\,10^{-16}$.

as $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2$, with $\widehat{\mathbf{x}}_k$ the kth floating point approximate. Therefore, it is of interest to know how much the computed ρ_k may differ from the exact residual norm reduction. The errors made in the computation of ρ_k are of order $\mathbf{u}$ and can be neglected. Since the computation of ρ_k and of $\widehat{\mathbf{x}}_k$ are based on the same inexact Lanczos process, (22) implies that

$$\left| \rho_k - \frac{\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2}{\|\mathbf{r}_0\|_2} \right| \leq 3\sqrt{3k}\,\mathbf{u}\,\kappa_2(\mathbf{A})^2 + k\sqrt{k}\,\mathbf{u}\,\kappa_2(\mathbf{A}).$$

The situation for GMRES is much better: the difference between ρ_k and the true residual reduction for GMRES can be bounded by the quantity in the right hand side of (14). In fact, as observed at the end of §2.2, except for the moderate constant $(3\sqrt{3} + k\sqrt{k})\,\|\mathbf{V}_{k+1}\|_2$, this is about the most accurate computation that can be expected.

2.5 Diagonal matrices

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

Rotating the matrix from diagonal to non-diagonal (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: experimental results (cf. Fig. 2) indicate that the errors in the MINRES residuals for diagonal matrices are of order $\mathbf{u}\,\kappa_2(\mathbf{A})$, as for GMRES. This can be understood as follows.

If we neglect $\mathcal{O}(\mathbf{u}^2)$ terms, then, according to (15), the error, due to the inversion of R_k, in the jth coordinate of the MINRES-$\mathbf{x}_k$ is given by

$$(\Delta\mathbf{x}_k)_j = (\widehat{w}_{j,:} - w_{j,:})z_k + (\Delta_2)_j = -v_{j,:}\,R_k^{-1}\Delta_{R_j}R_k^{-1}z_k + (\Delta_2)_j.$$

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 (6))

$$\begin{aligned}(\Delta\mathbf{r}_k)_j &= \lambda_j v_{j,:}\,R_k^{-1}\Delta_{R_j}R_k^{-1}z_k - \lambda_j(\Delta_2)_j = \mathbf{e}_j^T\mathbf{A}\mathbf{V}_kR_k^{-1}\Delta_{R_j}R_k^{-1}z_k - \lambda_j(\Delta_2)_j \\ &= \mathbf{e}_j^T\mathbf{V}_{k+1}\underline{Q}_k\Delta_{R_j}R_k^{-1}z_k - \lambda_j(\Delta_2)_j.\end{aligned} \tag{23}$$

Therefore, in view of (16), and including the error term for the multiplication with $\widehat{\mathbf{W}}_k$ (cf. (19)), we have for MINRES applied to a diagonal matrix:

$$\frac{\|\Delta\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq (3\sqrt{3} + k\sqrt{k})\;\mathbf{u}\,\|\mathbf{V}_{k+1}\|_2\,\kappa_2(\mathbf{A}),$$

which is the same upper bound as for the errors in the GMRES residuals in (14).

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, (23) 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 the bound (22).

2.6 The errors in the approximations

In exact arithmetic we have that $\|\mathbf{x}_k\|_2 = \|\mathbf{V}_kR_k^{-1}z\|_2 = \|R_k^{-1}z\|_2$. Assuming that, in finite precision, this also gives about the right order of magnitude, then the errors related to differences between MINRES and GMRES, for the approximate solutions in (11) and (20) can be bounded by essentially the same upper bound:

$$\frac{\|\Delta\mathbf{x}_k\|_2}{\|\widehat{\mathbf{x}}_k\|_2} \lesssim (3\sqrt{3} + k\sqrt{k})\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(R_k) \leq (3\sqrt{3} + k\sqrt{k})\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(\mathbf{A}). \tag{24}$$

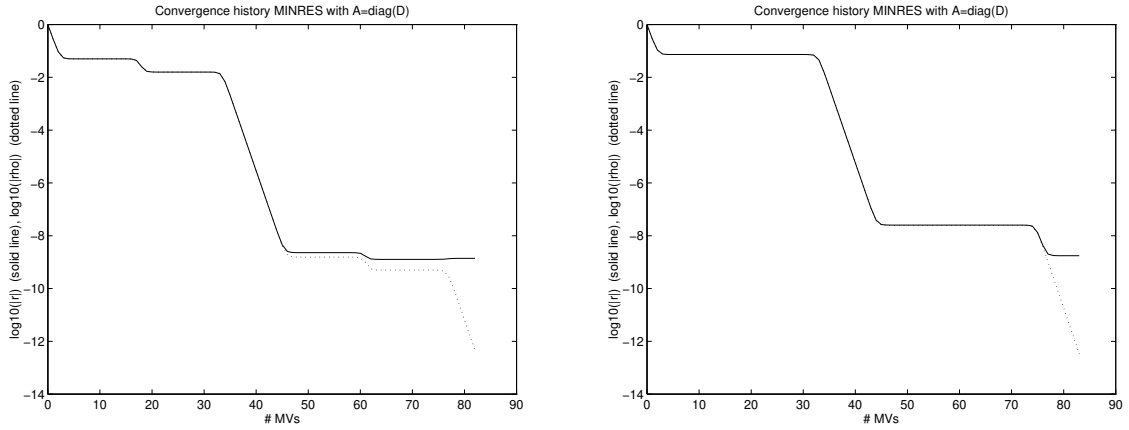

FIGURE 2. MINRES: *solid line* (—) $\log_{10}$ *of* $\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_k\|_2/\|\mathbf{r}_0\|_2$; *dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* ρ_k. *The pictures show the results for a positive definite diagonal system (the left picture) and for a non-definite diagonal system (the right picture). Except for the Givens rotation, the matrices in these examples are equal to the matrices of the examples in* Fig. 1*: here* $\mathbf{G} = \mathbf{I}$.

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 for numerical experiments, we think that this can be explained as follows. The error in the GMRES approximation has mainly large components in the direction of the small singular vectors of $\mathbf{A}$. These components are relatively reduced by 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 singular values of $\mathbf{A}$ and multiplication with $\mathbf{A}$ will make error components associated with larger singular values more dominating in the residual.

We will support our viewpoint by a numerical example. The results in FIG. 3 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 singular vectors with smallest singular value. 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 easier visible.

The convergence history of GMRES and of MINRES (not shown here) for this example with $\mathbf{x}_0 = \mathbf{0}$, is comparable to the ones in the left pictures of FIG. 1, 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).

FIG. 3 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 singular vectors with smallest singular value, 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

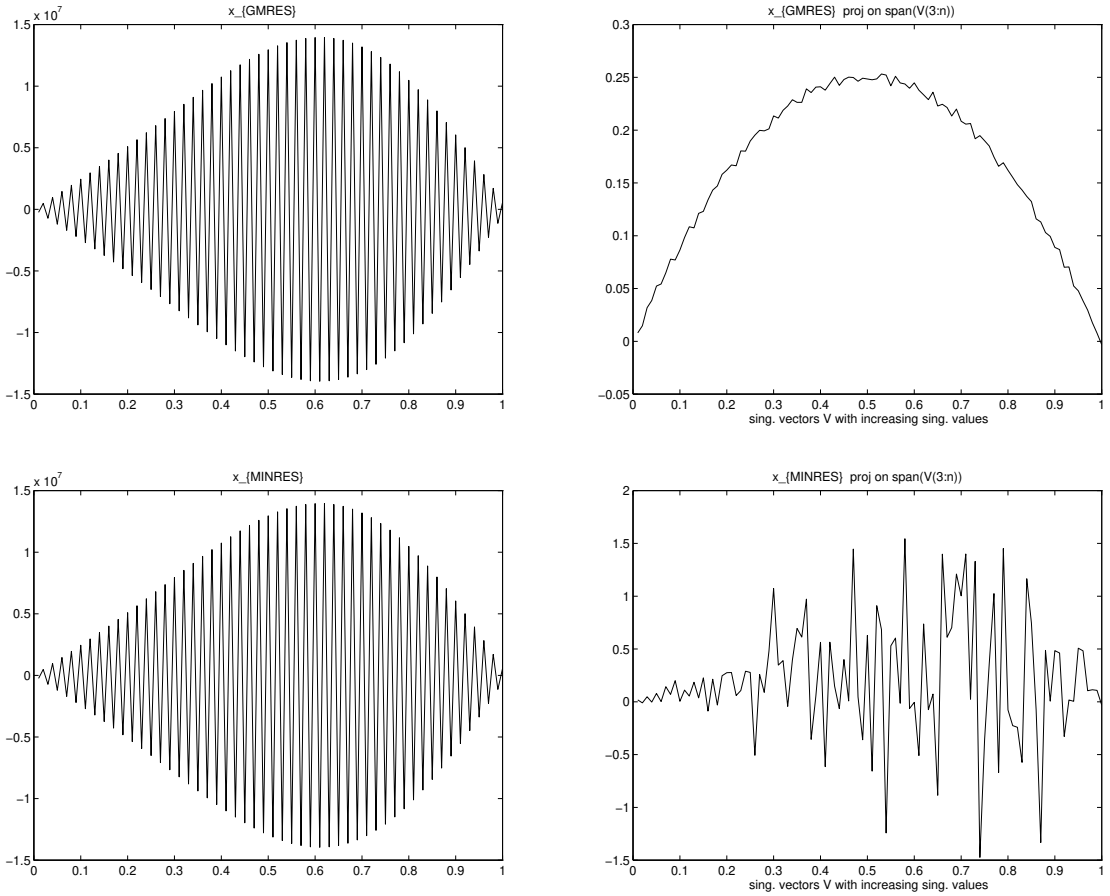

FIGURE 3. *The pictures show the solution* $\mathbf{x}$ *of* $\mathbf{Ax} = \mathbf{b}$*, computed with* 80 *steps of* GMRES (*top pictures*) *and of* MINRES (*bottom pictures*). *The* i*th coordinate of* $\mathbf{x}_k$ (*along the vertical axis*) *is plotted against* $\frac{i}{n}$ (*along the horizontal axis*). $\mathbf{A} = \mathbf{Q}^*\mathbf{D}\mathbf{Q}$ *with* $\mathbf{D} = diag(10^{-10}, 2\,10^{-10}, 2 : h : 3)$, $h = 1/97$ *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 singular vectors with smallest singular value, while the left pictures show the complete* $\mathbf{x}_k$.

mainly in the direction of the two small singular vectors. In contrast, the irregularities in the MINRES curve (bottom-right) are almost purely 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 $L\widetilde{Q}$ decomposition of $\underline{T}_k^T$, and this is equivalent to the transpose of the $\underline{Q}_k R_k$ decomposition of $\underline{T}_k$ (see (6)), which is constructed for GMRES and MINRES:

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

This leads to

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

from which the basic generating formula for SYMMLQ is obtained:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{x}_0 + \mathbf{A}\mathbf{V}_k R_k^{-1} R_k^{-T} \|\mathbf{r}_0\|_2 \, e_1 \\ &= \mathbf{x}_0 + \mathbf{V}_{k+1} \underline{T}_k R_k^{-1} R_k^{-T} \|\mathbf{r}_0\|_2 \, e_1 \\ &= \mathbf{x}_0 + (\mathbf{V}_{k+1} \underline{Q}_k) \, (L_k^{-1} \|\mathbf{r}_0\|_2 \, e_1), \end{aligned} \tag{25}$$

with $L_k \equiv R_k^T$. We will further assume that $\mathbf{x}_0 = \mathbf{0}$.

The actual implementation of SYMMLQ [14] is based on an update procedure for $\mathbf{V}_{k+1}\underline{Q}_k$, and on a three term recurrence relation for $\|\mathbf{r}_0\|_2 \, L_k^{-1} e_1$.
Note that SYMMLQ can be carried out with exactly the same computed values for $\mathbf{V}_{k+1}$, $\underline{Q}_k$, R_k, and $\mathbf{r}_0$, as for GMRES and MINRES. In fact, there is no good reason for using different values for each of the algorithms. Therefore, differences because of round-off, between the three methods, must be attributed to the additional rounding errors made in the evaluation of the right-hand side of (25).

The largest factor in the upper bound for these additional rounding errors in the construction of the SYMMLQ approximation $\mathbf{x}_k$ is caused by the inversion of L_k. The multiplication with $\underline{Q}_k$ and the assembly of $\mathbf{x}_k$, leads to a factor $k\sqrt{k}$ in the upper bound (similar as for MINRES and GMRES). In order to simplify the much more complicated analysis for SYMMLQ, we have chosen to study only the effect of the errors introduced by the inversion of L_k. The resulting error $\Delta\mathbf{x}_k$ is written as

$$\Delta\mathbf{x}_k = \mathbf{V}_{k+1} \, \underline{Q}_k (\widehat{g}_k - g_k) \quad \text{with} \quad L_k g_k = \|\mathbf{r}_0\|_2 \, e_1, \tag{26}$$

where g_k represents the exact solution and $\widehat{g}_k$ is the value obtained in finite precision arithmetic. We write $g_k/\|\mathbf{r}_0\|_2 = (\gamma_1, \ldots, \gamma_k)^T$, and likewise the coordinates of $\widehat{g}_k/\|\mathbf{r}_0\|_2$ are denoted by $\widehat{\gamma}_j$. These coordinates can be written as

$$\gamma_k = e_k^T L_k^{-1} e_1, \quad \widehat{\gamma}_k = e_k^T (L_k + \Delta_L)^{-1} e_1, \qquad \text{with} \quad |\Delta_L| \leq 3 \, \mathbf{u} \, |L_k|.$$

With $L_{k+2} = (\ell_{ij})$, some formula manipulation leads to

$$\mathbf{A}\mathbf{V}_{k+1} \, \underline{Q}_k = \mathbf{V}_{k+2} \, \underline{T}_{k+1} \, \underline{Q}_k = \mathbf{V}_k L_k + [\mathbf{v}_{k+1}, \mathbf{v}_{k+2}] \, M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix}, \tag{27}$$

where

$$M_k \equiv \begin{bmatrix} \ell_{k+1\,k-1} & \ell_{k+1\,k} \\ 0 & \ell_{k+2\,k} \end{bmatrix}. \tag{28}$$

From (25) it follows that

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

Hence, the error in the SYMMLQ residual $\mathbf{r}_k^{\mathrm{ME}}$ can be written as

$$\mathbf{V}_k L_k(\widehat{g}_k - g_k) + \|\mathbf{r}_0\|_2 \,[\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]\, M_k \begin{bmatrix} \widehat{\gamma}_{k-1} - \gamma_{k-1} \\ \widehat{\gamma}_k - \gamma_k \end{bmatrix}. \tag{30}$$

The first term can be treated as in GMRES:

$$\mathbf{V}_k L_k(\widehat{g}_k - g_k) = -\|\mathbf{r}_0\|_2 \,\mathbf{V}_k \,\Delta_L \, L_k^{-1} e_1 \quad \text{with} \quad |\Delta_L| \leq 3\,\mathbf{u}\,|L_k|.$$

We define

$$t_k \equiv M_k \begin{bmatrix} \gamma_{k-1} \\ \gamma_k \end{bmatrix}, \quad \text{and} \quad \widehat{t}_k \equiv M_k \begin{bmatrix} \widehat{\gamma}_{k-1} \\ \widehat{\gamma}_k \end{bmatrix}.$$

By combining (29), (27), and the definition for t_k, we conclude that

$$\mathbf{r}_k = \|\mathbf{r}_0\|_2 \,[\mathbf{v}_k, \mathbf{v}_{k+1}]\, t_k,$$

and because of the orthogonality of $\mathbf{v}_k$ and $\mathbf{v}_{k+1}$, we have that

$$\|\mathbf{r}_k\|_2 = \|\mathbf{r}_0\|_2 \,\|t_k\|_2. \tag{31}$$

The computed residual 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. Expression (30) can now be bounded realistically by

$$3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_0\|_2 + \|\widehat{t}_k - t_k\|_2\,\|\mathbf{r}_0\|_2. \tag{32}$$

Here we have used that $\|\,|L_k|\,\|_2 \|L_k^{-1}\|_2 \leq \sqrt{3}\,\kappa_2(\mathbf{A})$. Hence

$$\frac{\|\Delta \mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(\mathbf{A}) + \|\widehat{t}_k - t_k\|_2. \tag{33}$$

A straight-forward estimate is

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

which is much larger than the first term in (33). 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. (49)). Fig. 4 illustrates that the upper bound in (33), with $\|\widehat{t}_k - t_k\|_2 \approx 0$, is fairly sharp.

Accuracy. From (33) it follows that

$$\left| \|\widehat{t}_k\|_2 - \frac{\|\widehat{\mathbf{r}}_k\|_2}{\|\mathbf{r}_0\|_2} \right| \leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(\mathbf{A}) + 2\,\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2}, \tag{34}$$

where $\widehat{\mathbf{r}}_k$ is the SYMMLQ residual with respect to the computed SYMMLQ approximate and $\mathbf{r}_k$ is the SYMMLQ residual for the exact SYMMLQ approximate (for the finite precision Lanczos). Apparently, assuming that $\|\mathbf{r}_k\|_2 \to 0$ if k increases, SYMMLQ is rather accurate since, for any method, errors in the order $\mathbf{u}\,\kappa_2(\mathbf{A})$ should be expected anyway.

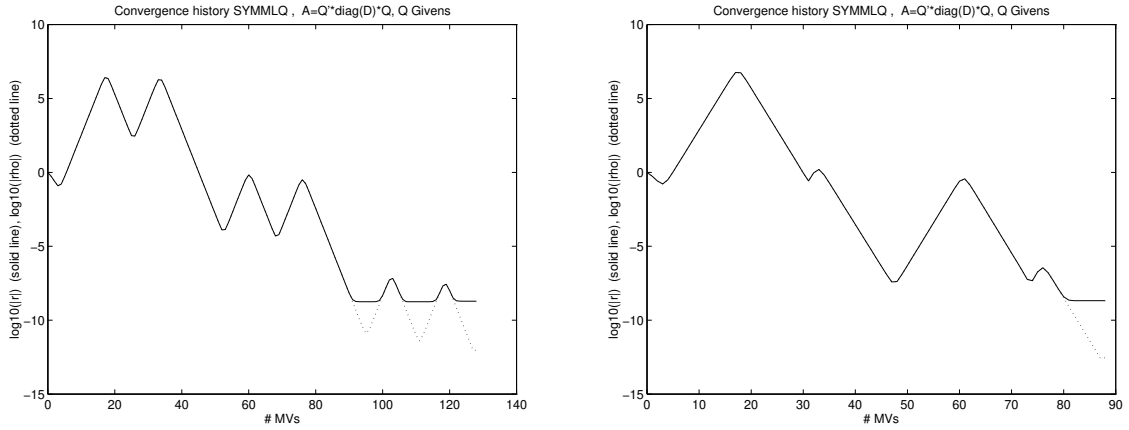

FIGURE 4. SYMMLQ: *solid line* (—) $\log_{10}$ *of* $\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_k\|_2/\|\mathbf{r}_0\|_2$; *dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* $\|\widehat{t}_k\|_2$. *The pictures show the results for the positive definite system* (*the left picture*) *and for the non-definite system* (*the right picture*) *of* Fig. 1. *Both systems have condition number* $3\,10^8$.

Convergence. It is not clear yet whether the convergence of SYMMLQ is insensitive to rounding errors. This would follow from (33) 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 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 analyse 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 section we will assume exact arithmetic, in particular the Lanczos process is assumed to be exact. The residuals $\mathbf{r}_k^{\mathrm{MR}}$ and $\mathbf{r}_k^{\mathrm{ME}}$ denote the residuals of MINRES and SYMMLQ, respectively.

The norm of the residual $\mathbf{b} - \mathbf{A}\mathbf{x}^{\mathrm{b}}$, with $\mathbf{x}^{\mathrm{b}}$ the best approximate of $\mathbf{x}$ in $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$, i.e., $\|\mathbf{x} - \mathbf{x}^{\mathrm{b}}\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{y} \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$, can be bounded in terms of the norm of the MINRES residual $\mathbf{r}_k^{\mathrm{MR}}$:

$$\frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}^{\mathrm{b}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \leq \kappa_2(\mathbf{A}). \tag{35}$$

This follows from the observation that $\mathbf{r}_k^{\mathrm{MR}} = \mathbf{b} - \mathbf{A}\mathbf{x}_k^{\mathrm{MR}}$ where $\mathbf{x}_k^{\mathrm{MR}}$ is from the same subspace

from which the best approximate $\mathbf{x}^{\rm b}$ has been selected, and furthermore that $\|\mathbf{b}-\mathbf{A}\mathbf{x}^{\rm b}\|_2 \leq \|\mathbf{A}\|_2\,\|\mathbf{x}-\mathbf{x}^{\rm b}\|_2$ and $\|\mathbf{x}-\mathbf{x}_k^{\rm MR}\|_2 \leq \|\mathbf{A}^{-1}\|_2\,\|\mathbf{r}_k^{\rm MR}\|_2$. Unfortunately, SYMMLQ selects its approximation $\mathbf{x}_k$ from a different subspace, namely $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$. This makes a comparison less straight forward.

The following lemma will be used for bounding the SYMMLQ error in terms of the MINRES error. Its proof uses the fact that $\mathbf{r}_k^{\rm MR}$ connects $\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$ and $\mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, that is, $\mathbf{r}_k^{\rm MR} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$, $\mathbf{r}_k^{\rm MR} \perp \mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, and hence $\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$ is spanned by $\mathbf{r}_k^{\rm MR}$ and $\mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$.

Lemma 3.1 *For each* $\mathbf{z} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$, *we have*

$$\|\mathbf{x}-\mathbf{x}_k^{\rm ME}\|_2^2 \leq \|\mathbf{x}-\mathbf{z}\|_2^2 + |\alpha_k|^2\,\|\mathbf{r}_k^{\rm MR}\|_2^2 \quad \textit{where} \quad \alpha_k \equiv \frac{(\mathbf{x},\mathbf{r}_k^{\rm MR})}{\|\mathbf{r}_k^{\rm MR}\|_2^2}. \tag{36}$$

Proof. For simplicity we will assume that $x_0 = \mathbf{0}$.
By construction $\mathbf{x}_k^{\rm ME}$ minimizes $\|\mathbf{x}-\mathbf{z}\|_2$ over all $\mathbf{z}$ in the space $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$. Hence $\mathbf{x}-\mathbf{x}_k^{\rm ME} \perp \mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$. Since $\mathbf{r}_k^{\rm MR} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, it follows that $(\mathbf{x}_k^{\rm ME},\mathbf{r}_k^{\rm MR}) = 0$, and therefore

$$\alpha_k = (\mathbf{x}-\mathbf{x}_k^{\rm ME},\mathbf{r}_k^{\rm MR})/\|\mathbf{r}_k^{\rm MR}\|_2^2 \quad \text{and} \quad \mathbf{x}-\mathbf{x}_k^{\rm ME} - \alpha_k\mathbf{r}_k^{\rm MR} \perp \mathbf{r}_k^{\rm MR}. \tag{37}$$

Since $\mathbf{x}-\mathbf{x}_k^{\rm ME} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ and $\mathbf{r}_k^{\rm MR} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, equation (37) implies that

$$\mathbf{x}-\mathbf{x}_k^{\rm ME} - \alpha_k\mathbf{r}_k^{\rm MR} \perp \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$$

By construction we have that $\mathbf{x}_k^{\rm ME} - \alpha_k\mathbf{r}_k^{\rm MR} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$ and as a consequence:

$$\|\mathbf{x}-\mathbf{x}_k^{\rm ME} - \alpha_k\mathbf{r}_k^{\rm MR}\|_2 \leq \|\mathbf{x}-\mathbf{z}\|_2 \quad \text{for all} \quad \mathbf{z} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0). \tag{38}$$

From Pythogoras' theorem, with (37), we conclude that

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

and (36) follows by combining this result with (38). □

Unfortunately, a combination of (36) with $\mathbf{z} = \mathbf{x}_k^{\rm MR}$ and the obvious estimate $|\alpha_k|\,\|\mathbf{r}_k^{\rm MR}\|_2 \leq \|\mathbf{x}-\mathbf{x}_k^{\rm ME}\|_2$, from (37) 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

$$\|\mathbf{r}_k^{\rm ME}\|_2 \leq \nu_{k+1}\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_k^{\rm MR}\|_2 \quad \textit{with} \quad \nu_k \equiv k + \tfrac{1}{2}\ln(k). \tag{39}$$

Proof. We use the relation

$$\mathbf{r}_k^{\rm MR} = s^2\mathbf{r}_{k-1}^{\rm MR} + c^2\mathbf{r}_k^{\rm CG}, \tag{40}$$

where $\mathbf{r}_k^{\rm CG}$ 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 [2, 22]. For symmetric $\mathbf{A}$, GMRES is equivalent with MINRES, and FOM is equivalent with CG. Since $\mathbf{r}_k^{\rm CG} = \|\mathbf{r}_k^{\rm CG}\|_2\mathbf{v}_{k+1}$, it follows that

$$\mathbf{r}_k^{\rm MR} = s^2\mathbf{r}_{k-1}^{\rm MR} + \gamma\mathbf{v}_{k+1} \quad \text{where} \quad s \equiv \frac{\|\mathbf{r}_k^{\rm MR}\|_2}{\|\mathbf{r}_{k-1}^{\rm MR}\|_2}, \tag{41}$$

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

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 $\|\gamma \mathbf{v}_{k+1}\|_2 \leq \|\mathbf{r}_k^{\mathrm{MR}}\|_2$. 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 (41) implies

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

and hence

$$|\alpha_k| \leq \frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{x} - \mathbf{x}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}. \tag{42}$$

A combination of (42) and (36) with $\mathbf{z} = \mathbf{x}_{k+1}^{\mathrm{MR}}$ leads to

$$\frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} \leq \frac{\|\mathbf{x} - \mathbf{x}_{k+1}^{\mathrm{MR}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} + \left(\frac{\|\mathbf{e}_{k-1}^{\mathrm{ME}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{x} - \mathbf{x}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \right)^2. \tag{43}$$

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 $\|\mathbf{r}_{k+1}^{\mathrm{MR}}\|_2 \leq \|\mathbf{r}_k^{\mathrm{MR}}\|_2$, we obtain the following recursive upper bound from (43):

$$\beta_k^2 \leq 1 + (\beta_{k-1} + 1)^2, \qquad \beta_0 = \frac{1}{\|\mathbf{A}^{-1}\|_2} \frac{\|\mathbf{e}_0^{\mathrm{ME}}\|_2}{\|\mathbf{r}_0^{\mathrm{MR}}\|_2} \leq 1.$$

A simple induction argument shows that $\beta_k \leq \nu_{k+1}$, and the definition of β_k implies

$$\frac{\|\mathbf{r}_k^{\mathrm{ME}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \leq \kappa_2(\mathbf{A}) \, \beta_k, \tag{44}$$

which completes the proof. □

For our analysis 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* $\mathbf{y}$.
For the best approximation $\mathbf{y}_k^{\mathrm{ME}}$ *of* $\mathbf{y}$ *in* $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$, *and for* $\mathbf{y}_k^{\mathrm{MR}} \in \mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$ *such that* $\mathbf{A}\mathbf{y}_k^{\mathrm{MR}}$ *is the best approximation of* $\mathbf{c}$ *in* $\mathbf{A}\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$, *with* ν_k *as in* (39), *we have*

$$\frac{\|\mathbf{c} - \mathbf{A}\,\mathbf{y}_k^{\mathrm{ME}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \leq \nu_{k+1} \, \kappa_2(\mathbf{A}) \, \mu_k, \quad \textit{where} \quad \mu_k \equiv \sup_{i \leq k} \frac{\|\mathbf{c} - \mathbf{A}\,\mathbf{y}_i^{\mathrm{MR}}\|_2}{\|\mathbf{r}_i^{\mathrm{MR}}\|_2}. \tag{45}$$

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, (36) 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

$$\frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} \leq \frac{\|\mathbf{y} - \mathbf{y}_{k+1}^{\mathrm{MR}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} + \left(\frac{\|\mathbf{e}_{k-1}^{\mathrm{ME}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{y} - \mathbf{y}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \right)^2 . \tag{46}$$

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

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

which implies (45). □

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 ([2, 22])

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

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. From relation (44) in combination with (31) we conclude that

$$\|t_k\|_2 \leq \rho_k\, \kappa_2(T_m)\, \nu_{k+1} \quad \text{with} \quad \nu_k = k + \tfrac{1}{2}\ln(k), \tag{47}$$

for all $m > k$.
Note that inequality (47) is correct for any symmetric tri-diagonal extension $\widetilde{T}_m$ of T_{k+1}: (47) holds with $\widetilde{T}_m$ instead of T_m. It has been shown in [6] 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

$$\|t_k\|_2 \lesssim \rho_k\, \kappa_2(\mathbf{A})\, \nu_{k+1} \quad \text{with} \quad \nu_k = k + \tfrac{1}{2}\ln(k). \tag{48}$$

In §3.1.1, we will show that

$$\|t_k - \widehat{t}_k\|_2 \lesssim 5\, \mathbf{u}\, \rho_k\, \kappa_2(\mathbf{A})^2 \left(\tfrac{1}{6}k^3 + \mathcal{O}(k^2 \ln k)\right). \tag{49}$$

The upper bound in (49) 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 (48) and (49), respectively, appear to be sharp (see Fig. 5).

Although the maximal values of the ratio of $\|t_k - \widehat{t}_k\|_2/\rho_k$ in Fig. 5 exhibit slowly growing behavior, the growth is not of order k^3. In the proof of (49) (cf. §3.1.1), upper bounds as in (48) are used in a consecutive number of steps. In view of the irregular convergence of SYMMLQ, the upper bound (48) 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.

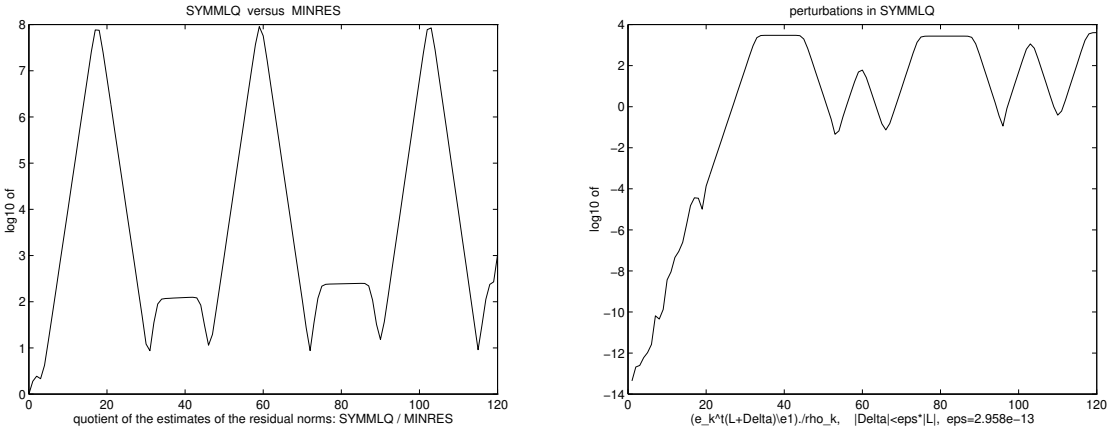

FIGURE 5. *Results for the non-definite matrix with condition number* $3\,10^8$ *(as in the right pictures) of* Fig. 1 *and* Fig. 4. *The left picture shows* $\log_{10}$ *of the ratio* $\|\widehat{t}_k\|_2/\rho_k$ *of the estimated residual norm reduction of* SYMMLQ *with the one of* MINRES*, the right picture models* $\|\widehat{t}_k - t_k\|_2/\rho_k$*: it shows the* $\log_{10}$ *of* $e_k^T(L_k + \Delta_L)^{-1}e_1/\rho_k$*, where* $|\Delta_L| \leq 3\,10^{-13}\,|L_k|$.

3.1.1 SYMMLQ recurrences

In this section we derive the upper bound (49).

Suppose that the jth recurrence for the γ_i's is perturbed by a relatively small δ and all other recurrence relation are exact:

$$\delta = \ell_{jj}\widetilde{\gamma}_j + \ell_{jj-1}\gamma_{j-1} + \ell_{jj-2}\gamma_{j-2} \quad \text{with} \quad |\delta| \leq \mu\,\mathbf{u}\,|\ell_{jj}|\,|\gamma_j|. \tag{50}$$

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

$$\widetilde{t}_k - t_k = \delta\, M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_j. \tag{51}$$

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 (48), 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, with $\mathbf{v}_j = e_j$, $\mathbf{A} = T_m$, and $\mathbf{c} = \mathbf{v}_{j+1}$, we have (in the notation of Theorem 3.3), for $j < k$,

$$\mathbf{y}_j^{\mathrm{ME}} = \mathbf{0}, \qquad \|\mathbf{c} - \mathbf{A}\mathbf{y}_k^{\mathrm{ME}}\|_2 = \|M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_{j+1}\|_2, \tag{52}$$

and

$$\|\mathbf{c} - \mathbf{A}\mathbf{y}_k^{\mathrm{MR}}\|_2 = c_j\,\frac{\rho_k}{\rho_j}, \tag{53}$$

with c_j the cosine in the jth Givens rotation. Therefore, by Theorem 3.3,

$$\|M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_{j+1}\|_2 \leq \kappa_2(L_k)\,c_j\,\nu_{k+1}\,\frac{\rho_k}{\rho_j}. \tag{54}$$

For this specific situation, the estimate for β_k in the last paragraph of the proof of Theorem 3.2 can be improved. It can be shown that $\beta_j \leq 1$ if $\beta_k \leq \nu_{k-j}$. Therefore, the ν_{k+1} in (54) can be replaced by ν_{k-j}.

A combination of (51) with (54) gives

$$\|\widetilde{t}_k - t_k\|_2 \leq \frac{|\delta|}{|\rho_{j-1}|}\,\rho_k\,\kappa_2(L_k)\,\nu_{k-j+1} \lesssim \frac{|\delta|}{|\rho_{j-1}|}\,\rho_k\,\kappa_2(\mathbf{A})\,\nu_{k-j+1}. \tag{55}$$

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 (48), we have that

$$|\ell_{jj}|\,\frac{|\gamma_j|}{\rho_{j-1}} \leq \frac{\|t_{j-1}\|_2}{\rho_{j-1}} \leq \kappa_2(\mathbf{A})\,\nu_j. \tag{56}$$

Hence (cf. (50))

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

and, with (55), this gives gives

$$\|\widetilde{t}_k - t_k\|_2 \leq \mu\,\mathbf{u}\,\rho_k\,\kappa_2(\mathbf{A})^2\,\nu_j\nu_{k-j+1}. \tag{57}$$

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 (50) then the estimate (49) appears as a cumulation of bounds as in (57). The vector $\widehat{t}_k$ in (49) 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 (50), with $\mu = 5$.
First we note that the $\widetilde{\gamma}_k$'s resulting from the perturbation

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

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 $(j-1)$st 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

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

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

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

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 $(j-1)$st 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 mis-convergence. 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 beyond 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.

In view of the fact that we may expect at least a linear factor $\kappa_2(\mathbf{A})$, when working with Euclidean norms, GMRES (§2.2) and SYMMLQ (§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. For ill-conditioned systems this may be considerable.

As has been pointed out in [14], the Conjugate Gradient iterates can be constructed with little effort from SYMMLQ information if the 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 has no minimization property (as for 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 $\kappa_2(\mathbf{A})^2$ (§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 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 [20]). 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 be even 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 gives often a quite good impression of what to expect for non-diagonal matrices with the same spectrum. However, as we have shown in our §2.5, for MINRES this may lead to a too optimistic picture, since floating point error perturbations with MINRES lead to errors in the residual (and the approximated solution) that are a factor $\kappa_2(\mathbf{A})$ smaller as for non-diagonal matrices.

References

[1] R. Barrett, M. Berry, T.F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H.A. van der Vorst, *Templates for the solution of linear systems:building blocks for iterative methods*, SIAM, Philadelphia, 1994.

[2] P.N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 58–78.

[3] A.M. Bruaset, *A survey of preconditioned iterative methods*, Wiley, New York, 1995.

[4] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley-Teubner, Stuttgart 1996.

[5] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Third edition, The John Hopkins University Press, Baltimore and London, 1996.

[6] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

[7] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[8] 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.

[9] M.R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1954), pp. 409–436.

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

[11] C.C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Maths. Applics 18 (1976), pp. 341–349.

[12] C.C. Paige, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.

[13] C.C. Paige, B.N. Parlett, and H.A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Num. Lin. Algebra with Appl., 2 (1995), pp. 115–134.

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

[15] B.N. Parlett, *The symmetric eigenvalue problem*, Prentice-Hall series in Computational Mathematics, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1980.

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

[17] G.L.G. Sleijpen, H.A. van der Vorst, and D.R. Fokkema, *BiCGstab(ℓ) and other Hybrid Bi-CG Methods*, Numerical Algor., 7 (1994), pp. 75–109.

[18] G.L.G. Sleijpen and H.A. van der Vorst, *Reliable updated residuals in hybrid Bi-CG methods*, Computing, 56 (1996), pp. 141–163.

[19] E. Stiefel, *Relaxiationsmethoden bester Strategie zur Lösung linearer Gleichungssysteme*, Comm. Math. Helv., 29 (1955), pp. 157–179.

[20] K. Turner and H. F. Walker, *Efficient High Accuracy Solutions with GMRES(m)*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 815–825.

[21] A. vand der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

[22] H.A. van der Vorst and C. Vuik, *The superlinear convergence behaviour of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.

5 Appendix

Lemma 5.1 *If, for a matrix* $\mathbf{C}$, $n_C = \min(n_c, n_r)$ *with* n_c *the maximum number of non-zero's per column and* n_r *the maximum number of non-zero's per row, then*

$$\| \, |\mathbf{C}| \, \|_2 \leq \sqrt{n_C} \, \|\mathbf{C}\|_2. \tag{58}$$

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 [21, Th. 4.2]), we have

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

GMRES

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 $k = 1$

$\vec{\ell} = [\,],\ R = [\,l_0\,]$

else

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

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

end if

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

end while

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

FIGURE 6. *The* GMRES *algorithm.*

The vector $\vec{0}$ *for the expansion of the upper triangular matrix* R *is a row vector of zero's of appropriate size (different size at different occurences).*

MINRES

Choose $\mathbf{x}_0$

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

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

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

while $|\rho| > tol$ do

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

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

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

$l_1 \leftarrow s\,\alpha - c\,\widetilde{\beta},\ l_2 \leftarrow s\,\beta$

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

$l_0 \leftarrow \sqrt{\widetilde{\alpha}^2 + \beta^2},\ c \leftarrow \widetilde{\alpha}/l_0,\ s \leftarrow \beta/l_0$

$\widetilde{\mathbf{w}} \leftarrow \widetilde{\widetilde{\mathbf{w}}} - l_1\,\mathbf{w},\ \widetilde{\widetilde{\mathbf{w}}} \leftarrow \mathbf{v} - l_2\,\mathbf{w},\ \mathbf{w} \leftarrow \widetilde{\mathbf{w}}/l_0$

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

end while

FIGURE 7. *The* MINRES *algorithm*

SYMMLQ

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

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

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

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

$l_1 \leftarrow s\,\alpha - c\,\widetilde{\beta},\ l_2 \leftarrow s\,\beta$

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

$l_0 \leftarrow \sqrt{\widetilde{\alpha}^2 + \beta^2},\ c \leftarrow \widetilde{\alpha}/l_0,\ s \leftarrow \beta/l_0$

$\widetilde{g} \leftarrow \widetilde{\widetilde{g}} - l_1\,g,\ \widetilde{\widetilde{g}} \leftarrow -l_2\,g,\ g \leftarrow \widetilde{g}/l_0$

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

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

end while

FIGURE 8. *The* SYMMLQ *algorithm*