SYMMLQ

```
Choose 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 $\kappa > \mathit{tol}$ 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

The main effects of rounding errors in Krylov solvers for symmetric linear systems

by

Gerard L.G. Sleijpen, Henk A. Van der Vorst, and Jan Modersitzki

Universiteit Utrecht

Department of Mathematics

Preprint

nr. 1006

March 10, 1997

THE MAIN EFFECTS OF ROUNDING ERRORS IN KRYLOV SOLVERS FOR SYMMETRIC LINEAR SYSTEMS

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

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 linear systems, by solving the reduced system in one way or another. This leads to well-known methods: MINRES (GMRES), CG, CR, and SYMMLQ. We will discuss in what way and to what extent the various approaches are sensitive to rounding errors.
In our analysis we will assume that the Lanczos basis is generated in exactly the same way for the different methods (except CR), and we will not consider the errors in the Lanczos process itself. These errors may lead to large perturbations with respect to the exact process, but convergence takes still place. Our attention is focussed to what happens in the solution phase. We will show that the way of solution may lead, under circumstances, to large additional errors, that are not corrected by continuing the iteration process.
Our findings are supported and illustrated by numerical examples.

1 A basis for the Krylov subspace

All the iterative methods that we consider construct an approximate solution, starting with $\mathbf{x}_0$, for the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, with $\mathbf{A}$ symmetric, in the k-dimensional Krylov subspace

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

with $\mathbf{r}_0 \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_0$.
With the standard 3-term Lanczos process, we generate an orthonormal basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$, with $\mathbf{v}_1 \equiv \mathbf{r}_0/\|\mathbf{r}_0\|_2$, for $\mathcal{K}_k(\mathbf{A}; \mathbf{r}_0)$. The result of the Lanczos process can be formally written as

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

We will present solution methods on the basis of this Lanczos process, and we will analyse what additional errors are introduced to working with the $\mathbf{v}_k$ and $\underline{T}_k$. The errors in the Lanczos process itself are not considered here, these errors have been analysed by Paige [10, 11]. It has been proven by Greenbaum and Strakos [5], that rounding errors in the Lanczos process may have a delaying effect on the convergence of iterative solvers, but these iterative solvers usually produce converging sequences $\{\mathbf{x}_k\}$. The question is how close the $\mathbf{x}_k$ can come to the best solution, with respect to the generated Krylov subspace, in the presence of rounding errors. Since the residuals are usually exploited for monitoring the convergence, we will derive sharp upper bounds for the rounding errors in the residuals.

We will consider the following iterative solution methods:

*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.ruu.nl, vorst@math.ruu.nl, modersitzki@informatik.mu-luebeck.de

1. **MINRES (GMRES)**: determine $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k\mathbf{y}_k$, such that the residual $\mathbf{b} - \mathbf{A}\mathbf{x}_k$ has minimal Euclidean length.
   This can be done with short recurrences, which avoids the necessity to store all basis vectors $\mathbf{v}_j$, and this leads to MINRES [13]. It can also be done with a long recurrence and this gives GMRES [15]. Although in practice one always prefers MINRES over GMRES for symmetric systems, we will show that the GMRES approach has certain advantages with respect to the effects of rounding errors.

2. **CG**: determine $\mathbf{x}_k = \mathbf{x}_0+\mathbf{V}_k\mathbf{y}_k$, such that the residual $\mathbf{b}-\mathbf{A}\mathbf{x}_k$ is orthogonal with respect to $\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ (a Galerkin condition). In actual computations one often uses a scaled basis: $\mathbf{W}_k = \mathbf{V}_k D_k$, where the diagonal matrix D_k is chosen such that $\mathbf{w}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j$. CG can be used safely only for symmetric positive definite systems. The CG method was originally proposed in [8], for up to date references see [7, 1].

3. **SYMMLQ**: determine $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{A}\mathbf{V}_k\mathbf{y}_k$, such that the error $\mathbf{x} - \mathbf{x}_k$ has minimal Euclidean length. The method was proposed in [13].

4. **CR**: determine $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k\mathbf{y}_k$, such that the residual $\mathbf{b} - \mathbf{A}\mathbf{x}_k$ is orthogonal with respect to $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$. By working with $\mathbf{A}$ times the Krylov subspace we can avoid the condition that $\mathbf{A}$ has to be positive definite, as in CG. However, in CR we construct an $\mathbf{A}$-orthogonal basis for the Krylov subspace, which leads to a different tridiagonal matrix as in the standard Lanczos method. CR was introduced in [19], for a modern coverage of the method see [6].

With respect to the choice of these methods, as far as numerical stability is concerned, one cannot find much in literature. For SYMMLQ and other approaches like MINRES, it is stated in [13, p. 625] that these approaches are not as accurate as SYMMLQ for the reason 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 of the solution depends on $\kappa_2(\mathbf{A})^2$ (it is even said 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). Both remarks miss solid ground since MINRES and GMRES, which also follows a minimal residual approach, both do not work with $\mathbf{A}^2$, and as we will see the sensitivity for GMRES is governed by $\kappa_2(\mathbf{A})$.

Because of its conceptual simplicity, we will start with analyzing the main errors in the GMRES approach.

2 The additional errors introduced in GMRES

Since

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\mathbf{x}_k\|_2 &= \|\mathbf{r}_0 - \mathbf{A}\mathbf{V}_k\mathbf{y}_k\|_2 \\ &= \|\mathbf{r}_0 - \mathbf{V}_{k+1}\underline{T}_k\mathbf{y}_k\|_2 \\ &= \|\underline{T}_k\mathbf{y}_k - \|\mathbf{r}_0\|_2 e_1\|_2, \end{aligned} \tag{2}$$

we see that $\mathbf{y}_k$ is the linear least squares solution of the $k+1$ by k overdetermined system

$$\underline{T}_k\mathbf{y}_k - \|\mathbf{r}_0\|_2 e_1,$$

and the usual way to solve this is to reduce $\underline{T}_k$ first to upper triangular form:

$$\underline{T}_k = \underline{Q}_k R_k. \tag{3}$$

The matrix R_k is k by k upper triangular with bandwidth 3, and $\underline{Q}_k$ is a $k+1$ by k orthogonal matrix. The reduction is in GMRES done with Givens rotations, which only introduces $\mathcal{O}(\mathbf{u})$ errors in the matrices involved [7, p. 217], and we will neglect these errors. The scalar $\mathbf{u}$ denotes the relative machine precision.
With the QR decomposition, we have that y_k can be solved from

$$R_k y_k = z_k \equiv \|\mathbf{r}_0\|_2 \, \underline{Q}_k^T e_1, \tag{4}$$

and $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k y_k$. Combining the two relations gives

$$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k (R_k^{-1} \underline{Q}_k^T \|\mathbf{r}_0\|_2 \, e_1).$$

According to [7, p. 89], in floating point arithmetic the computed solution $\widehat{y}_k$ satisfies

$$(R_k + \Delta_R)\widehat{y}_k = z_k, \quad \text{with} \quad |\Delta_R| \leq 3\,\mathbf{u}\,|R_k| + \mathcal{O}(\mathbf{u}^2). \tag{5}$$

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

$$\Delta y_k = -R_k^{-1} \Delta_R R_k^{-1} z_k.$$

Hence, the error in $\mathbf{x}_k$, due to the floating point computation of the solution y_k, is

$$\Delta \mathbf{x}_k = -\mathbf{V}_k R_k^{-1} \Delta_R R_k^{-1} z_k,$$

and for the corresponding floating point error in the residual we have that:

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

Using the bound in (5), we get

$$\begin{aligned} \|\Delta \mathbf{r}_k\|_2 &\leq \|\mathbf{V}_{k+1} \underline{Q}_k\|_2 \, 3\,\mathbf{u}\, \|\,|R_k|\,\|_2 \, \|R_k^{-1}\|_2 \, \|\mathbf{r}_0\|_2 \\ &\leq 3\,\mathbf{u}\, \|\mathbf{V}_{k+1}\|_2 \, \|\,|R_k|\,\|_2 \, \|R_k^{-1}\|_2 \, \|\mathbf{r}_0\|_2 \\ &\leq 3\sqrt{3}\,\mathbf{u}\, \|\mathbf{V}_{k+1}\|_2 \, \kappa_2(R_k) \, \|\mathbf{r}_0\|_2. \end{aligned} \tag{6}$$

Here we have used that $\|\,|R_k|\,\|_2 \leq \sqrt{3}\,\|R_k\|_2$ (which follows from [21, Th. 4.2]; see Lemma 8.1 for details). The factor κ_2 denotes the condition number with respect to the Euclidean norm.

Note that we may 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 [14, 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

$$\underline{T}_k^T \underline{T}_k = R_k^T R_k = \mathbf{V}_k^T \mathbf{A}^T \mathbf{A} \mathbf{V}_k.$$

Hence,

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

and

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

which implies $\kappa_2(R_k) \leq \kappa_2(\mathbf{A})$, which finally results in the upper bound for the error in the residual due to the GMRES computations on top of the Lanczos process:

$$\frac{\|\Delta \mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_{k+1}\|_2\,\kappa_2(\mathbf{A}). \tag{7}$$

In our Fig. 1, we see that, apart from the factor $3\sqrt{3}\,\|\mathbf{V}_{k+1}\|_2$, this predicted error is attained fairly well in actual floating point computations. Note that, due to rounding errors in the matrix-vector multiplication, even the error $\Delta\mathbf{x}$ in the exact solution will be in norm in the order of $\mathbf{u}\,\|\mathbf{A}^{-1}\|_2\,\|\mathbf{b}\|_2$, which corresponds to an error $\|\mathbf{A}\Delta\mathbf{x}\|_2$ in the order of $\mathbf{u}\,\kappa_2(\mathbf{A})\|\mathbf{b}\|_2$ in the residual. Therefore, the stability of GMRES may considered to be optimal.

3 The additional errors introduced in MINRES

MINRES follows basically the same approach as GMRES for the minimization of the residual, but it exploits the banded structure of R_k in the computation of the vector $\mathbf{x}_k$.
As we have seen in §2, $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k\, R_k^{-1} \underline{Q}_k^T\, \|\mathbf{r}_0\|_2\, e_1$, and this can be rearranged as

$$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k R_k^{-1} z_k = \mathbf{x}_0 + \mathbf{W}_k\, z_k.$$

Note that the interpretation of $\mathbf{x}_k$ as $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{V}_k y_k$ leads to the GMRES computation. The interpretation $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{W}_k\, z_k$ leads to MINRES.
We first compute the matrix $\mathbf{W}_k = \mathbf{V}_k R_k^{-1}$, and it is easy to see that the last column of $\mathbf{W}_k$ is obtained from the last 2 columns of $\mathbf{W}_{k-1}$ and the last column of $\mathbf{V}_k$. The vector $\mathbf{x}_k = \mathbf{x}_0 + \mathbf{W}_k\, z_k$ can now easily be updated, also since z_k follows from a simple Givens rotation on the last coordinate of z_{k-1} and 0.
Now we have to analyze the floating point errors introduced by the computation of the columns of $\mathbf{W}_k$. Note that the j-th row $w_{j,:}$ of $\mathbf{W}_k$ has to satisfy

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

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

$$\widehat{w}_{j,:}(R_k + \Delta_{R_j}) = v_{j,:}, \tag{8}$$

with

$$|\Delta_{R_j}| \leq 3\,\mathbf{u}\,|R_k| + \mathcal{O}(\mathbf{u}^2). \tag{9}$$

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

$$\widehat{\mathbf{W}}_k = (\mathbf{V}_k + \Delta_W) R_k^{-1}, \tag{10}$$

with

$$|\Delta_W| \leq 3\,\mathbf{u}\,|\widehat{\mathbf{W}}_k|\,|R_k| + \mathcal{O}(\mathbf{u}^2) \tag{11}$$

Note that by replacing $\widehat{\mathbf{W}}_k$ by $\mathbf{W}_k = \mathbf{V}_k R_k^{-1}$ in (11), we introduce only $\mathcal{O}(\mathbf{u}^2)$ errors. Now the error $\Delta\mathbf{x}_k$ in the MINRES-$\mathbf{x}_k$ becomes

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

and the error in the MINRES residual:

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

If we use the bounds on Δ_W, and use for other quantities similar bounds as for GMRES, we obtain

$$\begin{aligned}
\|\Delta\mathbf{r}_k\|_2 &\leq 3\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\,|\mathbf{V}_k R_k^{-1}|\,\|_2\,\|\,|R_k|\,\|_2\,\|R_k^{-1}\|_2\,\|\mathbf{r}_0\|_2 \\
&\leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\mathbf{V}_k\|_F\,\|R_k^{-1}\|_2\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_0\|_2 \\
&\leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_F\,\kappa_2(\mathbf{A})^2\,\|\mathbf{r}_0\|_2.
\end{aligned}$$

Here we have also used the fact that

$$\|\,|\mathbf{V}_k R_k^{-1}|\,\|_2 \leq \|\mathbf{V}_k R_k^{-1}\|_F \leq \|\mathbf{V}_k\|_F\,\|R_k^{-1}\|_2, \tag{12}$$

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 in the residual due to the MINRES computations on top of the Lanczos process:

$$\frac{\|\Delta\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq 3\sqrt{3k}\,\mathbf{u}\,\kappa_2(\mathbf{A})^2. \tag{13}$$

In our Fig. 1, we see that, apart from the factor $3\sqrt{3k}$, the effect of the square of the condition number is clearly visible in deviation in the residual for MINRES. This analysis implies that one has to be careful with MINRES when solving linear systems with an ill-conditioned matrix $\mathbf{A}$.

The residual norm reduction $\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2$ of the exact 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 never computed 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 differs from the true 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, (13) implies that

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

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 (7). In fact, as observed at the end of §2, except for the moderate constant $3\sqrt{3}\,\|\mathbf{V}_{k+1}\|_2$, this is about the most accurate computation that can be expected.

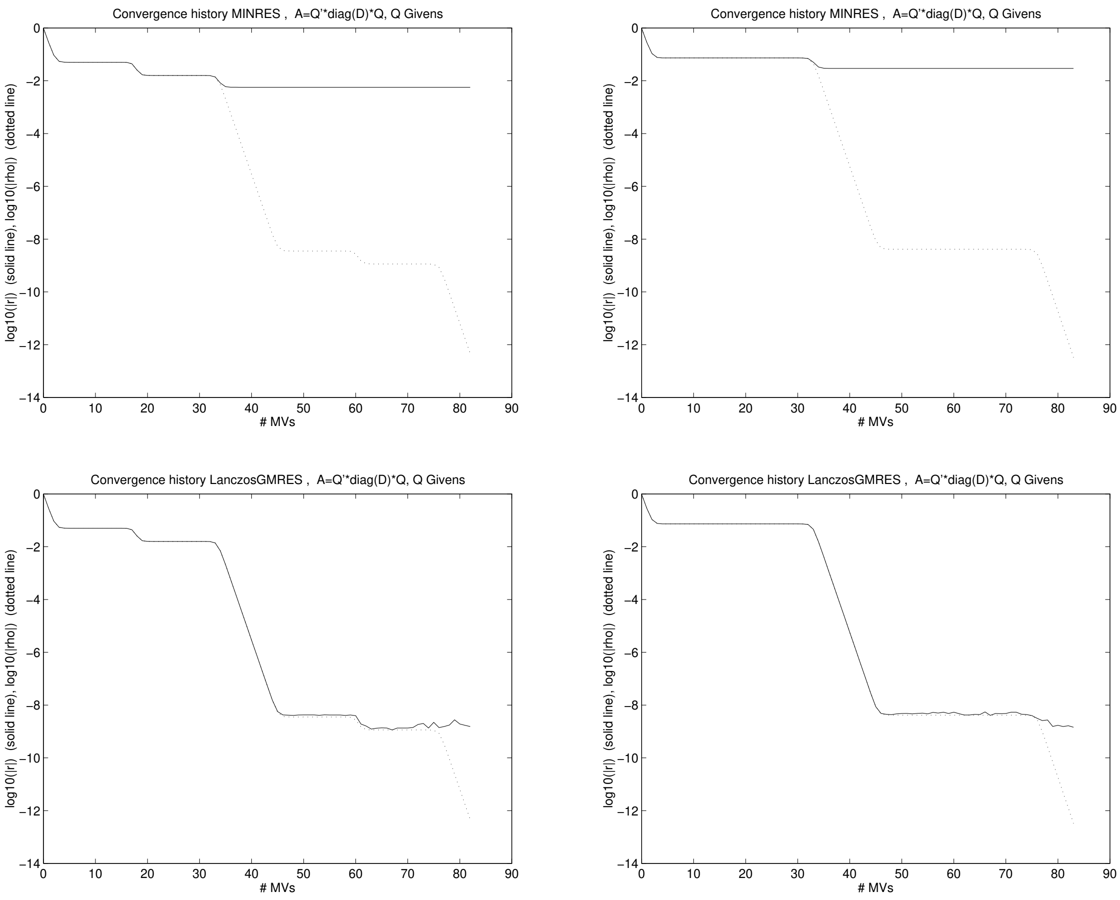


FIGURE 1. MINRES *(top) and* GMRES *(bottom): solid line (—)* $\log_{10}$ *of* $\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\|_2/\|\mathbf{r}_0\|_2$*, dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* ρ_k*. The pictures show the results for a positive definite system (the left pictures) 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* $\mathbf{A} = \mathbf{G}\mathbf{D}\mathbf{G}'$ *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* $\mathbf{A} = \mathbf{G}\mathbf{D}\mathbf{G}'$ *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* $\mathbf{u} = 1.1\,10^{-16}$*.*

3.1 Diagonal matrices

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 (8), the error in the jth coordinate of the MINRES-$\mathbf{x}_k$ is given by

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

When $\mathbf{A}$ is diagonal with (j,j)-entry λ_j, the error in the jth coordinate of the MINRES

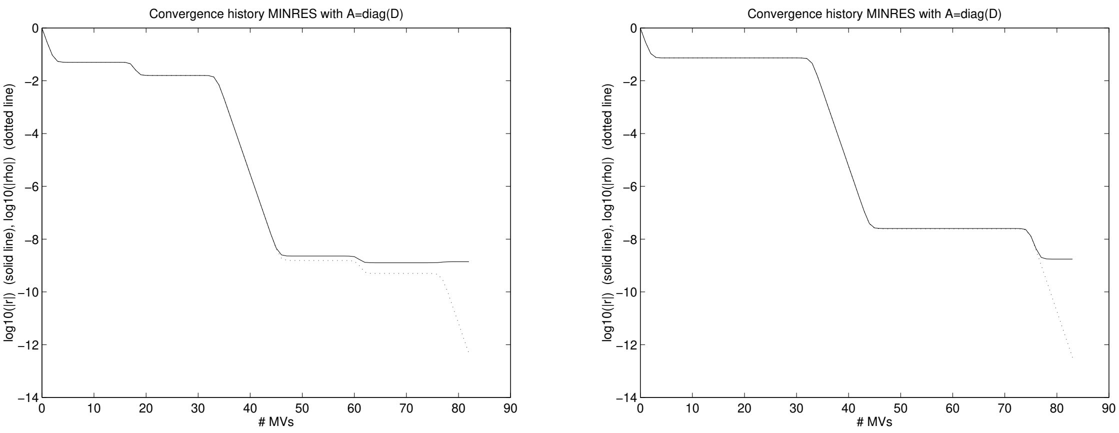


FIGURE 2. MINRES: *solid line (—)* $\log_{10}$ *of* $\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\|_2/\|\mathbf{r}_0\|_2$, *dotted line (⋯)* $\log_{10}$ *of the estimated residual norm reduction* ρ_k. *The pictures show the results for a positive definite diagonal system (the left picture) and for a non-definite 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* $\mathbf{G}=\mathbf{I}$.

residual is equal to (use (1) and (3))

$$(\Delta\mathbf{r}_k)_j = \lambda_j v_{j,\cdot} R_k^{-1}\Delta_{R_j}R_k^{-1}z_k = \mathbf{e}_j^T\mathbf{A}\mathbf{V}_kR_k^{-1}\Delta_{R_j}R_k^{-1}z_k = \mathbf{e}_j^T\mathbf{V}_{k+1}\underline{Q}_k\Delta_{R_j}R_k^{-1}z_k. \tag{14}$$

Therefore, in view of (9), for MINRES applied to a diagonal matrix we have,

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

which is the same upper bound as for the errors in the GMRES residuals in (7).

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, (14) may not 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 a mth coordinate and may not be damped by a small value $|\lambda_m|$. More precise, 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 $\Gamma\,\mathbf{u}\,\|R_k^{-1}\|_2\,\kappa_2(R_k^{-1}) \le \Gamma\,\mathbf{u}\,\|\mathbf{A}^{-1}\|_2\,\kappa_2(\mathbf{A})$. If $\Gamma\approx\|\mathbf{A}\|_2$, we recover the bound (13).

4 The additional errors in a CG-type approach

In this section we will assume that $\mathbf{A}$ is also positive definite.
The Galerkin condition for the residual $\mathbf{b}-\mathbf{A}\mathbf{x}_k$, with respect to $\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, for $\mathbf{x}_k=\mathbf{x}_0+\mathbf{V}_ky_k$ leads to

$$\mathbf{V}_k^T(\mathbf{r}_0-\mathbf{A}\mathbf{V}_ky_k)=0,$$

or

$$T_ky_k = \|\mathbf{r}_0\|_2e_1, \tag{15}$$

where T_k is the k by k upper block of $\underline{T}_k$. Hence,

$$\mathbf{x}_k = \mathbf{x}_0 + \|\mathbf{r}_0\|_2\mathbf{V}_kT_k^{-1}e_1. \tag{16}$$

The coupled 2-term recurrences, on which CG is based, arise if we factorize $T_k = L_k L_k^T$ (in actual implementations the factorization may be a variant of this). This leads to

$$\mathbf{x}_k = \mathbf{x}_0 + \|\mathbf{r}_0\|_2 (\mathbf{V}_k L_k^{-T})(L_k^{-1} e_1), \tag{17}$$

and the columns of $\mathbf{W}_k \equiv \mathbf{V}_k L_k^{-T}$ and the elements of $L_k^{-1} e_1$ can be obtained by two-term recurrences, since L_k is lower bidiagonal.

For the moment we ignore the errors made in the Choleski decomposition; they are of minor importance because of the stability of the Choleski algorithm for positive definite matrices(cf. [9, Theorem 10.3-4]).
We first consider the errors made in the floating point computation of $\mathbf{W}_k$, and we denote the floating point result by $\widehat{\mathbf{W}}_k$. Denoting the entries of L_k, $\widehat{\mathbf{W}}_k$, and $\mathbf{V}_k$ by ℓ_{ij}, $\widehat{w}_{ij}$, and v_{ij}, respectively, we have that

$$\widehat{w}_{ij} = \left(v_{ij} - \ell_{j,j-1} \widehat{w}_{i,j-1} (1 + \xi_{ij}^{(1)}) \right) (1 + 2\xi_{ij}^{(2)}) / \ell_{jj}, \tag{18}$$

with $|\xi| \leq \mathbf{u}$. This can be rewritten as

$$\widehat{w}_{ij} = \left(v_{ij}(1 + \xi_{ij}^{(3)}) - \ell_{j,j-1} \widehat{w}_{i,j-1} \right) / \ell_{jj}(1 + 3\xi_{ij}^{(4)}). \tag{19}$$

Taking the equations together in matrix form, we obtain

$$\widehat{\mathbf{W}}_k = (\mathbf{V}_k + \Delta_V) L_k^{-T} + \Delta_W, \quad \text{with} \quad |\Delta_V| \leq \mathbf{u}\, |\mathbf{V}_k| \quad \text{and} \quad |\Delta_W| \leq 3\, \mathbf{u}\, |\mathbf{W}_k|, \tag{20}$$

ignoring terms of $\mathcal{O}(\mathbf{u}^2)$. Note that these bounds are essentially better than those for MINRES (cf. (10) and (20)), which is due to the fact that in CG we have to do with (coupled) 2-term recurrences instead of a 3-term recurrence: in a 2-term recurrence, new local errors can be modeled as perturbations on local vector-elements only (cf. (18) and (19), which is impossible for 3-term recurrences.

These errors lead to a contribution $\Delta_1 \mathbf{r}_k$ in the residual of $\mathbf{r}_k$:

$$\Delta_1 \mathbf{r}_k = -\|\mathbf{r}_0\|_2\, \mathbf{A} (\Delta_V L_k^{-T} + \Delta_W) L_k^{-1} e_1. \tag{21}$$

We can bound this error contribution as (using that $\|\,|\mathbf{V}_k|\,\|_2 \leq \|\mathbf{V}_k\|_F$; cf. (12))

$$\|\Delta_1 \mathbf{r}_k\|_2 \leq \mathbf{u}\, \|\mathbf{A}\|_2 \left(\|\mathbf{V}_k\|_F\, \|L_k^{-T}\|_2\, \|L_k^{-1}\|_2 + 3\, \|\mathbf{V}_k\|_F\, \|L_k^{-T}\|_2\, \|L_k^{-1}\|_2 \right) \|\mathbf{r}_0\|_2. \tag{22}$$

For the Choleski process we have that $\|T_k^{-1}\|_2 = \|L_k^{-1}\|_2^2 = \|L_k^{-1}\|_2 \|L_k^{-T}\|_2$, and since the eigenvalues of T_k are inside the spectrum of $\mathbf{A}$, we have that $\|T_k^{-1}\|_2 \leq \|\mathbf{A}^{-1}\|_2$. Note that this argument does not apply if $\mathbf{A}$ is not definite: then T_k can have eigenvalues close to or equal to 0 and $\|T_k^{-1}\|_2$ can be arbitrarily large.
Thus we arrive at

$$\|\Delta_1 \mathbf{r}_k\|_2 \leq 4\, \mathbf{u}\, \|\mathbf{V}_k\|_F\, \kappa_2(\mathbf{A})\, \|\mathbf{r}_0\|_2. \tag{23}$$

Now we have to consider the errors in (17) due to the part $L_k^{-1} e_1$. Let us denote $z_k = L_k^{-1} e_1$, then the floating point solution $\widehat{z}_k$ satisfies:

$$\widehat{z}_k^{(1)} = 1/\ell_{11}(1 + \xi^{(1)}),$$

$$\widehat{z}_k^{(i)} = -\ell_{i,i-1} \widehat{z}_k^{(i-1)} / \ell_{ii}(1 + 2\xi^{(i)}),$$

which shows that $\widehat{z}_k = L_k^{-1} e_1 + \Delta z$, with $\|\Delta z\|_2 \leq 2\,\mathbf{u}\,\|z\|_2$.
These errors lead to another first order contribution $\Delta_2 \mathbf{r}_k$ in the residual:

$$\Delta_2 \mathbf{r}_k = \|\mathbf{r}_0\|_2\, \mathbf{A} \mathbf{V}_k L_k^{-T} \Delta_z, \tag{24}$$

and we obtain the following upper bound for the norm

$$\|\Delta_2 \mathbf{r}_k\|_2 \leq 2\,\mathbf{u}\,\|\mathbf{A}\|_2\,\|\mathbf{V}_k\|_2\,\|L_k^{-T}\|_2\,\|L_k^{-1}\|_2\,\|\mathbf{r}_0\|_2. \tag{25}$$

This leads to the upper bound for the error in $\mathbf{r}_k$;

$$\frac{\|\Delta \mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq \left(4\sqrt{k} + 2\,\|\mathbf{V}_k\|_2\right) \mathbf{u}\,\kappa_2(\mathbf{A}). \tag{26}$$

Actual implementations, see for instance [1], are not based on the $\mathbf{v}_j$-basis, but on the residuals $\mathbf{r}_j$ as a basis for the Krylov subspace. Also, but this is a technical detail, Choleski's decomposition is not exploited, but rather an LU-decomposition, with $\mathrm{diag}(U) = I$.
In our derivation of bounds on the norm of $\Delta_1 \mathbf{r}_k$, the Choleski factors play a role only in the bound for $\|\Delta_W L_k^{-1}\|_2$ with $|\Delta_W| \leq 3\,\mathbf{u}\,|\mathbf{V}_k L_k^{-T}|$ (cf. (21) and (22)).
Similarly, when working with $T_k = \widetilde{L}_k U_k$, the factors are relevant only for the bound on $\|\widetilde{\Delta}_W \widetilde{L}_k^{-1}\|_2$ with $|\widetilde{\Delta}_W| \leq 3\,\mathbf{u}\,|\mathbf{V}_k U_k^{-1}|$. The factors L_k and $\widetilde{L}_k$ are related as: $\widetilde{L}_k = L_k D_k$. For the nonsingular diagonal matrix D_k, we also have that $U_k = D_k^{-1} L_k^T$. Note that, $\|\widetilde{\Delta}_W \widetilde{L}_k^{-1}\|_2 = \|\widetilde{\Delta}_W D_k^{-1} L_k^{-1}\|_2$ and $|\widetilde{\Delta}_W D_k^{-1}| \leq 3\,\mathbf{u}\,|\mathbf{V}_k U_k^{-1}||D_k^{-1}| = 3\,\mathbf{u}\,|\mathbf{V}_k (D_k U_k)^{-1}| = 3\,\mathbf{u}\,|\mathbf{V}_k L_k^{-T}|$. A similar observation holds for $\Delta_2 \mathbf{r}_k$ (cf. (24)). Apparently, the diagonal scaling in the LU decomposition does not affect the upper bound.

The $\mathbf{r}_j$ basis has the effect that we obtain the following recurrence relation

$$\mathbf{A}\mathbf{R}_k = \mathbf{R}_{k+1} \underline{\widetilde{T}}_k,$$

with $\mathbf{R}_k = \mathbf{V}_k \bar{D}_k$. The diagonal matrix $\bar{D}_k$ has diagonal elements $\bar{d}_{jj} = \|\mathbf{r}_j\|_2$. It is easy to see that

$$\underline{\widetilde{T}}_k = \bar{D}_{k+1}^{-1} \underline{T}_k \bar{D}_k.$$

This means that the actual CG-implementations differs from the Lanczos-based one, analysed in this section, by a similarity scaling with $\bar{D}_k$, and by a slightly different factorization of the matrix. In our tests we have not seen any marked effect on the accuracy by this different scaling. A tentative idea is that although $\bar{D}_k$ is generally ill-conditioned (it represents the norms of the residuals), the diagonal elements are in general only slowly decaying so that successive parts of the tridiagonal matrix are not scaled disproportionally.

Note that for CG we have obtained an upper bound for the floating point rounding errors in the residual, without considering floating point errors in the approximate residual. Therefore, if we want to know how big the error might be in the *true residual* $\mathbf{b} - \mathbf{A}\mathbf{x}_k$, then we have to include the observation that $\mathbf{x}_k$ and $\mathbf{r}_k$ are updated separately, and that the updating of $\mathbf{x}_k$ is not corrected by any floating point error that we make in the updating for $\mathbf{r}_k$.
If we denote by $\widehat{\mathbf{r}}_k$ the updated residual that we obtain in floating point computation, and the computed approximate solution by $\widehat{\mathbf{x}}_k$, then, from [16, §2.1] using the fact that the CG-residuals decrease with respect to the $\mathbf{A}^{-1}$-norm, we have that (see Lemma 8.2):

$$|\,\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2 - \|\widehat{\mathbf{r}}_k\|_2| \leq 2\,k\,\mathbf{u}\,n_A^{3/2}\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_0\|_2 \tag{27}$$

in which n_A represents the (average) number of nonzero elements per row of $\mathbf{A}$.
Since $\widehat{\mathbf{r}}_k = \mathbf{r}_k + \Delta\mathbf{r}_k^{\mathrm{CG}}$, we arrive at the following upper bound for the difference in the computable true residual norm $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2$ and the norm of the (unknown) best residual $\mathbf{r}_k$ that we could have obtained in exact processing with the Lanczos results, then we simply have to add the two upper bounds:

$$\frac{|\,\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2 - \|\mathbf{r}_k\|_2|}{\|\mathbf{r}_0\|_2} \leq \mathbf{u}\left(2\,k\,n_A^{3/2} + 4\sqrt{k} + 2\,\|\mathbf{V}_k\|_2\right)\kappa_2(\mathbf{A}).$$

We see this behavior confirmed in our experiment for CG, in the left top picture of Fig. 3 for the positive definite case.

Since the actual CG-implementations use an LU-decomposition of the matrix $\underline{\widetilde{T}}_k$, instead of a Choleski decomposition, one may use this form of CG also for indefinite systems. Of course, one has no guarantee that the LU-decomposition does not break down, but this is a situation that seldomly occurs. More serious is the fact that $\widetilde{T}_k$ may be close to a singular matrix, for some values of k, which has the effect that the corresponding residual norms of $\mathbf{r}_k$ may be very large (see, for an explanation of this, [12]). Large intermediate residuals have a large effect on the accuracy of the CG-process, as has been shown in [18], and in [4].

The left top picture in Fig. 3 shows the results for CG to a positive definite problem. For the right top picture CG was applied to a non-definite problem. Here we used the classical implementation based coupled 2-term recurrences. Results based on the 3-term Lanczos recurrences were, for the positive definite problem, comparable. For non-definite matrices there were differences, but the differences did not indicate that one implementation is more stable or accurate than the other.

5 Additional errors in the CR-approach

In principle, the CR iterative process can be obtained from the CG process, by carrying out the required inner products in as $\mathbf{A}$-inner products: $(\mathbf{x}, \mathbf{y})_A \equiv (\mathbf{A}\mathbf{x}, \mathbf{y})$. This defines a proper inner product if $\mathbf{A}$ is symmetric positive definite. We will use this observation so that we can carry over the main results from the CG-analysis to the CR situation.
The naive implementation based upon the $\mathbf{A}$-inner product would be too expensive, since it requires 2 inner products per iteration step. For that reason, one works with slightly different implementations in practice. To be more precise, a typical iteration step for CG looks like

$$\begin{aligned}
\mathbf{p}_{j-1} &= \mathbf{r}_{j-1} + \beta_{j-1}\mathbf{p}_{j-2} \;\; ; \;\; \beta_{j-1} = \frac{\mathbf{r}_{j-1}^T\mathbf{r}_{j-1}}{\mathbf{r}_{j-2}^T\mathbf{r}_{j-2}} \\
\mathbf{x}_j &= \mathbf{x}_{j-1} + \alpha_j\mathbf{p}_{j-1} \quad ; \;\; \alpha_j = \frac{\mathbf{r}_{j-1}^T\mathbf{r}_{j-1}}{\mathbf{p}_{j-1}^T\mathbf{A}\mathbf{p}_{j-1}} \\
\mathbf{r}_j &= \mathbf{r}_{j-1} - \alpha_j\mathbf{A}\mathbf{p}_{j-1}
\end{aligned}$$

In contrast, for CR we would see a kernel like

$$\begin{aligned}
\mathbf{A}\mathbf{p}_{j-1} &= \mathbf{A}\mathbf{r}_{j-1} + \beta_{j-1}\mathbf{A}\mathbf{p}_{j-2} \;\; ; \;\; \beta_{j-1} = \frac{\mathbf{r}_{j-1}^T\mathbf{A}\mathbf{r}_{j-1}}{\mathbf{r}_{j-2}^T\mathbf{A}\mathbf{r}_{j-2}} \\
\mathbf{r}_j &= \mathbf{r}_{j-1} - \alpha_j\mathbf{A}\mathbf{p}_{j-1} \quad ; \;\; \alpha_j = \frac{\mathbf{r}_{j-1}^T\mathbf{A}\mathbf{r}_{j-1}}{\mathbf{p}_{j-1}^T\mathbf{A}^T\mathbf{A}\mathbf{p}_{j-1}}
\end{aligned}$$

$$\begin{aligned}
\mathbf{p}_{j-1} &= \mathbf{r}_{j-1} - \beta_{j-1}\mathbf{p}_{j-2} \\
\mathbf{x}_j &= \mathbf{x}_{j-1} + \alpha_j \mathbf{p}_{j-1}
\end{aligned} \tag{28}$$

[19, 6]. Note that the various quantities for CG and CR are different, but we have chosen not to use different notations.

The first two recurrences for CR have the same structure as the first two for CG; the main difference is that in CR the $\mathbf{r}_j$ are updated with an $\mathbf{A}\mathbf{p}_{j-1}$ obtained from a recursion that has been multiplied with $\mathbf{A}$.

As we have seen for CG, that process could be viewed as a solution process with the reduced system on top of a Lanczos process. A closer inspection of the CR process reveals that we have a process on top of a Lanczos process for $\mathbf{A}\mathbf{v}_j$ vectors, rather than $\mathbf{v}_j$ vectors. Namely, if we eliminate the $\mathbf{p}_j$ vectors from the first two recurrences in CR, then we arrive at

$$-\frac{1}{\alpha_j}\mathbf{A}\mathbf{r}_j = \mathbf{A}^2\mathbf{r}_{j-1} - \left(\frac{\beta_{j-1}}{\alpha_{j-1}} + \frac{1}{\alpha_j}\right)\mathbf{A}\mathbf{r}_{j-1} + \frac{\beta_{j-1}}{\alpha_{j-1}}\mathbf{A}\mathbf{r}_{j-2}.$$

In order to be more compatible with the starting point for CG, we rescale the $\mathbf{r}_j$ as

$$\mathbf{v}_j \equiv \frac{1}{\sqrt{\mathbf{r}_j^T\mathbf{A}\mathbf{r}_j}}\mathbf{r}_j,$$

and then the Lanczos recursion can be recast into

$$\mathbf{A}(\mathbf{A}\mathbf{V}_k) = \mathbf{A}\mathbf{V}_{k+1}\underline{T}_k$$

Using the $\mathbf{A}$-orthonormality of the $\mathbf{v}_j$, we see that

$$\mathbf{V}_k^T(\mathbf{A}(\mathbf{A}\mathbf{V}_k)y_k - \mathbf{A}\mathbf{r}_0) = 0,$$

or $T_k y_k = \|\mathbf{r}_0\|_A e_1$.

For the residual $\mathbf{r}_k$, this leads to

$$\mathbf{r}_k = \mathbf{r}_0 - \mathbf{A}\mathbf{V}_k y_k = \mathbf{r}_0 - \|\mathbf{r}_0\|_A(\mathbf{A}\mathbf{V}_k)(T_k^{-1}e_1).$$

Now we can simply take over the results from the CG-analysis, for the upper bound in the floating point perturbation $\Delta\mathbf{r}_k^{\mathrm{CR}}$in the residual, by skipping in (22) and (25) the factor $\|\mathbf{A}\|_2$, and replacing $\|\mathbf{V}_k\|_2$ by $\|\mathbf{A}\mathbf{V}_k\|_2$, $\|\mathbf{r}_0\|_2$ by $\|\mathbf{r}_0\|_A$.

By doing so, we arrive at

$$\|\Delta\mathbf{r}_k^{\mathrm{CR}}\|_2 \le (4\,\|\mathbf{A}\mathbf{V}_k\|_F + 2\,\|\mathbf{A}\mathbf{V}_k\|_2)\,\mathbf{u}\,\|L_k^{-T}\|_2\,\|L_k^{-1}\|_2\,\|\mathbf{r}_0\|_A.$$

From the positive definiteness of $\mathbf{A}$, we have that

$$\|\mathbf{A}\mathbf{V}\|_F \le \|\mathbf{A}^{\frac{1}{2}}\|_2\sqrt{k}, \qquad \|\mathbf{A}\mathbf{V}\|_2 \le \|\mathbf{A}^{\frac{1}{2}}\|_2\,\|\mathbf{A}^{\frac{1}{2}}\mathbf{V}_k\|_2,$$

and $\|\mathbf{r}_0\|_A \le \|\mathbf{A}^{\frac{1}{2}}\|_2\,\|\mathbf{r}_0\|_2$, so that we finally obtain

$$\frac{\|\Delta\mathbf{r}_k^{\mathrm{CR}}\|_2}{\|\mathbf{r}_0\|_2} \le \left(4\sqrt{k} + 2\,\|\mathbf{A}^{\frac{1}{2}}\mathbf{V}_k\|_2\right)\mathbf{u}\,\kappa_2(\mathbf{A}), \tag{29}$$

where we have used again the observation, made for CG, that $\|L_k^{-T}\|_2 \, \|L_k^{-1}\|_2 = \|T_k^{-1}\| \leq \|\mathbf{A}^{-1}\|_2$. Note that for CR the vectors $\mathbf{A}^{\frac{1}{2}}\mathbf{v}_k$ form an orthonormal basis so that this factor behaves the same as the factor $\|\mathbf{V}_k\|_2$ in CG. This implies that (29) represents virtually the same upper bound as for CG.

Note that for CR we have obtained an upper bound for the floating point rounding errors in the residual, without considering floating point errors in the approximate residual. Using the fact the norms of the CR residuals decrease, it can be shown that (27) is also correct for CR (i.e., (27) holds if $\widehat{\mathbf{x}}_k$ is the approximate solution and $\widehat{\mathbf{r}}_k$ is the updated residual obtained by CR in floating point computation; see Lemma 8.2). A combination of (27) and (29) leads to the following upper bound for the difference in the computable true residual norm $\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2$ and the norm of the (unknown) best residual $\mathbf{r}_k$:

$$\frac{|\,\|\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k\|_2 - \|\mathbf{r}_k\|_2|}{\|\mathbf{r}_0\|_2} \leq \mathbf{u}\left(2\,k\,n_A^{3/2} + 4\sqrt{k} + 2\,\|\mathbf{A}^{\frac{1}{2}}\mathbf{V}_k\|_2\right)\kappa_2(\mathbf{A}).$$

We see this behavior confirmed in our experiment for CR, in the left bottom picture of Fig. 3 for the positive definite case. In the right bottom picture CR was applied to a non-definite problem (note the scaling along the vertical axis) and then $\|\mathbf{A}\mathbf{V}_k\|_2$ may be arbitrarily large, if $\mathbf{V}_k^T\mathbf{A}\mathbf{V}_k = \mathbf{I}$, due to the indefiniteness of $\mathbf{A}$.

6 An analysis of the SYMMLQ method

In SYMMLQ we minimize the norm of the error $\mathbf{x} - \mathbf{x}_k$, 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 $\underline{T}_k^T$, but note that this is equivalent with the transpose of the $\underline{Q}_k R_k$ decomposition of $\underline{T}_k$ (see (3), 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:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{x}_0 + \mathbf{A}\mathbf{V}_k R_k^{-1} R_k^{-T}\|\mathbf{r}_0\|_2\, e_1 \\ &= \mathbf{x}_0 + \mathbf{V}_{k+1}\underline{T}_k R_k^{-1} R_k^{-T}\|\mathbf{r}_0\|_2\, e_1 \\ &= \mathbf{x}_0 + (\mathbf{V}_{k+1}\underline{Q}_k)\,(L_k^{-1}\|\mathbf{r}_0\|_2\, e_1), \end{aligned} \tag{30}$$

with $L_k \equiv R_k^T$. The actual implementation of SYMMLQ [13] is based on an update procedure for $\mathbf{W}_k \equiv \mathbf{V}_{k+1}\underline{Q}_k$, and on a three term recurrence relation for $\|\mathbf{r}_0\|_2\, L_k^{-1} e_1$.

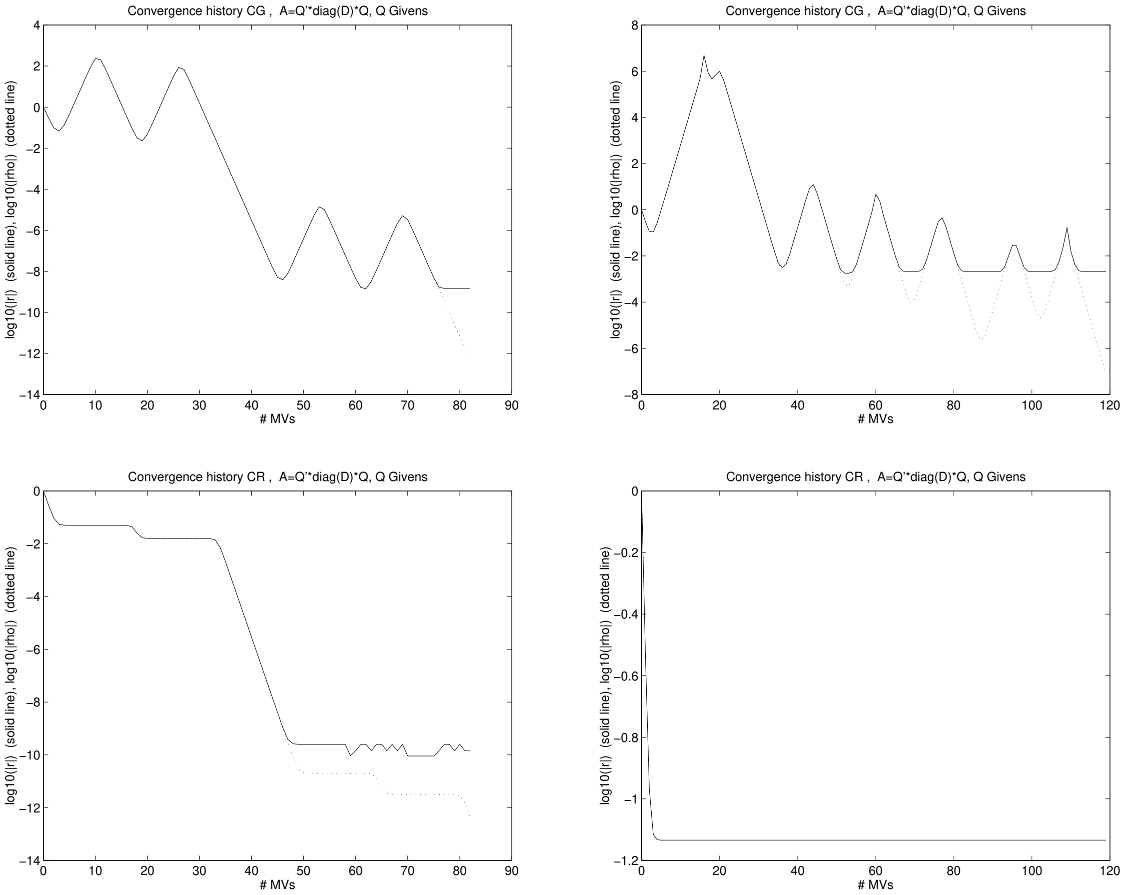


Figure 3. CG *(top) and* CR *(bottom): solid line (—)* $\log_{10}$ *of* $\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\|_2/\|\mathbf{r}_0\|_2$*, dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* $\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2$*, with* $\mathbf{r}_k$ *as computed recursively by the CG and CR algorithms. The pictures show the results for the positive definite system (the left pictures) and for the non-definite system (the right pictures) of Fig. 1. Both systems have condition number* $3\,10^8$*.*

Note that SYMMLQ can be carried out with exactly the same $\mathbf{V}_{k+1}$, $\underline{Q}_k$, and R_k, as for GMRES and MINRES.

In our analysis of the rounding errors, we were hampered by the fact that the residual for the SYMMLQ iterate is in a higher dimensional Krylov subspaces. To be more precise, from (30) it follows that

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

and since $\mathbf{A}\mathbf{V}_{k+1} = \mathbf{V}_{k+2}\underline{T}_{k+2}$, and we see that the residual is in a Krylov subspace of dimension $k+2$, instead of a $k+1$ dimensional subspace.
We have circumvented this problem by first establishing a relation between the norms of the residuals of SYMMLQ and MINRES, which is of interest by itself.

6.1 The 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 $\mathbf{b}-\mathbf{A}\mathbf{x}^{\flat}$, with $\mathbf{x}^{\flat}$ the best approximate of $\mathbf{x}$ in $\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, i.e., $\|\mathbf{x}-\mathbf{x}^{\flat}\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2$ 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 $\mathbf{r}_k^{\mathrm{MR}}$:

$$\frac{\|\mathbf{b}-\mathbf{A}\mathbf{x}^{\flat}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2} \leq \kappa_2(\mathbf{A}). \tag{31}$$

The proof exploits the fact that $\mathbf{r}_k^{\mathrm{MR}} = \mathbf{b}-\mathbf{A}\mathbf{x}_k^{\mathrm{MR}}$ with $\mathbf{x}_k^{\mathrm{MR}}$ in some subspace in which the best approximate $\mathbf{x}^{\flat}$ has been selected, and uses estimates as $\|\mathbf{b}-\mathbf{A}\mathbf{x}^{\flat}\|_2 \leq \|\mathbf{A}\|_2\,\|\mathbf{x}-\mathbf{x}^{\flat}\|_2$ and $\|\mathbf{x}-\mathbf{x}_k^{\mathrm{MR}}\|_2 \leq \|\mathbf{A}^{-1}\|_2\,\|\mathbf{r}_k^{\mathrm{MR}}\|_2$. Proving a similar relation for SYMMLQ and MINRES residuals (as is done in Theorem 6.3) requires some more effort: the approximate $\mathbf{x}_k^{\mathrm{ME}}$ delivered by SYMMLQ is the best approximate from the space $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$. Unfortunately, MINRES selects $\mathbf{x}_k^{\mathrm{MR}}$ from $\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ and this vector may be outside the space $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$.

The following lemma can be used to bound the SYMMLQ error in terms of the MINRES error. Its proof uses the fact that $\mathbf{r}_k^{\mathrm{MR}}$ "links" $\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$ and $\mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$:

$$\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0) = \mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0) \oplus \mathrm{span}(\mathbf{r}_k^{\mathrm{MR}})\,, \tag{32}$$

that is, $\mathbf{r}_k^{\mathrm{MR}} \perp \mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ and $\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$ is spanned by $\mathbf{r}_k^{\mathrm{MR}}$ and $\mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$.

Lemma 6.1 *For each* $\mathbf{z}\in\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$, *we have*

$$\|\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}\|_2^2 \leq \|\mathbf{x}-\mathbf{z}\|_2^2 + |\alpha_k|^2\,\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2 \quad \textit{where} \quad \alpha_k \equiv \frac{(\mathbf{x},\mathbf{r}_k^{\mathrm{MR}})}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2}. \tag{33}$$

Proof. Observe that $\mathbf{x}_k^{\mathrm{ME}}$ minimizes $\|\mathbf{x}-\mathbf{z}\|_2$ over all $\mathbf{z}$ in the space $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$. Hence $\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}} \perp \mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$. Moreover, (cf. (32)),

$$\alpha_k = (\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}},\mathbf{r}_k^{\mathrm{MR}})/\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2 \quad \text{and} \quad \mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}} \perp \mathbf{r}_k^{\mathrm{MR}}. \tag{34}$$

Therefore, (32) implies that

$$\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}} \perp \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0) \quad \text{and} \quad \mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0).$$

Consequently,

$$\|\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}-\alpha_k\mathbf{r}_k^{\mathrm{MR}}\|_2 \leq \|\mathbf{x}-\mathbf{z}\|_2 \quad \text{for all} \quad \mathbf{z}\in\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0) \tag{35}$$

and (33) follows by combining Pythagoras' theorem (cf. (34)) and (35). □

Note that $\alpha_k = (\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}},\mathbf{r}_k^{\mathrm{MR}}/\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2)$. Unfortunately, a combination of (33) with $\mathbf{z}=\mathbf{x}_k^{\mathrm{MR}}$ and the obvious estimate $|\alpha_k|\,\|\mathbf{r}_k^{\mathrm{MR}}\|_2 \leq \|\mathbf{x}-\mathbf{x}_k^{\mathrm{ME}}\|_2$ does not lead to an interesting result. A useful result follows from an upper bound for $|\alpha_k|$ obtained by exploiting a relation between two consecutive MINRES residuals and a Lanczos vector, stated in the Lemma 6.2 (see (36)). Actually, the MINRES residual $\mathbf{r}_k^{\mathrm{MR}}$ is a convex combination of the MINRES residual $\mathbf{r}_{k-1}^{\mathrm{MR}}$ and the CG residual $\mathbf{r}_k^{\mathrm{CG}}$: $\mathbf{r}_k^{\mathrm{MR}} = s^2\,\mathbf{r}_{k-1}^{\mathrm{MR}} + c^2\,\mathbf{r}_k^{\mathrm{CG}}$ (see, e.g., [2, 22, 17]; a detailed proof can also be found in the Appendix). Since $\mathbf{r}_k^{\mathrm{CG}}$ is a multiple of the $k+1$th Lanczos vector $\mathbf{v}_{k+1}$, the lemma follows.

Lemma 6.2 *For some scalar factor* γ *we have*

$$\mathbf{r}_k^{\mathrm{MR}} = s^2\,\mathbf{r}_{k-1}^{\mathrm{MR}} + \gamma\mathbf{v}_{k+1} \quad \textit{where} \quad s \equiv \frac{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2}. \tag{36}$$

Theorem 6.3

$$\|\mathbf{r}_k^{\rm ME}\|_2 \leq \nu_{k+1}\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_k^{\rm MR}\|_2 \quad \textit{with} \quad \nu_k \equiv k + \tfrac{1}{2}\ln(k). \tag{37}$$

Proof. A combination of (33) and an upper bound for $|\alpha_k|$ (see (38)) will lead to (37). We first consider relation (36).

Since $\gamma\mathbf{v}_{k+1} \perp \mathbf{r}_{k-1}^{\rm MR} \in \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, it follows that $\|\gamma\mathbf{v}_{k+1}\|_2 \leq \|\mathbf{r}_k^{\rm MR}\|_2$. Moreover, since $\mathbf{r}_{k-1}^{\rm MR} \perp \mathbf{A}\,\mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0)$ and $\gamma\mathbf{v}_{k+1} \perp \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, we have $\mathbf{r}_{k-1}^{\rm MR} \perp \mathbf{x}_{k-1}^{\rm ME}$ and $\gamma\mathbf{v}_{k+1} \perp \mathbf{x}_k^{\rm MR}$. Therefore, with $\mathbf{e}_j^{\rm ME} \equiv \mathbf{x} - \mathbf{x}_j^{\rm ME}$, (36) implies

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

and hence

$$|\alpha_k| \leq \frac{\|\mathbf{r}_k^{\rm MR}\|_2}{\|\mathbf{r}_{k-1}^{\rm MR}\|_2}\,\|\mathbf{e}_{k-1}^{\rm ME}\| + \|\mathbf{x} - \mathbf{x}_k^{\rm MR}\|\,. \tag{38}$$

A combination of (38) and (33) with $\mathbf{z} = \mathbf{x}_{k+1}^{\rm MR}$ leads to

$$\frac{\|\mathbf{e}_k^{\rm ME}\|_2^2}{\|\mathbf{r}_k^{\rm MR}\|_2^2} \leq \frac{\|\mathbf{x} - \mathbf{x}_{k+1}^{\rm MR}\|_2^2}{\|\mathbf{r}_k^{\rm MR}\|_2^2} + \left(\frac{\|\mathbf{e}_{k-1}^{\rm ME}\|_2}{\|\mathbf{r}_{k-1}^{\rm MR}\|_2} + \frac{\|\mathbf{x} - \mathbf{x}_k^{\rm MR}\|_2}{\|\mathbf{r}_k^{\rm MR}\|_2}\right)^2. \tag{39}$$

With

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

using the minimal residual property $\|\mathbf{r}_{k+1}^{\rm MR}\|_2 \leq \|\mathbf{r}_k^{\rm MR}\|_2$, we obtain the following recursive upper bound from (39):

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

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

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

this completes the proof of the theorem. □

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

Theorem 6.4 *Consider some vectors* $\mathbf{y}$ *and* $\mathbf{c}$ *with* $\mathbf{c} = \mathbf{A}\mathbf{y}$.
Then, for the best approximation $\mathbf{y}_k^{\rm ME}$ *of* $\mathbf{y}$ *in* $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$*, and for* $\mathbf{y}_k^{\rm MR} \in \mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ *such that* $\mathbf{A}\mathbf{y}_k^{\rm MR}$ *is the best approximation of* $\mathbf{c}$ *in* $\mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$*, with* ν_k *as in* (37)*, we have*

$$\frac{\|\mathbf{c} - \mathbf{A}\,\mathbf{y}_k^{\rm ME}\|_2}{\|\mathbf{r}_k^{\rm MR}\|_2} \leq \nu_{k+1}\,\kappa_2(\mathbf{A})\,\mu_k, \quad \textit{where} \quad \mu_k \equiv \sup_{i\leq k}\frac{\|\mathbf{c} - \mathbf{A}\,\mathbf{y}_i^{\rm MR}\|_2}{\|\mathbf{r}_i^{\rm MR}\|_2}. \tag{40}$$

Proof. The proof comes along the same lines as the proof of Theorem 6.3. 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, (33) 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

$$\frac{\|\mathbf{e}_k^{\mathrm{ME}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} \leq \frac{\|\mathbf{y} - \mathbf{y}_{k+1}^{\mathrm{MR}}\|_2^2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2^2} + \left(\frac{\|\mathbf{e}_{k-1}^{\mathrm{ME}}\|_2}{\|\mathbf{r}_{k-1}^{\mathrm{MR}}\|_2} + \frac{\|\mathbf{y} - \mathbf{y}_k^{\mathrm{MR}}\|_2}{\|\mathbf{r}_k^{\mathrm{MR}}\|_2}\right)^2. \tag{41}$$

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 (40). □

6.2 Additional errors introduced by SYMMLQ

The error in the SYMMLQ approximation $\mathbf{x}_k$ equals

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

g_k is the exact solution and $\widehat{g}_k$ is the solution in finite precision arithmetic. We denote the coordinates of $g_k/\|\mathbf{r}_0\|_2$ by γ_j, $g_k/\|\mathbf{r}_0\|_2 = (\gamma_1, \ldots, \gamma_k)^T$, and the ones of $\widehat{g}_k/\|\mathbf{r}_0\|_2$ by $\widehat{\gamma}_j$. Observe that coordinates γ_j and $\widehat{\gamma}_j$ are independent of k (for $j \leq k$). Furthermore,

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

With $L_{k+2} = (\ell_{ij})$, some manipulation shows that

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

where

$$M_k \equiv \begin{bmatrix} \ell_{k+1\,k-1} & \ell_{k+1\,k} \\ 0 & \ell_{k+2\,k} \end{bmatrix}. \tag{42}$$

Hence, the error in the SYMMLQ residual $\mathbf{r}_k^{\mathrm{ME}}$ equals

$$\mathbf{V}_k L_k(\widehat{g}_k - g_k) + \|\mathbf{r}_0\|_2\,[\mathbf{v}_{k+1}, \mathbf{v}_{k+2}]\, M_k \begin{bmatrix} \widehat{\gamma}_{k-1} - \gamma_{k-1} \\ \widehat{\gamma}_k - \gamma_k \end{bmatrix}. \tag{43}$$

The first term can be treated as in GMRES:

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

With

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

expression (43) can sharply be bounded by

$$3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_0\|_2 + \|\widehat{t}_k - t_k\|_2\,\|\mathbf{r}_0\|_2. \tag{44}$$

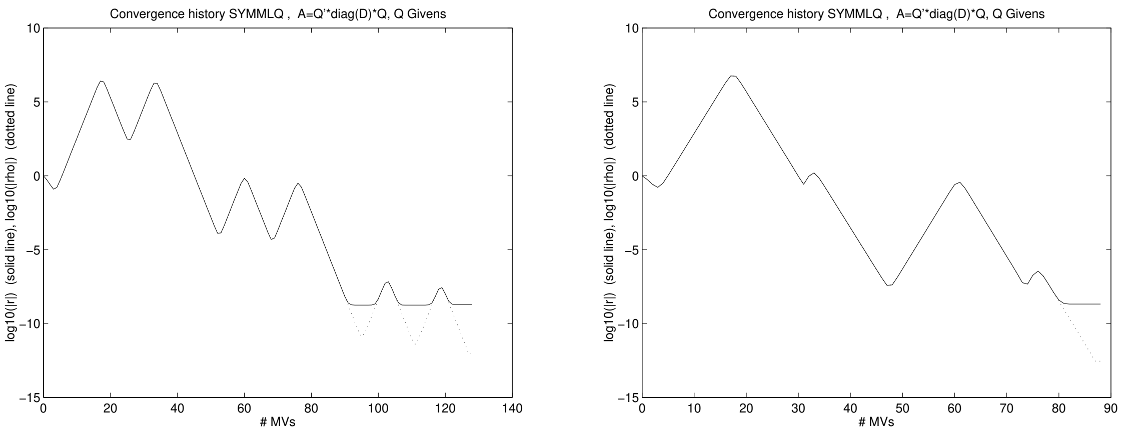


Figure 4. SYMMLQ*: solid line (—)* $\log_{10}$ *of* $\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\|_2/\|\mathbf{r}_0\|_2$*, dotted line* $(\cdots)$ $\log_{10}$ *of the estimated residual norm reduction* $\|\widehat{t}_k\|_2$*. The pictures show the results for the positive definite system (the left picture) and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

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

$$\frac{\|\Delta\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(\mathbf{A}) \;+\; \|\widehat{t}_k - t_k\|_2. \tag{45}$$

A straight-forward estimate yields

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

which is much worse than the first term in (45). 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. (48)). Fig. 4 illustrates that the upper bound in (45), with $\|\widehat{t}_k - t_k\|_2 \approx 0$, is fairly sharp.

Note that, in exact arithmetic, $\|\mathbf{r}_0\|_2\,\|t_k\|_2 = \|\mathbf{r}_k\|_2$ is the norm of the kth SYMMLQ residual. 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.

Accuracy. From (45) it follows that

$$\left| \|\widehat{t}_k\|_2 - \frac{\|\widehat{\mathbf{r}}_k\|_2}{\|\mathbf{r}_0\|_2} \right| \leq 3\sqrt{3}\,\mathbf{u}\,\|\mathbf{V}_k\|_2\,\kappa_2(\mathbf{A}) + 2\,\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2}, \tag{46}$$

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 may be called accurate: recall that, for any method, a loss of accuracy of order $\mathbf{u}\,\kappa_2(\mathbf{A})$ can considered to be acceptable.

Convergence. It is not clear yet whether the convergence of SYMMLQ is insensitive to rounding errors. This would follow from (45) 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 $\|t_k\|_2 \lesssim \|\widetilde{t}_k\|_2$. 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, in view of the results in §6.1, Theorem 6.3, it may be anticipated that the interesting quantities can be bounded in terms of the MINRES residual.

Recall that the results in §6.1 assumed exact arithmetic, that is an exact Lanczos process as well as an exact solve of the "L-equations". However, Theorem 6.3 holds also for systems involving T_m ($m > k$) and $\kappa_2(\mathbf{A})$ can be replaced by $\kappa_2(L_k)$. In this setting,

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

with s_j the sines in the jth Givens rotations employed in the QR decomposition of $\underline{T}_k$: ρ_k is the estimated reduction of the norms of the MINRES residuals. Moreover $\kappa_2(L_k) \lesssim \kappa_2(\mathbf{A})$. Therefore,

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

In §6.2.1, we will show that, with $\mu = 5$,

$$\|t_k - \widehat{t}_k\|_2 \leq \mu\,\mathbf{u}\,\rho_k\,\kappa_2(\mathbf{A})^2\,\left(\tfrac{1}{6}k^3 + \mathcal{O}(k^2\ln k)\right). \tag{48}$$

The upper bound in (48) 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. Note that the upper bound for $\|\widehat{t}_k - t_k\|_2$ differs from the bound for $\|t_k\|_2$ in order of $\mathbf{u}\,\kappa_2(\mathbf{A})$.

Remark 6.5 Except for the constants '$k + \mathcal{O}(k)$' and '$\frac{1}{6}k^3 + \mathcal{O}(k^2\ln k)$', the estimates (47) and (48), 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 (48) (cf. §6.2.1), upper bounds as in (47) are used in a consecutive number of steps. In view of the irregular convergence of SYMMLQ, the upper bound (47) will be sharp for at most a few steps. By exploiting this observation, one can show that a growth of order k (or k^2) will be more likely.

6.2.1 SYMMLQ recurrences

In this section we derive the upper bound (48).

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

$$\delta = \ell_{jj}\gamma_j^\star + \ell_{jj-1}\gamma_{j-1} + \ell_{jj-2}\gamma_{j-2} \quad \text{with} \quad |\delta| \leq \mu\,\mathbf{u}\,|\ell_{jj}|\,|\gamma_j|. \tag{49}$$

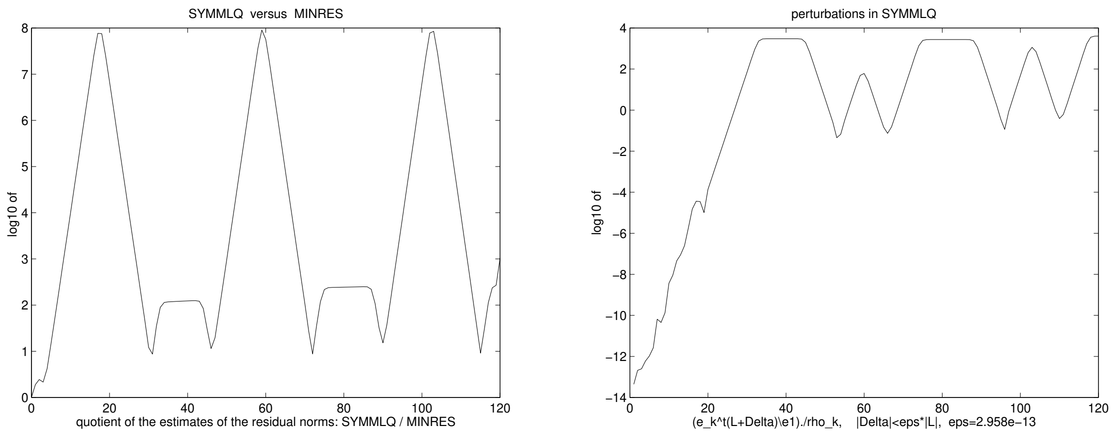


FIGURE 5. *Results for the non-definite matrix with condition number* $3\,10^8$ *(as in the right pictures) of* Fig. 1, Fig. 3, *and* Fig. 4. *The left picture shows* $\log_{10}$ *of the ratio* $\|\widehat{t}_k\|_2/\rho_k$ *of the estimated residual norm reduction of* SYMMLQ *with the one of* MINRES,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

The resulting perturbed quantities are labeled with $\cdot^\star$.
Then

$$t_k^\star - t_k = \delta\, M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_j. \tag{50}$$

For $j=1$, $t_k^\star - t_k$ is a multiple of the SYMMLQ residual for the T_m-system ($m>k$) and, as in the proof of Inequality (47), Theorem 6.3 could be applied to estimate $\|t_k^\star - t_k\|_2$. For the situation where $j \neq 1$, Theorem 6.4 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 6.4), for $j<k$,

$$\mathbf{y}_j^{\mathrm{ME}} = \mathbf{0}, \qquad \|\mathbf{c} - \mathbf{A}\mathbf{y}_k^{\mathrm{ME}}\|_2 = \|M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_{j+1}\|_2, \tag{51}$$

and

$$\|\mathbf{c} - \mathbf{A}\mathbf{y}_k^{\mathrm{MR}}\|_2 = c_j\,\frac{\rho_k}{\rho_j}, \tag{52}$$

with c_j the cosine in the jth Givens rotation. Therefore, by Theorem 6.4,

$$\|M_k \begin{bmatrix} e_{k-1}^T \\ e_k^T \end{bmatrix} L_k^{-1} e_{j+1}\|_2 \leq \kappa_2(L_k)\, c_j\, \nu_{k+1}\,\frac{\rho_k}{\rho_j}. \tag{53}$$

For this specific situation, the estimate for β_k in the last paragraph of the proof of Theorem 6.3 can be improved. It can be shown that $\beta_j \leq 1$, whence $\beta_k \leq \nu_{k-j}$. Therefore, the ν_{k+1} in (53) can be replaced by ν_{k-j}.

A combination of (50) with (53) gives

$$\|t_k^\star - t_k\|_2 \leq \frac{|\delta|}{|\rho_{j-1}|}\,\rho_k\,\kappa_2(L_k)\,\nu_{k-j+1} \lesssim \frac{|\delta|}{|\rho_{j-1}|}\,\rho_k\,\kappa_2(\mathbf{A})\,\nu_{k-j+1}. \tag{54}$$

Now, observe that

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

Therefore, from (47), we have that

$$|\ell_{jj}|\,\frac{|\gamma_j|}{\rho_{j-1}} \leq \frac{\|t_{j-1}\|_2}{\rho_{j-1}} \leq \kappa_2(\mathbf{A})\,\nu_j. \tag{55}$$

Hence (cf. (49))

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

and, with (54), this gives gives

$$\|t_k^\star - t_k\|_2 \leq \mu\,\mathbf{u}\,\rho_k\,\kappa_2(\mathbf{A})^2\,\nu_j\nu_{k-j+1}. \tag{56}$$

The recurrences are linear. Therefore, the effect of a number of perturbations is the cumulation of the effect of single perturbations. If each recurrence relation is perturbed as in (49), then the estimate (48) appears as a cumulation of bounds as in (56): now, the tilded quantities in (48) refer to the resulting perturbed quantities.

Finally, we will argue that the effect of rounding errors in solving $L^{-1}e_1$ with the recurrence relation can be described as indicated with $\mu = 5$.
Observe that the $\gamma_k^\star$'s resulting from the perturbation

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

(all operations exact) are the same as the ones resulting from the perturbation

$$\ell_{j-1j-1}\gamma_{j-1}^\star(1+\mu\,\xi) + \ell_{j-1j-2}\gamma_{j-2} + \ell_{j-1j-3}\gamma_{j-3} = 0$$

(again all other operations exact) which fits in the description (49). If $\widehat{\gamma}_j$ are the computed γ_j (in rounded arithmetic) then

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

whence

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

7 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 more serious point of concern; it has been the main theme

of this study. In our study we have restricted ourselves to symmetric 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), SYMMLQ (§6), and CG (§4), lead to acceptable approximate solutions. When these methods converge then the relative error in the approximate solution is, apart from modest factors, bounded by $\mathbf{u}\,\kappa_2(\mathbf{A})$. It should be noted that application of CG is limited to positive definite matrices $\mathbf{A}$. For indefinite systems, the relative errors with CG can be much larger, due to possibly large intermediate residuals, or close to singular reduced systems T_k. 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 6.3. For ill-conditioned systems this may be considerable. Because of the economy of storage, we conclude that for *positive definite systems* CG is the best choice.
For *indefinite symmetric systems* we see that MINRES (and CR) may lead to large perturbation errors: for MINRES the upper bound contains a factor $\kappa_2(\mathbf{A})^2$ (§3), for CR (§5) the situation may be even worse because of the indefiniteness of the reduced system (near-break down situations). 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 §3.1, 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] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, To appear in SIAM J. Matrix Anal. Appl. (1997).

[5] 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.

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

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

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

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

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

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

[12] 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.

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

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

[15] 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.

[16] 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.

[17] G.L.G. Sleijpen and H.A. van der Vorst, *Maintaining convergence properties of BiCGstab methods in finite precision arithmetic*, Numerical Algor., 10 (1995), pp. 203–223.

[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.

8 Appendix

Lemma 8.1 *If, for a matrix* $\mathbf{C}$*,* n_C *is the maximum number of non-zero's per column, then*

$$\|\,|\mathbf{C}|\,\|_2 \leq \sqrt{n_C}\,\|\mathbf{C}\|_2. \tag{57}$$

Proof. Since $\|\,|\mathbf{C}|\,\|_2^2 \leq n_C \max_j \left(\sum_i |c_{ij}|^2\right)$ (see [21, Th. 4.2]), we have

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

Lemma 8.2 *For the computed approximate solution* $\widehat{\mathbf{x}}_k$ *and the updated residual* $\widehat{\mathbf{r}}_k$ *that we obtain in floating point computation with CG or with CR, we have that*

$$\|(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k) - \widehat{\mathbf{r}}_k\|_2 \leq 2\,k\,\mathbf{u}\,n_A^{3/2}\,\kappa_2(\mathbf{A})\,\|\mathbf{r}_0\|_2.$$

Proof. In [16, §2.1], it has been shown that (see also [4])

$$\|(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_k) - \widehat{\mathbf{r}}_k\|_2 \leq 2\,k\,\mathbf{u}\,n_A\,\|\,|\mathbf{A}|\,\|_2 \max_{j\leq k} \|\mathbf{x} - \mathbf{x}_k\|_2.$$

The CG residuals decrease with respect to the $\mathbf{A}^{-1}$-norm, while the CR residuals decrease with respect to the Euclidian norm: $\|\mathbf{r}_k\|_{A^{2\gamma-2}} \leq \|\mathbf{r}_0\|_{A^{2\gamma-2}}$, with $\gamma = \frac{1}{2}$ for CG, and with $\gamma = 1$ for CR. Hence

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_k\|_2 &\leq \|\mathbf{A}^{-\gamma}\|_2\,\|\mathbf{x} - \mathbf{x}_k\|_{A^{2\gamma}} \leq \|\mathbf{A}^{-\gamma}\|_2\,\|\mathbf{r}_k\|_{A^{2\gamma-2}} \\ &\leq \|\mathbf{A}^{-\gamma}\|_2\,\|\mathbf{r}_0\|_{A^{2\gamma-2}} \leq \|\mathbf{A}^{-\gamma}\|_2\,\|\mathbf{A}^{\gamma-1}\|_2\,\|\mathbf{r}_0\|_2 = \|\mathbf{A}^{-1}\|_2\,\|\mathbf{r}_0\|_2, \end{aligned}$$

and the lemma follows from an induction argument and (57). $\square$

Proof of Lemma 6.2. If $\mathbf{w} \in \mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ is orthogonal to $\mathbf{A}\,\mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0)$, then

$$\mathbf{A}\,\mathcal{K}_k(\mathbf{A};\mathbf{r}_0) = \mathbf{A}\,\mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0) \oplus \mathrm{span}(\mathbf{w})\,. \tag{58}$$

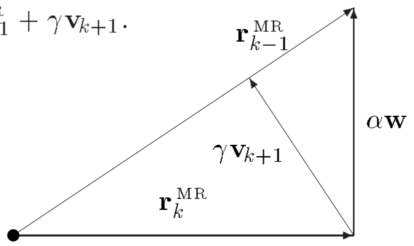
With $\beta := (\mathbf{r}_{k-1}^{\mathrm{MR}}, \mathbf{r}_{k-1}^{\mathrm{MR}})/(\mathbf{w}, \mathbf{r}_{k-1}^{\mathrm{MR}})$ we have $\mathbf{r}_{k-1}^{\mathrm{MR}} - \beta\,\mathbf{w} \perp \mathbf{r}_{k-1}^{\mathrm{MR}}$. Since $\mathbf{r}_{k-1}^{\mathrm{MR}} - \beta\,\mathbf{w} \perp \mathbf{A}\,\mathcal{K}_{k-1}(\mathbf{A};\mathbf{r}_0)$, $\mathbf{r}_{k-1}^{\mathrm{MR}} - \beta\,\mathbf{w} \in \mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0)$, and $\mathcal{K}_{k+1}(\mathbf{A};\mathbf{r}_0) = \mathcal{K}_k(\mathbf{A};\mathbf{r}_0) \oplus \mathrm{span}(\mathbf{v}_{k+1})$, identity (32) (with $j = k-1$) implies

$$\mathbf{r}_{k-1}^{\mathrm{MR}} - \beta\,\mathbf{w} \in \mathrm{span}(\mathbf{v}_{k+1})\,.$$

A similar argument, using (58) and (32), shows that, with $\alpha := (\mathbf{r}_{k-1}^{\mathrm{MR}}, \mathbf{w})/(\mathbf{w}, \mathbf{w})$, we have $\mathbf{r}_k^{\mathrm{MR}} = \mathbf{r}_{k-1}^{\mathrm{MR}} - \alpha\,\mathbf{w}$. Therefore, with scalars γ' such that $\gamma'\mathbf{v}_{k+1} := \mathbf{r}_{k-1}^{\mathrm{MR}} - \beta\,\mathbf{w}$ and $\gamma := \frac{\alpha}{\beta}\gamma'$, we see that

$$\mathbf{r}_k^{\mathrm{MR}} = \mathbf{r}_{k-1}^{\mathrm{MR}} - \frac{\alpha}{\beta}(\mathbf{r}_{k-1}^{\mathrm{MR}} - \gamma'\mathbf{v}_{k+1}) = \left(1 - \frac{\alpha}{\beta}\right)\mathbf{r}_{k-1}^{\mathrm{MR}} + \gamma\mathbf{v}_{k+1}.$$

A simple geometrical argument, considering the cosine of the angle between $\mathbf{r}_k^{\mathrm{MR}}$ and $\mathbf{r}_{k-1}^{\mathrm{MR}}$ (see the picture to the right) shows that $s^2 = 1 - \alpha/\beta$. $\square$



Proof of (51). Consider $\mathbf{c} := \mathbf{v}_{j+1}$ and $\mathbf{y}$ such that $\mathbf{c} = \mathbf{A}\mathbf{y}$.
Consider $\mathbf{y}_k^{\mathrm{ME}} \in \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$ for which $\|\mathbf{y} - \mathbf{y}_k^{\mathrm{ME}}\|$ is minimal.
Then $\mathbf{y} - \mathbf{y}_k^{\mathrm{ME}} \perp \mathbf{A}\mathcal{K}_k(\mathbf{A};\mathbf{r}_0)$, or, equivalently, with $\mathbf{y}_k^{\mathrm{ME}} = \mathbf{A}\mathbf{V}_k g_k$,

$$0 = (\mathbf{A}\mathbf{V}_k)^T\mathbf{w} - (\mathbf{A}\mathbf{V}_k)^T\mathbf{A}\mathbf{V}_k\,g_k = e_{j+1} - \underline{T}_k^T\,\underline{T}_k\,g_k \quad \text{if} \quad j+1 \le k$$

and

$$0 = -\underline{T}_k^T\,\underline{T}_k\,g_k \quad \text{if} \quad j+1 > k.$$

In particular $\mathbf{y}_k^{\mathrm{ME}} = \mathbf{0}$ if $j+1 > k$. $\square$

CG

```
Choose x_0
x = x_0, r = b - Ax
u = 0, ρ = 1
while ||r|| > tol do
    σ ← -ρ
    ρ ← (r, r), β ← ρ/σ
    u ← r - β u
    c ← Au
    σ ← (c, u), α ← ρ/σ
    x ← x + α u
    r ← r - α c
end while
```

MINRES

```
Choose x_0
```

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

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

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

while $|\rho| > \mathit{tol}$ do

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

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

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

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

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

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

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

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

end while