Numer. Math. (2005) 100: 711–733
Digital Object Identifier (DOI) 10.1007/s00211-005-0603-8

Numerische Mathematik

The effect of non-optimal bases on the convergence of Krylov subspace methods

Valeria Simoncini[1], **Daniel B. Szyld**[2]

[1] Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, 40127 Bologna, Italy; and also IMATI-CNR, Pavia and CIRSA, Ravenna, Italy; e-mail: valeria@dm.unibo.it

[2] Department of Mathematics, Temple University (038-16), 1805 N. Broad Street, Philadelphia, Pennsylvania 19122-6094, USA; e-mail: szyld@math.temple.edu

Received June 23, 2004 / Revised version received January 3, 2005 /
Published online: May 25, 2005 –

Summary. There are many examples where non-orthogonality of a basis for Krylov subspace methods arises naturally. These methods usually require less storage or computational effort per iteration than methods using an orthonormal basis (*optimal* methods), but the convergence may be *delayed*. Truncated Krylov subspace methods and other examples of *non-optimal* methods have been shown to converge in many situations, often with small delay, but not in others. We explore the question of what is the effect of having a non-optimal basis. We prove certain identities for the relative residual gap, i.e., the relative difference between the residuals of the optimal and non-optimal methods. These identities and related bounds provide insight into when the delay is small and convergence is achieved. Further understanding is gained by using a general theory of superlinear convergence recently developed. Our analysis confirms the observed fact that in exact arithmetic the orthogonality of the basis is not important, only the need to maintain linear independence is. Numerical examples illustrate our theoretical results.

Mathematics Subject Classification (2000): 65F10, 65F15, 15A06, 15A18

1 Introduction

Krylov subspace methods for the solution of $n \times n$ nonsymmetric linear systems of equations of the form $Ax = b$ are extensively used nowadays. Let

Correspondence to: D.B. Szyld

$r_0 = b - Ax_0$ be the initial residual, with x_0 an initial approximation to the solution, and let $\mathcal{K}_m(A, r_0) = \text{span}\{r_0, Ar_0, A^2r_0, \ldots, A^{m-1}r_0\}$ be the Krylov subspace of dimension m defined by A and r_0. If $\{w_1, \ldots, w_m\}$ is a basis of $\mathcal{K}_m(A, r_0)$, let $W_m = [w_1, \ldots, w_m]$; then the new approximation x_m to the solution of the linear system is taken to be of the form

$$x_m = x_0 + W_m y_m \tag{1.1}$$

for some $y_m \in \mathbb{R}^m$ determined so that the associated residual $r_m = b - Ax_m$ satisfies some constraint such as Galerkin or Petrov-Galerkin condition; see, e.g., [2], [13], [27], [34]. Whenever W_m has orthonormal columns, we call the corresponding schemes *optimal* methods. In order to save either storage or computational effort, there are many methods where a non-orthogonal basis of $\mathcal{K}_m(A, r_0)$ is used. We call them *non-optimal* methods; see section 2 for examples. This can be carried out for instance by either relaxing the residual constraint or by specifically choosing the auxiliary spaces within the Petrov-Galerkin condition.

In this paper we analyze the convergence of non-optimal Krylov subspace methods as compared to the optimal ones. While these non-optimal methods are widely used, there is little analysis of them in the literature. Even though we do not provide a complete picture of their behavior, we present several equalities and some bounds which are a first attempt at their understanding. We study the residual gap, i.e., the norm of the difference between the residuals of the optimal and non-optimal methods, and in sections 3 and 4 we obtain new identities for this relative residual gap. In particular, our analysis confirms the fact that as long as the new element of the non-orthogonal basis has a sufficiently large angle with the previous subspace, these methods will work well in exact arithmetic; orthogonality is not necessary. We are motivated in part by the observation that when a very good preconditioner is available, or more generally, when fast convergence of the optimal Krylov subspace method is observed, the non-optimal methods may be very competitive: the savings in storage and computational effort more than compensate the delay in convergence due to the use of a non-orthogonal basis. We remark that fast convergence here does not necessarily mean just a few, say 10, iterations, as illustrated with several examples throughout the paper; see in particular Remark 4.1. Furthermore, we show that the general theory of superlinear convergence of Krylov methods recently developed in [33] applies to these non-optimal methods as well, providing us with further understanding of the problem settings where we can expect these methods to behave efficiently (section 5). Although our analysis assumes exact arithmetic, we present several numerical examples illustrating the theoretical results.

We caution the reader that we do not advocate the use of the non-optimal methods for all problems, and in fact they fail in many cases (see, e.g., Example 5.4), but we do believe they have a role to play in the roster of methods for

the solution of nonsymmetric problems. We hope that the new understanding we provide here helps finding their proper role.

2 Notation, descriptions, and examples

Given a vector w_1 satisfying $w_1 = \alpha r_0$ for some nonzero scalar α, we consider the following general recurrence

$$AW_m = W_{m+1}T_{m+1,m} = W_m T_m + \tau_{m+1,m} w_{m+1} e_m^*, \tag{2.1}$$

where W_m is an $n \times m$ full column rank matrix, whose columns span $\mathcal{K}_m(A, r_0)$. Here and throughout the paper e_m indicates the mth column of the identity matrix of appropriate order, depending on the context; when this order needs to be specified we write I_m for the identity in the $m \times m$ case. The symbol e^* denotes the transpose of e. We emphasize that here the matrix W_{m+1} satisfying (2.1) is not required to have orthonormal columns but only to be full column rank, i.e., that its columns span the subspace $\mathcal{K}_{m+1}(A, r_0)$. The matrix $T_{m+1,m} \in \mathbb{R}^{(m+1)\times m}$ here is upper Hessenberg, and in general we are interested in the case in which $T_{m+1,m}$ is also banded. The matrix T_m is the $m \times m$ principal submatrix of $T_{m+1,m}$.

We concentrate in this paper on two classes of approximation methods. In the first class the vector y_m in (1.1) solves the linear system

$$T_m y_m = \beta e_1, \tag{2.2}$$

where $\beta = \|r_0\|$. The procedure above corresponds to the FOM method whenever W_m has orthonormal columns, and it corresponds to the Lanczos method when in addition A is symmetric. A second class of methods is obtained by requiring that y_m is the minimizer in

$$\min_y \|\beta e_1 - T_{m+1,m} y\|, \tag{2.3}$$

i.e., $y_m = T_{m+1,m}^+ \beta e_1$, where $T_{m+1,m}^+ = (T_{m+1,m}^* T_{m+1,m})^{-1} T_{m+1,m}^*$ is the pseudoinverse of $T_{m+1,m}$; see, e.g., [11]. When W_m has orthonormal columns, this corresponds to the GMRES method (or MINRES in the symmetric case).

We formally call *a non-optimal Krylov subspace method* a method that produces a non-optimal (or non-orthogonal) basis of a Krylov subspace as in (2.1), where $T_{m+1,m}$ is upper Hessenberg with a possibly banded structure. Several methods in the literature can be recast in terms of (2.1), such as the nonsymmetric Lanczos process, truncated methods, restarted FOM and restarted GMRES, and so on. For further details on these methods, see, e.g., [2], [13], [27], [34], and the discussion below.

Our aim is to relate these non-optimal approaches to the *optimal* methods that explicitly generate an orthonormal basis. In this latter class fall the (full)

FOM and GMRES methods, which will be used as reference optimal methods throughout the paper. We remark that strictly speaking, the nonsymmetric Lanczos method is an "optimal" method in the sense that the generated vectors satisfy a global Petrov-Galerkin variational property. Nonetheless, the method does not construct an orthonormal basis, and thus for the definition we have used of *optimal*, we will include it in the much wider class of schemes satisfying (2.1).

Our first motivating example relates to a symmetric and positive definite system $\widehat{A}\hat{x} = \hat{b}$ which is unsymmetrically preconditioned. General preconditioning consists of replacing this system with an equivalent one, e.g., of the form $Ax = b$ with $A = L^{-1}\widehat{A}L^{-*}$, $b = L^{-1}\hat{b}$, and $\hat{x} = L^{-*}x$, for some easily invertible matrix L. We consider the case where the application problem is such that a *nonsymmetric* preconditioner is employed, either because of inherent functional properties or because of the chosen preconditioning strategy. In this case, let P be the nonsymmetric preconditioner and let $P = LU$ be its LU factorization. If P is not too far from being symmetric, one can still think of applying the symmetric Lanczos process to the preconditioned problem

$$Ax := L^{-1}\widehat{A}U^{-1}x = L^{-1}\hat{b}, \tag{2.4}$$

as if the preconditioned matrix were symmetric; cf. [4]. By doing so, after m steps we obtain the matrices W_{m+1}, $T_{m+1,m}$ satisfying (2.1), where T_m is $m \times m$ symmetric and tridiagonal, while the matrix $W_{m+1} = [W_m, w_{m+1}] \in \mathbb{R}^{n\times(m+1)}$ has *locally* orthogonal columns, but in general, not all columns of W_{m+1} are orthogonal. Thus, to solve the system (2.4) one can consider two alternatives. One can use an optimal method for this nonsymmetric system, disregarding the original symmetry of the problem or, one could use, e.g., symmetric Lanczos, obtaining the non-orthogonal basis W_m of $\mathcal{K}_m(A, r_0)$ and computing $x_m = x_0 + W_m y_m$, where $T_m y_m = \beta e_1$; see [3] for an example of such approach. One step further consists in relaxing the symmetry property of T_m, and perhaps imposing more vectors to be locally orthogonal, [1], [24]. The overall process is a truncation of the corresponding optimal (globally orthogonal) method, where the orthogonalization process is incomplete.

In this paper we aim to show that the preconditioner P does not have to be close to symmetric for the non-optimal recurrence to successfully converge at a convergence rate that is not too far from the optimal one. Moreover, our considerations go far beyond the example above, and can be applied to a wide class of methods, such as

- The nonsymmetric Lanczos method [19] and its variations, e.g., QMR [10];
- Incomplete (truncated) methods [26], [27, §6.4.2 and §6.5.6], [29], and restarted methods [27], [31];

- Indefinite inner product methods for complex symmetric matrices [9], [35].

In truncated methods, the orthogonalization of a new basis vector is performed only with respect to a selected number of previous vectors of the basis, say k vectors. We refer to the number k as the truncation parameter. In this case, the upper Hessenberg matrix $T_{m+1,m}$ in (2.1) is banded with k nonzero superdiagonals; see, e.g., [26], [27, §6.4.2 and §6.5.6], [29].

We remark that for complex symmetric matrices, the methods proposed in [9], [35], replace the complex inner product $\langle x, y\rangle = \bar{x}^* y$ with the indefinite inner product $x^* y$ (transposition without complex conjugation), obtaining a basis of the Krylov subspace which is orthogonal with respect to the latter inner product, but not with respect to the underlying (complex) inner product.

In the examples just described, the lack of orthogonality of W_m is intrinsic to the methods used and is not related to the loss of orthogonality due to round-off errors, cf. [13]. In fact, we assume throughout our analysis that exact arithmetic is employed.

One could interpret the matrices W_m as the matrices of an optimal method computed with floating point arithmetic, and thus having non-orthogonal columns. In this case, our theoretical identities and bounds would relate to the delay of convergence due to round-off errors, i.e., the delay in comparison with an ideal exact arithmetic computation. We do not emphasize this situation, but mention that this is related to similar studies where this type of delay is analyzed in detail e.g., in [15], [20], [21], [22], [25]. Moreover, our study supports the common practice of using classical Gram-Schmidt and Modified Gram-Schmidt in Krylov subspace methods using finite precision arithmetic. This follows by considering the bases generated by these methods as the non-optimal bases W_m. As already pointed out in the case of GMRES [14], what matters for convergence and stability is the linear independence of these bases; see also [20] and other references therein. Therefore, it is not necessary for W_m to be close to being orthogonal. In fact $\|I - W_m^* W_m\|$ can be very large while the method may have good convergence properties; see, e.g., Example 5.5.

We mention that in certain circumstances, our theory applies to implicitly restarted augmented methods as well, since in those cases, the resulting space is a Krylov subspace; see [23] and references therein.

As we already mentioned, the non-optimal methods mentioned in this section do not always converge (due to stagnation or breakdown), and furthermore in many cases the convergence may be slow (large delay). Sometimes, these problems are caused by the fact that $W_{m+1} = [W_m, w_{m+1}]$ fails to be of full rank, i.e., its columns fail to span $\mathcal{K}_{m+1}(A, r_0)$. We assume throughout the paper though that the new vector of the basis, w_{m+1}, is linearly independent with respect to the previous ones. In fact, we provide theoretical and

experimental evidence that W_{m+1} may be severely ill-conditioned, while the effect on the delay of convergence is minimal.

Other authors have specifically looked at comparisons of quasi-minimal residual methods with optimal methods; see, e.g., [10], [18], [27] and the references therein. Although our study does not concentrate on these methods, our identities are still valid for them, and our bounds in some cases improve the existing results. We also mention several papers where certain comparisons are made between methods for linear systems or eigenvalue problems based on an Arnoldi relation (3.1) and those based on Lanczos biorthogonalizations (including QMR) [5], [6], [7], [8], [30].

We remark that our motivational problem may be addressed in a different, although related, manner. More precisely, efforts have been devoted to the analysis of the performance of symmetric Lanczos as a function of the lack of symmetry in the (preconditioned) coefficient matrix [12], [24]. In this case, what is measured is the distance from the *optimal* (symmetric) problem, keeping the method fixed. In our context, we assume that the problem cannot (need not) be modified, while we wish to measure the distance from a corresponding optimal, orthogonal, method for nonsymmetric problems; cf. also Example 5.5.

By $\mathcal{R}(M)$ we denote the range of the matrix M, i.e., its column space, by $\mathcal{N}(M)$ its null space, and by $\kappa(M)$ its condition number, i.e., the ratio of its (nonzero) smallest and largest singular values $\sigma_{\max}(M)/\sigma_{\min}(M)$.

3 Relative gaps between optimal and non-optimal residuals

In this section we present identities and bounds for the relative distance between the residuals of the optimal and non-optimal methods, that is, for $m > 0$, we evaluate the ratio

$$\frac{\|r_m^{opt} - r_m^{nonopt}\|}{\|r_m^{opt}\|},$$

in terms of the properties of the generated non-optimal (non-orthogonal) basis. A ratio much larger than one means that the non-optimal method is delayed. As already stated, we assume that W_{m+1} is full rank, although it may be highly ill-conditioned.

Let $V_m = [v_1, v_2, \ldots, v_m]$ with $v_1 = \beta r_0$, $\beta \neq 0$, and v_i, $i = 1, \ldots, m$, being orthonormal vectors. If V_m is generated by the Arnoldi process, it satisfies the following so-called Arnoldi relation

$$AV_m = V_{m+1} H_{m+1,m} = V_m H_m + h_{m+1,m} v_{m+1} e_m^*, \tag{3.1}$$

where $H_{m+1,m} = (h_{ij}) \in \mathbb{R}^{(m+1)\times m}$ is upper Hessenberg, H_m is its principal $m \times m$ submatrix. Clearly, (3.1) is equivalent to (2.1) whenever W_m has orthonormal columns.

Our first result relates the residual of the optimal and non-optimal methods using (2.2).

Theorem 3.1 *Assume that m iterations of a non-optimal Krylov subspace method using* (2.2) *have been performed with the matrix A. Let* $P_A = AV_m(V_m^*AV_m)^{-1}V_m^*$ *be the projector*[1] *onto* $A\mathcal{K}_m(A, r_0)$ *along the subspace orthogonal to* $\mathcal{K}_m(A, r_0)$. *Let* r_m *and* r_m^F, *be the residuals of the non-optimal method and of FOM, respectively. Then*

$$\frac{\|r_m^F - r_m\|}{\|r_m^F\|} = \frac{\|P_A w_{m+1}\|}{\|(I - P_A)w_{m+1}\|} \tag{3.2}$$

$$\leq \frac{1}{\|(I - P_A)w_{m+1}\|} + 1 = \frac{\|r_m\| + \|r_m^F\|}{\|r_m^F\|}\,. \tag{3.3}$$

Proof. We have $(I-P_A)r_m \in r_0 + A\mathcal{K}_m(A, r_0)$ and $(I-P_A)r_m \perp \mathcal{K}_m(A, r_0)$. Hence, $(I-P_A)r_m$ is equal to the optimal residual r_m^F, that is, $r_m^F = (I-P_A)r_m$. It can be easily verified that $r_m = \pm\|r_m\|w_{m+1}$, therefore we also obtain $\|r_m^F\| = \|(I - P_A)w_{m+1}\|\,\|r_m\|$. Hence, $r_m^F - r_m = -P_A r_m$ from which

$$\|r_m^F - r_m\| = \|P_A r_m\| = \|P_A w_{m+1}\|\,\|r_m\| = \frac{\|P_A w_{m+1}\|}{\|(I - P_A)w_{m+1}\|}\,\|r_m^F\|.$$

This proves the equality. The bound follows immediately. □

We emphasize the fact that our result (3.2) is an equality, and in particular, for W_m having orthonormal columns, we have that $\|r_m^F - r_m\| = 0$. Our result indicates that as long as the new direction vector w_{m+1} provides significant new information with respect to the subspace $\mathcal{K}_m(A, r_0)$ generated so far, then, the two residuals will be very close to each other and the relative residual gap is small, i.e., the delay in the convergence is small. We observe that for FOM, this result indicates that the orthogonality of the basis up to the previous step is of no importance, only the angle between the new vector and the subspace generated so far is.

In the following examples we illustrate the fact that the bound in (3.3) may be very sharp. This fact can be appreciated by noticing that the inequality in the bound is only due to the step from (3.2) to (3.3). In the figures, the quantity ν stands for $\|(I - P_A)w_{m+1}\|$ in (3.3).

Example 3.2 We consider the 100×100 matrix stemming from a centered finite differences discretization of the operator $\mathcal{L}(u) = -\Delta u + 100u_x$ in the unit square, with homogeneous boundary conditions. The right-hand side is a nonzero vector of equal entries. In Figure 1, on the left, we report the convergence history for FOM and for the truncated method with $k = 10$, while on the right we show the relative gap in (3.2) and its bound (3.3). It can

[1] The same projector P_A can alternatively be defined as $AW_m(W_m^*AW_m)^{-1}W_m^*$.

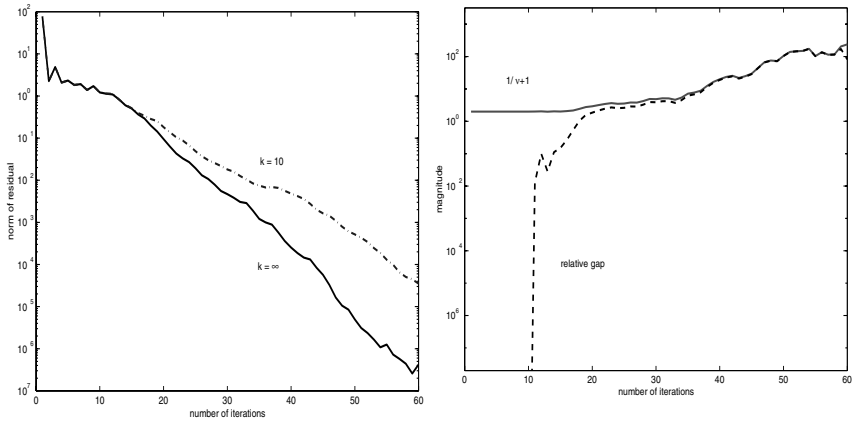

Fig. 1. Example 3.2: FOM and truncated FOM ($k = 10$). Left: Convergence history. Right: Relative gap (3.2) and bound (3.3)

be appreciated that the non-optimal method has good convergence behavior, and that the bound (3.3) can be very sharp.

We next show that the result of Theorem 3.1 applies to other examples, as mentioned in section 2.

Example 3.3 We consider the same matrix as in Example 3.2, the right-hand side $b = e_1$, and we apply the nonsymmetric Lanczos algorithm. The nonsymmetric Lanczos method determines a basis W_m satisfying (2.1), that is bi-orthogonal with respect to a basis spanning $K_m(A^*, p)$, where p is an a-priori chosen vector; we selected $p = b$. It is well known that the method is very sensitive to the choice of p, which influences the quality of the basis W_m. An approximate solution is determined by means of (2.2). The corresponding convergence curve, together with that of the optimal FOM method is reported in the left plot of Figure 2. It can be appreciated that the non-optimal method has a delay but converges in about 10–20% more iterations. The residual norm for the Lanczos method significantly deviates from that of FOM, and this gap is accurately bounded by the term (3.3) (see the right plot of Figure 2), as predicted by Theorem 3.1.

Example 3.4 We next consider the performance of the restarted FOM method. It was shown in [31] that this method can be globally represented as a truncated method with a special truncation strategy. In particular, the whole process can be represented by (2.1) with $T_{m+1,m}$ block diagonal, with upper Hessenberg blocks. We again consider the same matrix and right-hand side as in Example 3.2 and we report in the left plot of Figure 3, the convergence history of the optimal FOM method and the restarted FOM method with maximum Krylov subspace dimension before restarting, equal to 10.

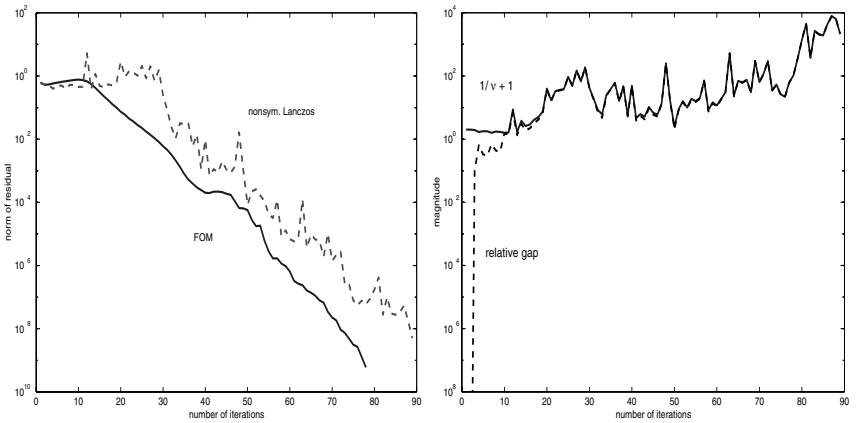

Fig. 2. Example 3.3. Left: convergence history for FOM and nonsymmetric Lanczos. Right: Relative gap (3.2) and bound (3.3)

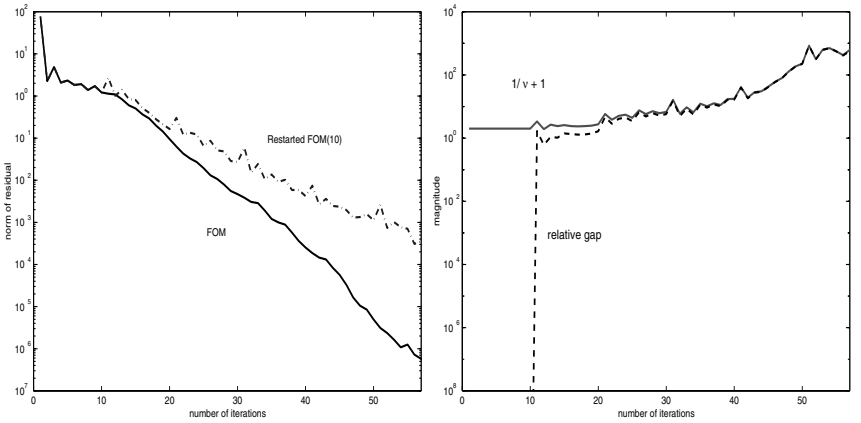

Fig. 3. Example 3.4. Left: convergence history for FOM and Restarted FOM with restarting parameter equal to 10. Right: Relative gap (3.2) and bound (3.3)

We can see that convergence is similar, but not equal, to that of the regularly truncated method. We can also see that in spite of the restarting procedure, the method is capable of constructing a full Krylov subspace basis, so that the gap between the optimal and non-optimal residuals remains moderate; see the right plot of Figure 3. We refer to [31] for a more detailed analysis of the restarted methods.

We proceed now to study the residual gap in the case of Petrov-Galerkin, i.e., of methods such as GMRES [28]. To that end, consider the approximation using the non-optimal basis, i.e., $x_m = x_0 + W_m y_m$, where y_m is the minimizer of (2.3), and that obtained with the optimal basis, i.e., the GMRES approxima-

tion $x_m^G = x_0 + V_m y_m^G$, where y_m^G minimizes $\|\beta e_1 - H_{m+1,m} y\| = \|b - AV_m y\|$ over all $y \in \mathbb{R}^m$. The corresponding residuals $r_m = b - Ax_m = r_0 - AW_m y_m$ and $r_m^G = r_0 - AV_m y_m^G$ both lie in the affine space $r_0 + A\mathcal{K}_m(A, r_0)$ and are related as in the following lemma.

Lemma 3.5 *Let $r \in r_0 + A\mathcal{K}_m(A, r_0)$, and $r_m^G = r_0 - Ax_m^G$, where x_m^G is the minimizer of $\min_{x \in \mathcal{K}_m(A,r_0)} \|r_0 - Ax\|$. Then $(r - r_m^G)^* r_m^G = 0$ and thus*

$$\|r - r_m^G\|^2 + \|r_m^G\|^2 = \|r\|^2. \tag{3.4}$$

Proof. Both r and r_m^G lie in $r_0 + A\mathcal{K}_m(A, r_0)$ and therefore their difference $r - r_m^G \in A\mathcal{K}_m(A, r_0)$. The GMRES residual minimizes the 2-norm over all elements in $r_0 + A\mathcal{K}_m(A, r_0)$, and therefore $r_m^G \perp A\mathcal{K}_m(A, r_0)$, from which $(r_m^G)^*(r_m - r_m^G) = 0$. The equality (3.4) follows using Pythagoras' theorem. □

In general, we can relate the two equations (3.1) and (2.1) as follows. Consider a reduced QR factorization $W_i = V_i U_i$, $i = m, m+1$, with $V_i \in \mathbb{R}^{n \times i}$ having orthonormal columns and $U_i \in \mathbb{R}^{i \times i}$ upper triangular. Let

$$\widehat{T}_{m+1,m} = U_{m+1} T_{m+1,m}, \tag{3.5}$$

which is an upper Hessenberg matrix when $T_{m+1,m}$ is upper Hessenberg, and, as before, we denote by $\widehat{T}_m$ its principal $m \times m$ submatrix.

Theorem 3.6 *Assume that m iterations of a non-optimal Krylov subspace method have been performed with the matrix A and that $U_{m+1} = V_{m+1}^* W_{m+1}$ is nonsingular. Let $r_m = b - AW_m T_{m+1,m}^+ e_1 \beta$ and $r_m^G = b - AW_m y_m^G \equiv V_{m+1} g_m^G$ be the GMRES residual, with $y_m^G = \widehat{T}_{m+1,m}^+ e_1 \beta$. Then*

$$\|r_m - r_m^G\| = \left(\alpha_m^2 - 1\right)^{\frac{1}{2}} \|r_m^G\|, \tag{3.6}$$

where

$$\alpha_m = \frac{\|r_m\|}{\|r_m^G\|} = \frac{\|U_{m+1} t\|}{\|U_{m+1}^* d\|}, \tag{3.7}$$

with $t \in \mathcal{N}(T_{m+1,m}^*)$, $d \in \mathcal{N}(\widehat{T}_{m+1,m}^*)$, both of unit length.

Proof. The identity (3.6) follows directly from Lemma 3.5. Let $T = T_{m+1,m}$ and $\widehat{T} = \widehat{T}_{m+1,m}$, for short. For $d \in \mathcal{N}(\widehat{T}^*)$, $\|d\| = 1$, we can write (by choosing the appropriate direction of d)

$$g_m^G = (I - \widehat{T}\widehat{T}^+) e_1 \beta = d d^* e_1 \beta, \quad \text{so that} \quad \|g_m^G\| = \|r_m^G\| = |d^* e_1 \beta|. \tag{3.8}$$

To prove the second equality in (3.7), for $t \in \mathcal{N}(T^*)$, $\|t\| = 1$, we write $\|r_m\| = \|W_{m+1} t t^* e_1 \beta\| = \|W_{m+1} t\| |t^* e_1 \beta| = \|U_{m+1} t\| |t^* e_1 \beta|$. Using (3.5)

we can write (by choosing the appropriate direction of t) $t = U_{m+1}^* d / \|U_{m+1}^* d\|$, from which, since $U_{m+1} e_1 = e_1$, using (3.8) we obtain

$$\|r_m\| = \|U_{m+1} t\| |t^* e_1 \beta| = \frac{\|U_{m+1} t\|}{\|U_{m+1}^* d\|} |d^* U_{m+1} e_1 \beta| = \frac{\|U_{m+1} t\|}{\|U_{m+1}^* d\|} \|r_m^G\|.$$

□

We emphasize that as in the case of Theorem 3.1, our result (3.6) with (3.7) is an equality, but unlike that case, here its direct interpretation is less clear. What we can say in general is that since $U_{m+1} = V_{m+1}^* W_{m+1}$ and t is of unit norm, we have that $\|U_{m+1} t\| \leq \sqrt{m+1}$, and thus its maximum growth is well understood. The analysis of the key quantity $\|U_{m+1}^* d\|$ is taken up in the next section for the particular case of GMRES.

We end this section with a general bound for the relative residual gap. By Lemma 3.5, we have that $r_m^* r_m^G = \|r_m^G\|^2$, so that we can write $\cos\theta = r_m^* r_m^G / (\|r_m\| \|r_m^G\|) = 1/\alpha_m$, and thus $|\tan\theta| = \sqrt{\alpha_m^2 - 1}$. Therefore, the relative gap only depends on the angle between the two residual vectors, and not on the norm of r_m. A well-known bound for α_m is $\kappa(U_{m+1})$, see, e.g., [13, Theorem 5.3.1], so that $|\tan\theta| < \alpha_m \leq \kappa(U_{m+1})$. Here we present a slightly better bound, whose proof is given in the Appendix.

Proposition 3.7 *With the notation of Theorem 3.6,*

$$\frac{\|r_m - r_m^G\|}{\|r_m^G\|} = |\tan\theta| \leq \frac{1}{2} \kappa(U_{m+1}) \left(1 - \frac{1}{\kappa(U_{m+1})}\right) < \kappa(U_{m+1}). \tag{3.9}$$

The estimate of Proposition 3.7 is in line with earlier results for quasi-minimal residual type methods, see, e.g., [10], [27], which relate the loss of optimality to the ill-conditioning of the non-orthogonal basis. Experimental results indicate that this bound is very pessimistic, and that the relative gap may be several orders of magnitude lower than $\kappa(U_{m+1})$ in practice. In the next section we analyze the quantity $\|U_{m+1}^* d\|$, resulting in an understanding of the behavior of the relative residual gap in the particular case of GMRES.

4 Analysis of the residual gap for GMRES

In this section we analyze the vector $U_{m+1}^* d$, which determines the size of the relative residual gap for GMRES in the bound

$$\frac{\|r_m - r_m^G\|}{\|r_m^G\|} \leq \frac{\sqrt{m+1}}{\|U_{m+1}^* d\|}\ ; \tag{4.1}$$

see Theorem 3.6 and the comments following it. We want to see when we can expect $\|U_{m+1}^* d\|$ to remain sufficiently away from zero, so that the relative

residual gap is not amplified as m grows. Thus, we can obtain a least qualitatively, a bound not as pessimistic as (3.9). The structure of the null vector d plays a crucial role in this analysis.

Consider the QR factorization of $\widehat{T}_{m+1,m} = Q_{m+1}R$, and observe that d is the $(m+1)$st column of Q_{m+1} and that the orthogonal matrix Q_{m+1} is the same as one would obtain in the QR factorization of $H_{m+1,m}$. Indeed, $H_{m+1,m} = T_{m+1,m}U_m^{-1} = Q_{m+1}RU_m^{-1} := Q_{m+1}R_1$. We recall that

$$Q_{m+1}^* = \Omega_m \Omega_{m-1} \cdots \Omega_1, \quad \text{where} \quad \Omega_i = \begin{bmatrix} I_{i-1} & & & \\ 0 & c_i & s_i & 0 \\ 0 & -s_i & c_i & 0 \\ & & & I \end{bmatrix}$$

is the appropriate Givens rotation; see, e.g., [27]. Explicit computation shows that

$$d = Q_{m+1}e_{m+1} = \begin{bmatrix} s_m s_{m-1} \cdots s_1 \\ s_m s_{m-1} \cdots s_2 c_1 \\ s_m s_{m-1} \cdots s_3 c_2 \\ \vdots \\ s_m c_{m-1} \\ c_m \end{bmatrix} = \begin{bmatrix} s_m s_{m-1} \cdots s_1 \\ s_m s_{m-1} \cdots s_1 (c_1/s_1) \\ s_m s_{m-1} \cdots s_1 (c_2/(s_1 s_2)) \\ \vdots \\ s_m s_{m-1} \cdots s_1 (c_{m-1}/(s_1 s_2 \cdots s_{m-1})) \\ c_m \end{bmatrix}.$$

We also recall that $\beta |s_k s_{k-1} \cdots s_1| = \|r_k^G\|$ with $\beta = \|r_0^G\|$, so that

$$d = \begin{bmatrix} \|r_m^G\|/\|r_0^G\| \\ (\|r_m^G\|/\|r_1^G\|)c_1 \\ \vdots \\ (\|r_m^G\|/\|r_{m-1}^G\|)c_{m-1} \\ c_m \end{bmatrix}.$$

Therefore, as $\|r_j^G\|$ decreases with j, the components of $|d|$ are expected to have an increasing pattern. In particular, the ith component $\delta_i = e_i^* d$ is characterized by the ratio $\|r_m^G\|/\|r_{i-1}^G\|$, which indicates how much the residual has decreased in the last $m-i$ iterations. In fact, if we let ρ be the smallest positive number such that $\|r_j^G\|/\|r_{j-1}^G\| \le \rho$, for all $j \le m$, we have

$$|\delta_i| \le \rho^{m-i+1}, \ i = 1, \ldots, m+1. \tag{4.2}$$

We call ρ the convergence rate of GMRES. Moreover, writing $U_{m+1}^* d = W_{m+1}^* V_{m+1} d$, we see that the gap between the optimal and non-optimal residuals solely depends on how well the "residual" direction vector $V_{m+1}d$ is

represented in the non-optimal basis. A related important consideration is that, apart from the sign, we can write

$$\frac{1}{\|r_m^G\|} r_m^G = V_{m+1} d = \sum_{i=1}^{m+1} \delta_i v_i \ . \tag{4.3}$$

The linear combination in (4.3) and the bounds (4.2) lead to the following remark which applies to the cases of interest described in the introduction, i.e., fast convergence of the optimal method. In the following remark we attempt to quantify what we mean by fast convergence, and illustrate this with Example 4.2 below.

Remark 4.1 If GMRES converges with convergence rate $\rho \ll 1$, then the residual direction vector $r_m^G / \|r_m^G\|$ lies very close to the subspace generated by the last computed basis vectors. In other words, if the convergence is fast, i.e., $\rho \ll 1$, we may not need to orthogonalize with respect to older vectors, since r_m^G may have very small components in them, i.e., r_m^G may be essentially a linear combination of the last few vectors (but not necessarily only a multiple of the last one).

Remark 4.1 has some insightful consequences. Firstly, if the rate ρ is significantly less than one, then for $\|r_m - r_m^G\| / \|r_m^G\|$ to be bounded is sufficient that the last columns in W_{m+1} have a non-negligible projection onto the corresponding portion of V_{m+1}. More precisely, for $\ell \geq 0$, we write

$$d = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}, \quad d_2 \in \mathbb{R}^{\ell+1}, \quad U_{m+1} = \begin{bmatrix} \mathcal{U}_1 & \mathcal{U}_2 \\ 0 & \mathcal{U}_3 \end{bmatrix}, \quad \mathcal{U}_3 \in \mathbb{R}^{(\ell+1)\times(\ell+1)} \ .$$

We have $\|d_1\|^2 \leq \rho^{2m} + \rho^{2m-2} + \cdots + \rho^{2(\ell+1)}$, so that $\|d_1\| \leq \sqrt{m-\ell}\, \rho^{\ell+1} \equiv \varepsilon$. Then, for $\varepsilon \ll 1$,

$$U_{m+1}^* d = \begin{bmatrix} \mathcal{U}_1^* d_1 \\ \mathcal{U}_2^* d_1 + \mathcal{U}_3^* d_2 \end{bmatrix}, \quad \text{with} \quad \begin{array}{l} \|\mathcal{U}_1^* d_1\| \approx \varepsilon \\ \|\mathcal{U}_2^* d_1 + \mathcal{U}_3^* d_2\| \approx \|\mathcal{U}_3^* d_2\|. \end{array}$$

To maintain $\|U_{m+1}^* d\|$ sufficiently larger than zero, it appears that it is only required that $\mathcal{U}_3 = V_{m-k:m+1}^* W_{m-k:m+1}$ be not too ill-conditioned (and this need is mitigated somehow by the fact that $\|d_2\| \approx \rho$, and no smaller). This goal may be achieved if the subspace dimension is ensured to increase while using the non-optimal basis W_{m+1}. In this respect, full orthogonality of the basis seems to be an unnecessarily strict condition.

A second, though related, consequence of Remark 4.1 is that if $\rho \ll 1$, then loss of orthogonality with respect to older vectors in the basis W_{m+1} may be harmless. Indeed, $\|\mathcal{U}_1^* d_1\| \approx \varepsilon$ even for $\mathcal{U}_1$ close to the identity matrix. The first block of the vector $U_{m+1}^* d$ does not contribute to the size of the vector, as it may be $\|\mathcal{U}_1^* d_1\| \ll \|U_{m+1}^* d\|$. Hence, enforcing orthogonality seems

to be beneficial only as a means to enlarge the subspace dimension, i.e., to maintain linear independence. We mention here again that orthogonality of the basis is not needed for stability of GMRES, only linear independence is; see, e.g., [14], [20].

The two considerations just discussed remain valid without the assumption (4.2) if $\|r_m^G\|$ is much smaller than $\|r_0^G\|, \|r_1^G\|, \ldots$, yielding small first components in d. This is the case for example when the GMRES residual norm shows initial stagnation, followed by a substantial decrease. Our experiments in the rest of the paper indeed show that the non-optimal methods are not significantly affected by an initial stagnation of the optimal residual, as long as they are able to capture the reduction in the residual norm, cf. [20], [21], [22].

Finally, since $U_{m+1}e_1 = e_1$, an obvious lower bound for $\|U_{m+1}^* d\|$ can be obtained by setting $\ell = m-1$ in the relation of the paragraph after Remark 4.1, so that $\|U_{m+1}^* d\| \geq \|\mathcal{U}_1^* \delta_1\| = \|r_m^G\|/\|r_0\|$, and thus

$$\|U_{m+1}^* d\| \geq \max \left\{ \sigma_{\min}(U_{m+1}), \frac{\|r_m^G\|}{\|r_0\|} \right\}.$$

This bound ensures that the non-optimal method does not significantly deviate from the optimal iteration, when, for instance, GMRES convergence experiences a slow initial phase.

A more optimistic estimate can be obtained when using truncated methods. Indeed, if $k+1$ consecutive basis vectors are maintained orthogonal, then the leading $(k+1) \times (k+1)$ block of U_{m+1} is the identity matrix, so that

$$\|U_{m+1}^* d\| \geq \|[U_{k+1}^*, O]d\| \approx \rho^{m-k+1}. \tag{4.4}$$

We illustrate our results and in particular this estimate with the following example.

Example 4.2 The aim of this example is to show that in (4.1), the quantity $\|U_{m+1}^* d\|$ may be much larger than $\sigma_{\min}(U_{m+1})$, and that (4.4) may provide a computable quantity to estimate $\|U_{m+1}^* d\|$ when using a truncated method.

We consider the same problem as in Example 3.2. In the left plot of Figure 4 we report the convergence history for GMRES and for the truncated method, with $k = 10$. In the right plot we report the values of $\sigma_{\min}(U_{m+1})$ (dashed) and of $\|U_{m+1}^* d\|$ (solid), as the iteration proceeds. As anticipated, the estimate using $\sigma_{\min}(U_{m+1})$ would provide an extremely pessimistic picture. The dash-dotted line is the graph of ρ_k^{m-k+1}, where $\rho_k = \|r_k\|/\|r_{k-1}\|$ is the residual decrease at the last step before truncation takes place, that is, at the last iteration the optimal residual is available. If the optimal method is converging fast, and $\rho_k \leq \rho$, then we expect that ρ_k^{m-k+1} be a good measure of $\|U_{m+1}^* d\|$, which cannot be computed when running the non-optimal

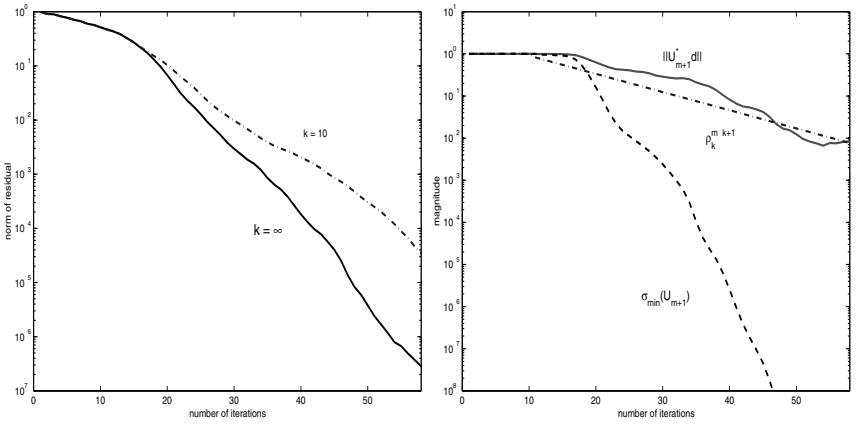

Fig. 4. Example 4.2:GMRES and truncated GMRES ($k = 10$). Left: Convergence history. Right: Illustration of estimate (4.4)

method. This statement is confirmed by the curve in the right plot of Figure 4. It should be remarked, however, that this is a very rough measure, since it strongly depends on the convergence rate at the kth iteration. A more reliable measure should take into account all available information, while possibly dynamically adapting the parameter as the iteration proceeds.

5 Delay and Invariant Subspaces

We return to the question on when one can expect larger or smaller delays, depending on the matrix A and the method used. We begin by rewriting (2.1) as

$$\begin{aligned} AW_m &= W_{m+1}T_{m+1,m} = V_{m+1}T_{m+1,m} + (W_{m+1} - V_{m+1})T_{m+1,m} \\ &=: V_{m+1}T_{m+1,m} + E_m\ , \end{aligned} \tag{5.1}$$

implying the relation

$$(A + \mathcal{E}_m)W_m = V_{m+1}T_{m+1,m}, \tag{5.2}$$

with $\mathcal{E}_m = -E_m U_m^{-1} V_m^*$. We highlight the difference between this last relation and the identities (3.1) and (2.1). In these earlier relations we represented the matrix A using optimal (orthogonal) or non-optimal bases of $\mathcal{K}_m(A, r_0)$, via their representations $H_{m+1,m}$ or $T_{m+1,m}$. In (5.2) we keep the non-optimal representation $T_{m+1,m}$, which becomes "optimal" for a modified matrix $(A + \mathcal{E}_m)$, i.e., the result of a projection with an orthogonal matrix V_{m+1}.

In the context of this paper $\|E_m\| = \|(U_{m+1} - I)T_{m+1,m}\|$ tends to grow with m (as the columns of W_m are less and less orthogonal); this fact and the

relations (5.1) and (5.2) resemble those for inexact Krylov subspace methods [32]. This similarity allows us to naturally generalize the theory developed for those inexact methods in [33] to the non-optimal approaches studied here.

The general idea of this superlinear convergence theory is that the iterative method changes convergence rate, i.e., the residual curve gets steeper, if the Krylov subspace contains a good approximation to an invariant subspace of A. For the unperturbed problem, i.e., for $\mathcal{E}_m = 0$, it is shown in [33] that the norm of the $(m + j)$th GMRES residual is comparable to that of another GMRES method whose initial vector is the mth GMRES residual stripped of components on the appropriate invariant subspace of A; see Theorem 5.1 below and the comments following it. Furthermore, the perturbed problem with $A + \mathcal{E}_m$ behaves as the unperturbed one as long as the iterative method is able to capture similar invariant subspace information. This is particularly apparent when considering truncated methods: if one orthogonalizes with respect to the previous k vectors, we have that $\mathcal{E}_m = 0$, $m = 1, \ldots, k$, so that the perturbed problem coincides with the unperturbed one, for $m \leq k$.

The following result formalizes the application of our theory from [33]. We comment further on its implications later in this section. Let r_m, as before, be the residual at the mth step of a non-optimal minimal residual method based on (2.1), or equivalently on (5.2). Here $\|\cdot\|$ represents the norm induced by the underlying inner product $\langle x, y\rangle$.

Theorem 5.1 *Let P_Q be a spectral projector onto a simple invariant subspace $\mathcal{R}(Q)$ of $A + \mathcal{E}_m$ of dimension s, and Π_Y the orthogonal projector onto an s-dimensional subspace $\mathcal{R}(Y) \subset (A + \mathcal{E}_m)\mathcal{K}_m(A + \mathcal{E}_m, r_0)$. Then, the following bound holds for $j = 1, \ldots,$*

$$\begin{aligned} \|r_{m+j}\| &= \min_{d \in (A+\mathcal{E}_{m+j})\mathcal{K}_{m+j}(A+\mathcal{E}_{m+j}, r_0)} \|r_0 - d\| \\ &\leq \sqrt{2} \min_{d \in (A+\mathcal{E}_{m+j})\mathcal{K}_j(A+\mathcal{E}_{m+j}, r_m)} \left\| \begin{bmatrix} I - P_Q \\ \gamma P_Q \end{bmatrix} (r_m - d) \right\|, \end{aligned} \tag{5.3}$$

where $\gamma = \|(I - \Pi_Y)P_Q\|$.

Proof. Observe first that $\mathcal{R}(AW_m) = \mathcal{R}(AV_m) = A\mathcal{K}_m(A, r_0)$, and similarly

$$\begin{aligned} \mathcal{R}((A + \mathcal{E}_m)W_m) &= \mathcal{R}((A + \mathcal{E}_m)V_m) = (A + \mathcal{E}_m)\mathcal{K}_m(A + \mathcal{E}_m, r_0) \\ &= \mathcal{R}(V_{m+1}T_{m+1,m}). \end{aligned}$$

Furthermore, using the form of the upper Hessenberg matrix $T_{m+j+1,m+j}$, whose $(m + 1) \times m$ submatrix is $T_{m+1,m}$ and by considering $V_{m+j+1} = [V_{m+1}, \hat{V}_j]$ for some matrix $\hat{V}_j$, we have $\mathcal{R}(V_{m+1}T_{m+1,m}) \subset \mathcal{R}(V_{m+j+1}$ $T_{m+j+1,m+j})$, for $j = 1, \ldots$. The theorem follows now by applying [33, Theo. 3.1]. □

The value of γ quantifies the proximity of two subspaces: the invariant subspace, and a subspace $\mathcal{R}(Y)$ of the same dimension in $(A+\mathcal{E}_m)\mathcal{K}_m(A+\mathcal{E}_m, r_0)$. The better approximated the invariant subspace is in $\mathcal{R}(Y)$, the smaller γ is; see further [33]. When γ is small, Proposition 3.4 and Corollary 3.6 of [33] ensure that the solution of the least squares problem (5.3) is very close to the solution of the minimization problem with initial vector $\bar{r}_m = (I - P_Q)r_m$, namely, to the solution of

$$\min_{d \in (A+\mathcal{E}_{m+j})\mathcal{K}_j(A+\mathcal{E}_{m+j}, \bar{r}_m)} \|\bar{r}_m - d\| .$$

In summary, the spectral properties of matrix A and the initial residual r_0 determine if the non-optimal iterative method has a superlinear convergence behavior similar to that of the optimal method or if there is a convergence delay. If A is such that an invariant subspace of A is well approximated in $\mathcal{K}_m(A, r_0)$ and if similarly an invariant subspace of the perturbed matrix $A+\mathcal{E}_m$ is well approximated in $\mathcal{K}_m(A+\mathcal{E}_m, r_0)$, then the delay is small. This will happen if, for example, the truncation parameter k is large enough.

In general, and not necessarily related to the superlinear convergence described in Theorem 5.1, small delay occurs when the spectrum of A is sufficiently clustered and not too sensitive to perturbations, so that the spectral properties of $A+\mathcal{E}_m$ do not significantly differ from those of A. In this case, a small number of iterations is enough to capture the information for the optimal and non-optimal methods to have fast convergence. This is the case, for example, of optimally preconditioned problems.

In the rest of the section we report on some additional numerical experiments which illustrate and support our theoretical analysis. Again, we remind the reader that all our analysis assumes finite precision arithmetic, whereas the experiments obviously do not. In the following two examples we analyze the dependence of the non-optimal method on the clustering of the spectrum.

Example 5.2 Consider the 100×100 matrix $A = XJ_1X^{-1}$, where X is a random matrix with normally distributed entries (Matlab function randn with initial random generator state), with condition number of about 500. Matrix J_1 is a block diagonal matrix with 2×2 blocks of the form

$$I_2 + \frac{1}{2}\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

and θ uniformly distributed in $[-\pi, \pi]$. The eigenvalues are depicted in the left plot of Figure 5. We compare the residual history for optimal GMRES and for truncated GMRES, with truncation parameter $k = 5, 10$, and right-hand side equal to a nonzero vector of equal entries. Due to this particular eigenvalue distribution, the convergence of GMRES is linear, with convergence rate depending on the disk radius [27]. Both truncated methods are

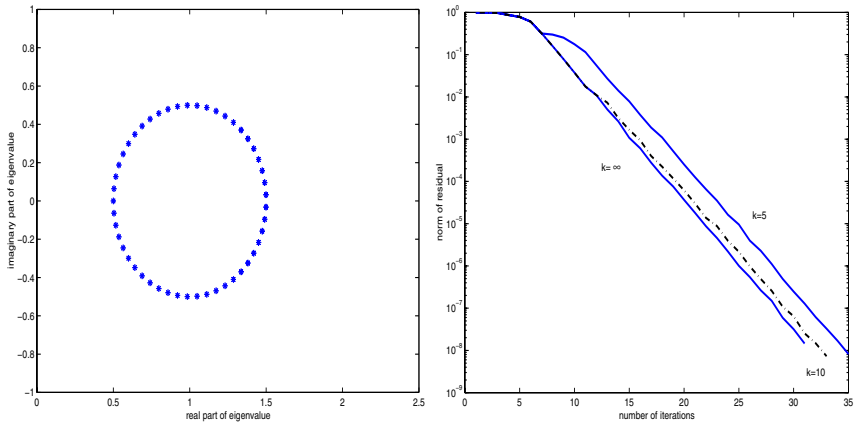

Fig. 5. Example 5.2. Left: Spectrum of the matrix. Right: Convergence history of GMRES and truncated GMRES ($k = 5,\ 10$)

able to correctly identify the perturbed cluster, maintaining the same rate of convergence, even for the strict truncation parameter, $k = 5$.

Example 5.3 We next modify some of the eigenvalues of A in the previous example, that is, we consider the 100×100 matrix $A = X J_2 X^{-1}$, with X the same as before, where J_2 differs from J_1 only in the first and last four eigenvalues. These are real, as reported in the left plot of Figure 6. The right-hand side is as before. The performance is now significantly different for the truncated methods, as can be deduced from the right plot of Figure 6. On the one hand, GMRES is able to capture the invariant subspaces associated with the

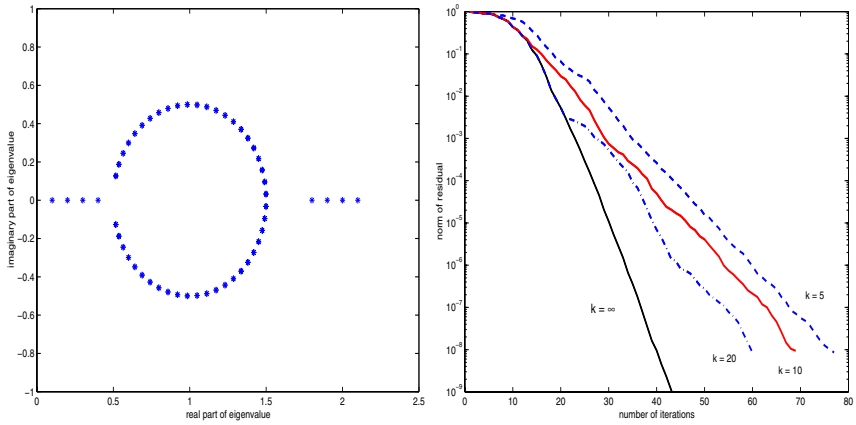

Fig. 6. Example 5.3. Left: Spectrum of the matrix. Right: Convergence history of GMRES and truncated GMRES ($k = 5,\ 10,\ 20$)

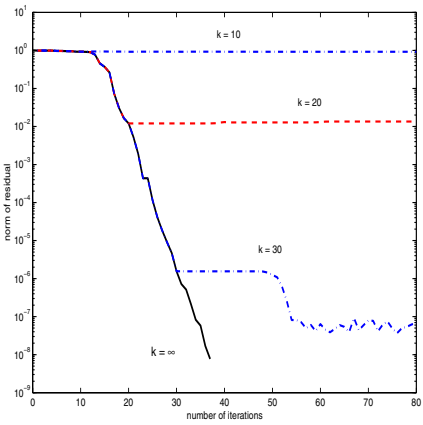

Fig. 7. Example 5.4. Convergence history of GMRES and truncated GMRES ($k = 10,\ 20,\ 30$)

outlying eigenvalues, showing superlinear convergence. On the other hand, according to the setting of (5.2), the truncated procedures solve a perturbed problem with eigenvalues in a larger region, with no outliers. Hence, superlinearity does not occur. It is also remarkable that the convergence slope does not change considerably when passing from 10 to 5 orthogonal vectors, showing that similar information is carried on in the two cases.

Example 5.4 We again modify the matrix in Example 5.2. This time, we maintain the same matrix J_1, yielding the clustered spectrum as in the left plot of Figure 5, whereas we consider an eigenvector matrix X_1 whose condition number is $\kappa(X_1) = 10^6$, providing a matrix $A = X_1 J_1 X_1^{-1}$ with condition number $\kappa(A) \approx 3 \cdot 10^{10}$. The matrix X_1 was generated by means of the the Matlab function randsvd in the Higham's TestMatrix Toolbox [17], by requiring geometrically distributed singular values (a zero seed was used for the random number generator before calling randsvd). The convergence curves for the optimal GMRES method and for its truncated version ($k = 10, 20, 30$) are reported in Figure 7. They show a dramatic difference in the performance, as compared to Figure 5, due to the conditioning of X_1. As discussed earlier in this section, the non-optimal method can be viewed as an inexact method. It was shown in [33] that the condition number of the eigenvector matrix plays a fundamental role in the performance of inexact methods. Our numerical results are thus in agreement with the arguments in [33]. Finally, we remark that in this experiment, the Krylov basis in the non-optimal case looses rank soon after truncation takes place.

Example 5.5 In this last example, we consider a 100×100 symmetric positive definite matrix, corresponding to the finite differences discretization of the

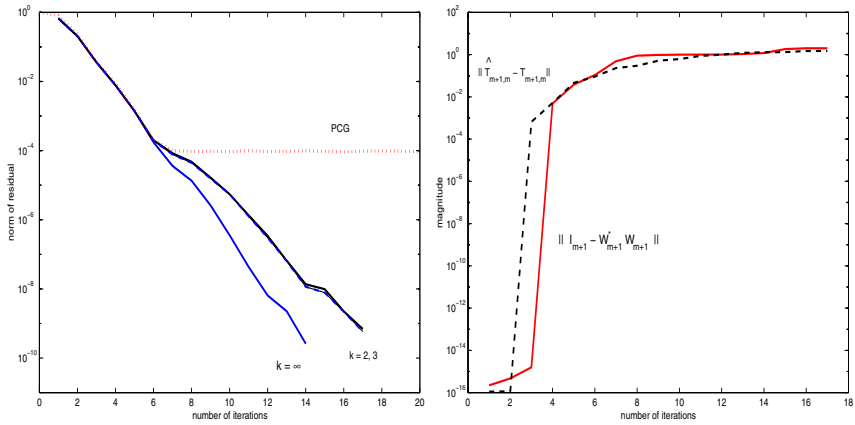

Fig. 8. Example 5.5: convergence history for GMRES and truncated GMRES, $k = 2,\ 3$

2D Laplacian on the unit square. The right-hand side is once again a constant nonzero vector. We explore the typical situation that arises in practice when the preconditioner is only known via a function that, given a vector v, returns the action of the preconditioner as $w = \mathcal{P}(v)$, as is the case for instance in multilevel preconditioning techniques. To simulate this situation, we considered an incomplete Cholesky preconditioner with threshold 10^{-1} (matrix L^* obtained with the Matlab function `cholinc`), together with a perturbation, that is, $w = L^{-*}L^{-1}v + \varepsilon e$, with $\varepsilon = 10^{-5}$ and e a vector of all ones. Here the situation is completely analogous to that described in [1] and [24].

The left plot of Figure 8 reports the convergence history of standard PCG (vertical dashes), as implemented for instance in [11]. As expected, the final attainable accuracy is reached at a level that depends on the nonsymmetry of the preconditioner. In the plot, the convergence history of FOM and that of the truncated version with $k = 2, 3$, are also reported. Right preconditioning is employed. The highly truncated versions are able to capture the fast convergence of the full method. It is important to observe that this is not due to a quasi-orthogonality of the basis, nor to an almost banded form of the representation matrix $T_{m+1,m}$. Indeed, the right plot of Figure 8 shows that for $k = 2$, the basis vectors in W_{m+1} are far from being orthogonal for $m > 3$ (solid curve), and that $\|E_m\| = \|\widehat{T}_{m+1,m} - T_{m+1,m}\|$ is highly non-negligible (dashed curve). Our results earlier in this section help explain the good observed behavior of the truncated methods. The clustering of the preconditioned problem is such that high perturbations in the operator can only mildly alter the linear convergence rate.

6 Conclusion

We have presented new results which help our understanding of the convergence behavior of truncated and other non-optimal Krylov subspace methods. In the non-optimal case, the matrix W_m whose columns span $\mathcal{K}_m(A, r_0)$ may be very far from orthogonal, and we may still have good convergence. The symmetry of A or of its restriction to $\mathcal{K}_m(A, r_0)$, T_m, does not enter into our analysis, we only require that T_m be nonsingular or $T_{m+1,m}$ be full rank, so that the solutions in (2.2) and (2.3) respectively, are well defined. On the other hand, certain properties of A do matter, for example its spectral properties. Furthermore, if A is very sensitive to perturbations, then the non-optimal method may have a considerable delay. In particular, our experiments show that for the non-optimal methods to be competitive, the eigenvector matrix should not be too ill-conditioned; this property should be taken into account when designing effective preconditioning procedures.

Appendix

In this appendix we prove Proposition 3.7.

Proof of Proposition 3.7. Let $T = T_{m+1,m}$ and $\widehat{T} = \widehat{T}_{m+1,m}$, for short, and let d, t be two unit norm vectors spanning $\mathcal{N}(\widehat{T}^*)$ and $\mathcal{N}(T^*)$, respectively. As in the proof of Theorem 3.6, we have $g_m = t\|g_m\|$ with $t = U_{m+1}^* d/\|U_{m+1}^* d\|$, and $r_m^G = \pm V_{m+1} d \|r_m^G\|$. Hence, letting $\sigma_{\min} \neq 0$, $\sigma_{\max}$ denote the smallest and largest singular values of U_{m+1}, we have

$$\frac{1}{\alpha} = |\cos\theta| = \frac{|r_m^* r_m^G|}{\|r_m\|\,\|r_m^G\|} = \frac{|(U_{m+1}g_m)^* d|}{\|U_{m+1}g_m\|} = \frac{(U_{m+1}U_{m+1}^* d)^* d}{\|U_{m+1}U_{m+1}^* d\|}$$
$$= \frac{d^* U_{m+1}U_{m+1}^* d}{\|U_{m+1}U_{m+1}^* d\|} \geq \min_{\|z\|=1} \frac{z^* U_{m+1}U_{m+1}^* z}{\|U_{m+1}U_{m+1}^* z\|} = \frac{2\sigma_{\max}\sigma_{\min}}{\sigma_{\min}^2 + \sigma_{\max}^2},$$

where the last equality follows from [16, Theorem 3.1-2]. The result follows from using the bound above in $|\tan\theta| = \sqrt{\alpha^2 - 1}$. □

Acknowledgements. Work on this paper was supported in part by the National Science Foundation under grant DMS-0207525. Part of this paper was prepared while the second author visited the Università di Bologna, within the 2004 I.N.d.A.M. project "Modellistica Numerica per il calcolo scientifico e applicazioni avanzate". We thank J. Liesen for his questions and suggestions which helped improve our presentation and simplify some results. We also thank S. Serra-Capizzano and F. Di Benedetto for pointing to [3].

References

1. Axelsson, O., Vassilevski, P.S.: Variable-step multilevel preconditioning methods. I. Self-adjoint and positive definite elliptic problems. Numer. Linear Algebra Appl. **1**, 75–101 (1994)

2. Barrett, R. et al.: Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods. SIAM, Philadelphia, 1994
3. Di Benedetto, F., Fiorentino, G., Serra, S.: C.G. preconditioning for Toeplitz matrices. Computers Math. Applic. **25**, 35–45 (1993)
4. Chan, T.F., Chow, E., Saad, Y., Yeung, M.C.: Preserving symmetry in preconditioned Krylov subspace methods. SIAM J. Scientific Comput. **20**, 568–581 (1998)
5. Cullum, J.: Peaks, plateaux, and numerical stabilities in a Galerkin/minimal residual pair of methods for solving $Ax = b$. Appl. Numer. Math. **19**, 255–278 (1995)
6. Cullum, J.: Arnoldi versus nonsymmetric Lanczos algorithms for solving matrix eigenvalue problems. BIT **36**, 470–493 (1996)
7. Cullum, J.: Iterative methods for solving $Ax = b$, GMRES/FOM versus QMR/BiCG. Advances Comput. Math. **6**, 1–24 (1996)
8. Cullum, J., Greenbaum, A.: Relations between Galerkin and norm-minimizing iterative methods for solving linear systems. SIAM J. Matrix Anal. Appl. **17**, 223–247 (1996)
9. Freund, R.W.: Conjugate Gradient-type methods for linear systems with complex symmetric coefficient matrices. SIAM J. Scientific Comput. **13**, 425–448 (1992)
10. Freund, R.W., Nachtigal, N.M.: QMR: A quasi-minimal residual method for non-Hermitian linear systems. Numerische Mathematik **60**, 315–339 (1991)
11. Golub, G.H., Van Loan, C.F.: Matrix Computations. The John Hopkins University Press, Baltimore, Third Edition, 1996
12. Golub, G.H., Ye, Q.: Inexact preconditioned conjugate gradient method with inner–outer iteration. SIAM J. Sci. Comput. **21**, 1305–1320 (1999)
13. Greenbaum, A.: Iterative Methods for Solving Linear Systems. SIAM, Philadelphia, 1997
14. Greenbaum, A., Rozložník, M., Strakoš, Z.: Numerical Behaviour of the Modified Gram-Schmidt GMRES Implementation. BIT **37**, 706–719 (1997)
15. Greenbaum, A., Strakoš, Z.: Predicting the behavior of finite precision Lanczos and conjugate gradient computations. SIAM J. Matrix Anal. Appl. **13**, 121–137 (1992)
16. Gustafson, K.E., Rao, D.K.M.: Numerical Range: The Field of Values of Linear Operators and Matrices. Springer, New York, 1997
17. Higham, N.J.: The Matrix Computation Toolbox. Technical report, Manchester Centre for Computational Mathematics, 2002. In: www.ma.man.uc.uk/~higham/mctoolbox
18. Hochbruck, M., Lubich, C.: Error analysis of Krylov methods in a nutshell. SIAM J. Scientific Comput. **19**, 695–701 (1998)
19. Lanczos, C.: Solution of linear equations by minimized iterations. J. Res. Natl. Bur. Stand. **49**, 33–53 (1952)
20. Liesen, J., Rozložník, M., Strakoš, Z.: Least Squares Residuals and Minimal Residual Methods. SIAM J. Scientific Comput. **23**, 1503–1525 (2003)
21. Liesen, J., Strakoš, Z.: Convergence of GMRES for tridiagonal Toeplitz matrices. SIAM J. Matrix Anal. Appl. **26**, 233–251 (2004)
22. Liesen, J., Strakoš, Z.: GMRES convergence analysis for a convection-diffusion model problem. Technical Report 26-2003, Institute of Mathematics, Technical University of Berlin, 2003. To appear in SIAM J. Scientific Comput.
23. Morgan, R.: Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations. SIAM J. Matrix Anal. Appl. **21**, 1112–1135 (2000)
24. Notay, Y.: Flexible conjugate gradient. SIAM J. Scientific Comput. **22**, 1444–1460 (2000)
25. Paige, C.C., Strakoš, Z.: Residual and Backward Error Bounds in Minimum Residual Krylov Subspace Methods. SIAM J. Scientific Comput. **23**, 1899–1924 (2002)

26. Saad, Y.: Practical use of some Krylov subspace methods for solving indefinite and unsymmetric linear systems. SIAM J. Scientific Stat. Comput. **5**, 203–228 (1984)
27. Saad, Y.: Iterative Methods for Sparse Linear Systems. The PWS Publishing Company, Boston, 1996. Second Edition, SIAM, Philadelphia, 2003
28. Saad, Y., Schultz, M.H.: GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. SIAM J. Scientific Stat. Comput. **7**, 856–869 (1986)
29. Saad, Y., Wu, K.: DQGMRES: a direct quasi-minimal residual algorithm based on incomplete orthogonalization. Numerical Linear Algebra Appl. **3**, 329–343 (1996)
30. Simoncini, V.: A matrix analysis of Arnoldi and Lanczos methods. Numer. Math. **81**, 125–141 (1998)
31. Simoncini, V.: On the convergence of restarted Krylov subspace methods. SIAM J. Matrix Anal. Appl. **22**, 430–452 (2000)
32. Simoncini, V., Szyld, D.B.: Theory of Inexact Krylov Subspace Methods and Applications to Scientific Computing. SIAM J. Scientific Comput. **25**, 454–477 (2003)
33. Simoncini, V., Szyld, D.B.: On the Occurrence of Superlinear Convergence of Exact and Inexact Krylov Subspace Methods. SIAM Review, **47**, 247–272 (2005)
34. van der Vorst, H.A.: Iterative Krylov Methods for Large Linear Systems. Cambridge University Press, Cambridge, 2003
35. van der Vorst, H.A., Melissen, J.B.M.: A Petrov-Galerkin type method for solving $Ax = b$, where A is symmetric complex. IEEE Trans. Magnetics **26**, 706–708 (1990)