SIAM J. SCI. COMPUT.
Vol. 25, No. 2, pp. 454–477

THEORY OF INEXACT KRYLOV SUBSPACE METHODS AND APPLICATIONS TO SCIENTIFIC COMPUTING[*]

VALERIA SIMONCINI[†] AND DANIEL B. SZYLD[‡]

Abstract. We provide a general framework for the understanding of inexact Krylov subspace methods for the solution of symmetric and nonsymmetric linear systems of equations, as well as for certain eigenvalue calculations. This framework allows us to explain the empirical results reported in a series of CERFACS technical reports by Bouras, Frayssé, and Giraud in 2000. Furthermore, assuming exact arithmetic, our analysis can be used to produce computable criteria to bound the inexactness of the matrix-vector multiplication in such a way as to maintain the convergence of the Krylov subspace method. The theory developed is applied to several problems including the solution of Schur complement systems, linear systems which depend on a parameter, and eigenvalue problems. Numerical experiments for some of these scientific applications are reported.

Key words. inexact matrix-vector multiplies, Krylov subspace methods, inexact preconditioning, inner-outer methods, iterative nonsymmetric solver

AMS subject classifications. 65F10, 65F15, 15A06, 15A18

DOI. 10.1137/S1064827502406415

1. Introduction.

Consider the iterative solution of large sparse (symmetric or) nonsymmetric $n \times n$ linear systems of the form

$$Ax = b \tag{1.1}$$

with a Krylov subspace method. In a series of papers, Bouras and Frayssé [3], [4] and Bouras, Frayssé, and Giraud [5] reported experiments in which the matrix-vector multiplication with A (at each step of the Krylov subspace method) is not performed exactly. Instead of the exact matrix-vector multiplication Av, the product

$$\mathcal{A}v = (A + E)v \tag{1.2}$$

is computed, where E is an error matrix which changes every time the operator is applied.

In the experiments reported in the mentioned papers, and also in [14], [32], the norm $\|E\|$ is allowed to grow as the Krylov iteration progresses, without apparent degradation of the convergence of the iterative method. This counterintuitive situation is in contrast to other inexact or two-stage methods, where the inner tolerance has to stay at least constant [12], or it needs to decrease as the iterates get closer to the solution [7], [30].

There are many scientific applications where the inexact matrix-vector product (1.2) appears naturally. For example, when operating with A implies a solution of a linear system, as is the case in Schur complement computations (see, e.g., [19], [20], [37]), and in certain eigenvalue algorithms [16], [32], or when the matrix is very large

[*] Received by the editors April 25, 2002; accepted for publication (in revised form) February 9, 2003; published electronically November 11, 2003.
http://www.siam.org/journals/sisc/25-2/40641.html

[†] Dipartimento di Matematica, Università di Bologna, and Istituto di Matematica Applicata e Tecnologie Informatiche del CNR, Pavia, Italy (val@dragon.ian.pv.cnr.it).

[‡] Department of Mathematics, Temple University (038-16), 1805 N. Broad Street, Philadelphia, PA 19122-6094 (szyld@math.temple.edu). This author's research was supported in part by the National Science Foundation under grant DMS-0207525.

(and/or dense), and a reasonable approximation can be used [2], [11]. Several authors have studied different aspects of the use of inexact matrix-vector multiplication in iterative methods, sometimes in the context of small perturbations, and in some other instances allowing for large tolerances (though not letting them grow); see, e.g., [13], [14], [15], [38].

The problem we consider in this paper is the case in which $\|E\|$ can be monitored, usually through an additional (inner) tolerance. We are interested in evaluating how large $\|E\|$ can be at each step while still achieving convergence of the Krylov subspace method to the sought-after solution x. In [3], [4], [5], and [41], ad hoc criteria were used to determine an appropriate inner tolerance.

In this paper, we address essentially three questions:

1. What are the variational properties of the computed approximate solution and associated residual when using inexact matrix-vector products?
2. Can we explain why $\|E\|$ can be allowed to grow?
3. Can we give computable criteria to bound $\|E\|$ at each step (i.e., criteria for the inner tolerance) of the Krylov subspace method, so that the global residual norm falls below a prescribed tolerance?

Our answers to these questions provide the theory for the experiments in [3], [4], [5], [41], as well as more general computable criteria than the ones hitherto proposed.

We present a general framework for inexact Krylov subspace methods, including FOM (full orthogonalization method) (or CG), Lanczos, MINRES (minimal residual), GMRES, and QMR. For general description of these methods, see, e.g., [17], [29]. The main results of this paper were first presented in [35] and were developed independently of similar results in [36], where another point of view is adopted; see section 6 for some details.

In the next section, we begin with the analysis of the inexact Krylov subspace methods. In section 3 we answer question 1. In section 4 we produce some bounds on the true residual of the methods, setting the stage to tackle questions 2 and 3 in section 5. In section 6 we comment on related inexact implementations. In later sections, we apply the theory we developed to several scientific applications and report on some computational experiments.

We note that even though we show that the matrix-vector multiplication may be performed in an increasingly inexact or approximate manner as the iteration progresses and still allow convergence to the solution of the original problem (1.1), the rate of convergence of the inexact Krylov subspace method may deteriorate in comparison to that with exact matrix-vector products. Example 5.6 illustrates this.

In section 9 we discuss flexible Krylov subspace methods, where the matrix-vector multiplication is exact, but the preconditioner is allowed to change; see, e.g., [34] and the references given therein. These methods can be combined with the inexact Krylov methods treated in this paper, and in fact this is used in [41].

Matlab notation is used throughout the paper. Given a vector u, u^T denotes its transpose, while $(u)_{k:m}$ denotes its kth through mth components; its kth single component is denoted by $(u)_k$. An analogous notation is used for matrices. Vector subscripts are instead used for recurrence iterates, except for e_k which is the kth column of the identity. We denote by I_k the $k \times k$ identity matrix; we omit the subscript when the dimension is clear from the context. By $\sigma_k(H)$ we denote the kth singular value of matrix $H \in \mathbb{R}^{m_1 \times m_2}$, while $\kappa(H) = \sigma_1(H)/\sigma_p(H)$, with $p = \min(m_1, m_2)$ and H full rank. By $\mathcal{R}(V)$ we denote the range of the matrix V. The 2-norm is used for vectors and the induced 2-norm is employed for matrices. The only exception is when one considers QMR; see Remark 3.1. Our analysis is presented in real arithmetic. Nonetheless, our theory can be equally stated in complex arithmetic; our numerical experiments in Example 10.2 were actually done using complex arithmetic. Exact arithmetic is assumed, although some results can be applied to finite precision arithmetic as well; see in particular Remark 4.4. Note that except for the phrase "exact arithmetic," when we use the word "exact" throughout the paper, it is in contrast to "inexact," for example, when we talk about the exact FOM method vs. the inexact FOM method.

2. General description of inexact Krylov subspace methods.

We assume that the reader is familiar with Krylov subspace methods and refer to the mentioned references [17], [29] for their description. In these iterative methods, such as CG, Lanczos, MINRES, GMRES, FOM, one finds at the mth iteration an approximation to the solution of (1.1) in a Krylov subspace $\mathcal{K}_m(A, v_1)$ spanned by the first m vectors of the Krylov sequence $v_1, Av_1, A^2v_1, \ldots$. At the mth step one has a typical relation

$$AV_m = V_{m+1}H_m, \tag{2.1}$$

where $V_m = [v_1, v_2, \ldots, v_m]$ and $v_1 = b/\beta$, $\beta = \|b\|$ $(x_0 = 0)$, and the $(m+1) \times m$ matrix H_m is either tridiagonal or upper Hessenberg. For certain methods, such as FOM or GMRES, the matrix V_m is orthogonal; for others, such as two-sided Lanczos or QMR, its columns are biorthogonal to a basis of another Krylov subspace generated with matrix-vector products with A^T. In all cases, we can assume $\|v_k\| = 1$, $k = 1, \ldots, m$. The relation (2.1) is called an Arnoldi relation when V_m is orthogonal.

When the matrix-vector multiplication is not exact, the relation (2.1) does not hold. The left-hand matrix needs to be replaced by $[(A+E_1)v_1, (A+E_2)v_2, \ldots, (A+E_m)v_m]$ so that (2.1) becomes

$$\begin{aligned} [(A+E_1)v_1, (A+E_2)v_2, \ldots, (A+E_m)v_m] &= V_{m+1}H_m, \\ AV_m + [E_1v_1, E_2v_2, \ldots, E_mv_m] &= V_{m+1}H_m. \end{aligned} \tag{2.2}$$

The subspace $\mathcal{R}(V_m)$ is no longer a Krylov subspace generated by A. However, V_m has the same properties as in the exact method, i.e., it is an orthogonal matrix in the case of FOM, GMRES, etc., or its columns are biorthogonal to a basis of another subspace generated using A^T. In particular, each basis vector v_m is explicitly (bi)orthogonalized with respect to all previous basis vectors so that orthogonality properties are preserved; see section 6 for further comments. We also point out that even in the case that A and all perturbation matrices $E_1, \ldots, E_m$ are symmetric, in general, the inexact H_m in (2.2) no longer maintains symmetry. We call (2.2) an inexact Arnoldi relation.

In most of our analysis we use the expression (2.2). Another way to look at the inexact Arnoldi relation when V_m is orthogonal is to write

$$(A + \mathcal{E}_m)V_m = V_{m+1}H_m, \qquad \mathcal{E}_m = \sum_{k=1}^{m} E_k v_k v_k^T. \tag{2.3}$$

The relation (2.3) shows that V_m is a basis for a subspace spanned by elements of a Krylov sequence obtained by a (possibly large) perturbation of the matrix A, where this perturbation is updated at each iteration, i.e., we have that

$$\mathcal{R}(V_m) = \mathcal{K}_m(A + \mathcal{E}_m, v_1). \tag{2.4}$$

We end this section with a brief comment on the techniques not used in this paper. The expression (2.3) indicates that when V_m is orthogonal one could analyze the inexact Krylov subspace method considering the consequences of a perturbation of the matrix A, e.g., as in [38]. Since the norm of the perturbation is indeed large, the usual techniques cannot be used. In addition, neither the use of orthogonal polynomials nor the consideration of the spectrum of the matrices as driving the iterative methods can be applied in this context [17], [29]. The large norm of the perturbations considered here, and the fact that they change from step to step, may also limit the applicability of pseudospectra analysis [40].

3. Properties of the computed approximate solution.

General Krylov subspace methods produce their approximation to the solution of (1.1) by either a minimization (or in some cases quasi minimization) or a projection procedure over a subspace of the form $A\mathcal{K}_m(A, v_1)$; cf. [9]. In the inexact case, the appropriate subspace is $\mathcal{R}(W_m)$, where $W_m = V_{m+1}H_m$; cf. (2.2) or (2.3). In fact, we show that given $r_0 = b-(A+E_0)x_0$, the inexact Krylov subspace methods determine an approximation to r_0 in $\mathcal{R}(W_m)$ by either of the following.

1. *Oblique projection.* Find $q_m \in \mathcal{R}(W_m)$ such that

$$r_0 - q_m \perp \mathcal{R}(Y_m), \tag{3.1}$$

where $Y_m = V_m$ when V_m is orthogonal, or else it represents the biorthogonal basis, i.e., $V_m^T Y_m$ is a diagonal matrix. Note that this is a general Petrov–Galerkin projection method; see, e.g., [29].

2. *Minimization procedure.* Solve

$$\min_{q \in \mathcal{R}(W_m)} \|r_0 - q\|. \tag{3.2}$$

Remark 3.1. The descriptions of the methods so far apply to a variety of Krylov subspace methods. For clarity of exposition, in the rest of the paper, we specialize in the FOM method (as well as CG) and in GMRES. Nevertheless, our results have wider application. In particular, the QMR method satisfies (3.2) using a different norm [9].

An inexact minimal residual method, such as GMRES, determines a solution $z_m = V_m y_m$ for y_m the minimizer of

$$\min_y \|e_1\beta - H_m y\|, \tag{3.3}$$

where $v_1 = r_0/\beta$, $\beta = \|r_0\|$. Unlike the exact case, this minimization does not imply a minimization of the residual $b - Az_m$. We next show that z_m is associated with a minimization procedure as in (3.2).

PROPOSITION 3.2. *Let $v_1 = r_0/\beta$, $\beta = \|r_0\|$. Let y_m^{gm} be the solution of the GMRES equation* (3.3) *at the mth step, i.e., the solution of $H_m^T H_m y = H_m^T e_1\beta$, and let the inexact Arnoldi relation* (2.2) *hold. Then $q_m = W_m y_m^{gm}$ satisfies* (3.2).

Proof. We have that since $W_m = V_{m+1}H_m$, with $V_{m+1}^T V_{m+1} = I$, y_m^{gm} solves

$$W_m^T(r_0 - q_m) = H_m^T e_1\beta - H_m^T H_m y_m^{gm} = 0.$$

Thus we have that $W_m^T(r_0 - q_m) = 0$, that is, $r_0 - q_m \perp \mathcal{R}(W_m)$, so that $q_m = \arg\min_{q\in\mathcal{R}(W_m)} \|r_0 - q\|$, and the proof is complete. $\square$

The exact FOM method determines z_m as $z_m = V_m y_m$ with y_m the solution of

$$\widehat{H}_m y_m = e_1\beta, \tag{3.4}$$

where $\widehat{H}_m = [I, 0]H_m$ is the principal square part of H_m. In the inexact case, the matrix H_m is as in (2.2), and z_m is computed as in the exact case, that is, $z_m = V_m\widehat{H}_m^{-1}e_1\beta$. However, z_m no longer satisfies a Galerkin condition on the residual $b - Az_m$ in $\mathcal{K}_m(A, r_0)$. We characterize now this solution in terms of the subspace $\mathcal{R}(W_m)$.

Proposition 3.3. *Let $v_1 = r_0/\beta$, $\beta = \|r_0\|$. Let y_m^{fom} be the solution of the FOM equation* (3.4)*, and let the inexact Arnoldi relation* (2.2) *hold. Then $q_m = W_m y_m^{fom}$ satisfies* (3.1).

Proof. The FOM solution y_m^{fom} solves (3.4), and since $q_m = V_{m+1}H_m y_m^{fom}$, we have $V_m^T(r_0 - q_m) = e_1\beta - \widehat{H}_m y_m^{fom} = 0$. This corresponds to $r_0 - q_m \perp \mathcal{R}(V_m)$, and the proof is complete. ▯

The approximation error $r_0 - q_m \equiv r_0 - V_{m+1}H_m y_m^{fom}$ satisfies

$$r_0 - q_m = r_0 - V_{m+1}H_m\widehat{H}_m^{-1}e_1\beta = -v_{m+1}h_{m+1,m}e_m^T y_m^{fom}. \tag{3.5}$$

It follows from (3.5) that as long as the subspace $\mathcal{R}(W_m)$ keeps growing, the norm of the error in the approximation, namely $\|\tilde{r}_m\| = \|r_0 - q_m\| = |h_{m+1,m}e_m^T y_m|$, decreases, although possibly not monotonically, and it eventually becomes zero for $m = n$. Unless $h_{m+1,m}$ becomes small, which occurs when $\{v_1, v_2, \dots, v_m, (A + E_m)v_m\}$ are almost linearly dependent, this means that $|e_m^T y_m|$ keeps decreasing. We conclude that the norm of the computed residual $\|\tilde{r}_m\|$ converges to zero. This can be clearly seen in the examples presented in later sections.

If $\{v_1, v_2, \dots, v_m, (A + E_m)v_m\}$ are linearly dependent, then $h_{m+1,m} = 0$ and we have breakdown, just as in the case of exact matrix-vector products, although in the inexact case breakdown does not necessarily imply convergence. Also note that since $\|E_m\|$ may be large, it is possible that for some m, $A + E_m$ is singular. This in itself does not cause breakdown, as long as $(A + E_m)v_m \notin \mathcal{R}(V_m)$.

We end this section with a comment on the accuracy required to compute the initial residual r_0. It was observed in [3] that r_0 should be computed with high accuracy; see also [5]. The same applies to the initial residual at each restarting phase of a restarted method, such as restarted GMRES. Our discussion in this section gives an explanation for such a requirement. Thus let x_0 be a starting guess and set $r_0 = b - (A + E_0)x_0$. If $x_m = x_0 + z_m$ is an approximate solution obtained with the inexact method, then

$$r_m = b - Ax_m = b - Ax_0 - Az_m = E_0x_0 + (r_0 - Az_m).$$

The inexact method aims to approximate r_0 in $\mathcal{R}(V_{m+1}H_m)$, and thus the term E_0x_0 remains untouched by the inexact procedure; hence, its norm should be smaller than the required convergence tolerance. Keeping this consideration in mind, in the following we shall always work with r_0 assuming that $\|E_0x_0\|$ is sufficiently small; at the start-up phase, this is easily achieved by setting x_0 equal to the zero vector.

4. Bounds on the true residual.

Let the inexact Arnoldi relation (2.2) hold, and let $z_m = V_m y_m$ be the approximate solution to $Az = r_0$ obtained by projecting r_0 onto $\mathcal{R}(V_m)$. Then the true residual at the mth iteration is

$$r_m = r_0 - AV_m y_m = (r_0 - V_{m+1}H_m y_m) + [E_1v_1, \dots, E_mv_m]y_m \tag{4.1}$$

$$= \tilde{r}_m + [E_1v_1, \dots, E_mv_m]y_m, \tag{4.2}$$

which defines the distance between the computed and true residuals (see also [36]) as

$$\delta_m := \|r_m - \tilde{r}_m\| = \|[E_1 v_1, \ldots, E_m v_m] y_m\|. \tag{4.3}$$

The next result will be applied to inexact GMRES and FOM (and implicitly to the other methods) and is the basis for our answers to questions 2 and 3.

Proposition 4.1. *Assume that m iterations of the inexact Arnoldi method have been carried out. Let $y_m = [\eta_1^{(m)}, \eta_2^{(m)}, \ldots, \eta_m^{(m)}]^T$. Then*

$$\delta_m = \|r_m - \tilde{r}_m\| \le \sum_{k=1}^{m} |\eta_k^{(m)}| \, \|E_k\|,$$

where r_m and $\tilde{r}_m$ are defined in (4.1) *and* (4.2)*, respectively. Moreover, let $U_m \in \mathbb{R}^{n\times m}$ be full column rank. If $\tilde{r}_m \perp \mathcal{R}(U_m)$, then*

$$\|U_m^T r_m\| \le \|U_m\| \sum_{k=1}^{m} |\eta_k^{(m)}| \, \|E_k\|.$$

Proof. The first result follows from (4.3) and the following bound, where the fact that $\|v_k\| = 1, k = 1, \ldots, m$, is used:

$$\|[E_1 v_1, \ldots, E_m v_m] y_m\| = \left\| \sum_{k=1}^{m} E_k v_k \eta_k^{(m)} \right\| \le \sum_{k=1}^{m} \|E_k\| \, |\eta_k^{(m)}|. \tag{4.4}$$

By using (4.1) together with the orthogonality relation $U_m^T \tilde{r}_m = 0$, we obtain

$$\|U_m^T r_m\| = \|U_m^T [E_1 v_1, \ldots, E_m v_m] y_m\| \le \|U_m\| \, \|[E_1 v_1, \ldots, E_m v_m] y_m\|.$$

Applying the bound (4.4), the result follows. □

We next specialize the result above to the cases of GMRES and FOM. Corollary 4.2 applies Proposition 4.1 to the GMRES solution, with $U_m = W_m = V_{m+1} H_m$.

Corollary 4.2. *Assume that m iterations of the inexact Arnoldi method have been carried out. Let $W_m = V_{m+1} H_m$ and $y_m^{gm} = \arg\min_y \|e_1 \beta - H_m y\|$, $y_m^{gm} = [\eta_1^{(m)}, \eta_2^{(m)}, \ldots, \eta_m^{(m)}]^T$. Then the true GMRES residual $r_m^{gm} = b - A V_m y_m^{gm}$ satisfies*

$$\delta_m = \|r_m^{gm} - \tilde{r}_m^{gm}\| \le \sum_{k=1}^{m} |\eta_k^{(m)}| \, \|E_k\|, \quad \|W_m^T r_m^{gm}\| \le \|H_m\| \sum_{k=1}^{m} |\eta_k^{(m)}| \, \|E_k\|, \tag{4.5}$$

where $\tilde{r}_m^{gm}$ is the computed residual $\tilde{r}_m^{gm} = r_0 - V_{m+1} H_m y_m^{gm}$.

The following result can be obtained by applying Proposition 4.1 to the FOM solution, with $U_m = V_m$.

Corollary 4.3. *Assume that after m iterations of the inexact Arnoldi method, the FOM solution $y_m^{fom} = \widehat{H}_m^{-1} e_1 \beta$ exists. Then the true FOM residual $r_m^{fom} = b - A V_m y_m^{fom}$ satisfies*

$$\delta_m = \|r_m^{fom} - \tilde{r}_m^{fom}\| \le \sum_{k=1}^{m} |\eta_k^{(m)}| \, \|E_k\|, \quad \|V_m^T r_m^{fom}\| \le \sum_{k=1}^{m} |\eta_k^{(m)}| \, \|E_k\|, \tag{4.6}$$

where $y_m^{fom} = [\eta_1^{(m)}, \eta_2^{(m)}, \ldots, \eta_m^{(m)}]^T$ and $\tilde{r}_m^{fom} = r_0 - V_{m+1} H_m y_m^{fom}$.

It follows directly from (4.6) and (4.5) that if one wants δ_m and/or $V_m^T r_m^{fom}$ (or $W^T r_m^{gm}$) to be small, what matters in these bounds is that the product $|\eta_k^{(m)}|\,\|E_k\|$ is small but not necessarily both factors! Thus, if $|\eta_k^{(m)}|$ is decreasing (as we establish in the next section), then $\|E_m\|$ can be allowed to grow as the iteration progresses, hence the accuracy of the matrix-vector product can be relaxed. This will explain the issue raised in question 2.

Remark 4.4. The bound (4.4) is often sharper than

$$\|[E_1v_1,\ldots,E_mv_m]y_m\| \le \|[E_1v_1,\ldots,E_mv_m]\|\,\|y_m\|.$$

This suggests that the techniques used here may lead to improvements in the bounds used in round-off error analysis of Krylov subspace methods; see, e.g., [17, Chap. 4].

From $\|V_m^T r_m^{fom}\| \le \|r_m^{fom}\|$ it follows that $\|V_m^T r_m^{fom}\|$ represents a lower bound for the attainable final norm of the true residual. It is thus important to be able to monitor such a quantity. We do this in the experiments with inexact FOM reported in later sections.

The true residual r_m may not be computationally available; however, the simple relation $\|r_m\| \le \|\tilde{r}_m\| + \delta_m$ together with (4.6) or (4.5) provides the computable bound

$$\begin{aligned} \|r_m\| &\le \|\tilde{r}_m\| + \|[E_1v_1,\ldots,E_mv_m]y_m\| \\ &\le \|\tilde{r}_m\| + \sum_{k=1}^{m} |\eta_k^{(m)}|\,\|E_k\|. \end{aligned} \tag{4.7}$$

When considering backward errors, the ratio $\|r_m\|/(\|A\|\,\|V_my_m\|)$ is the quantity that one would like to monitor. Plugging this into (4.7) and using the fact that V_m is orthogonal, we get

$$\frac{\|r_m\|}{\|A\|\,\|V_my_m\|} \le \frac{\|\tilde{r}_m\|}{\|A\|\,\|y_m\|} + \frac{\|[E_1v_1,\ldots,E_mv_m]y_m\|}{\|A\|\,\|y_m\|}\ .$$

Following backward error analysis, the final attainable residual norm should be monitored in terms of the relative quantity $\|[E_1v_1,\ldots,E_mv_m]y_m\|/(\|A\|\,\|y_m\|)$. It is thus natural to impose relative conditions on the norm of the error matrices. Given $\epsilon \in (0,1)$, we can consider perturbation matrices $E_1,\ldots,E_m$, satisfying

$$\|[E_1v_1,\ldots,E_mv_m]y_m\| \le \epsilon\|A\|, \qquad m = 1,2,\ldots.$$

To simplify the presentation, in this paper we shall always incorporate the norm of A in the relaxation parameter, thus defining $\varepsilon = \epsilon\|A\|$, $\varepsilon \in (0,\|A\|)$.

5. Relaxing the accuracy of the matrix-vector product.

We proceed now with questions 2 and 3, using as starting points Corollaries 4.2 and 4.3. We want to prescribe a relaxation strategy, i.e., a maximum value for $\|E_k\|$, $k = 1,\ldots,m$, so that $\sum_{k=1}^m |\eta_k^{(m)}|\,\|E_k\| \le \varepsilon$ (or $\le \varepsilon/\|H_m\|$) and thus, from (4.5) and (4.6), both δ_m and the norm of the projected residuals are less than ε.

From now on we drop the superscript in $\eta_k^{(m)}$, but the reader should remember that $\eta_k = e_k^T y_m$, $k = 1,\ldots,m$, are computed at the mth iteration, hence they are not available when we need to bound the matrices E_k, $k = 1,\ldots,m$, earlier in the process. We next provide bounds for η_k that depend on quantities at iteration $k-1$.

LEMMA 5.1. *Assume that m iterations of the inexact Arnoldi method have been carried out and let y_m^{gm} be the GMRES solution. Then, for any $k = 1, \ldots, m$,*

$$|\eta_k^{gm}| = |(y_m^{gm})_k| \leq \frac{1}{\sigma_m(H_m)} \|\tilde{r}_{k-1}^{gm}\|,$$

where $\tilde{r}_{k-1}^{gm} = r_0 - V_{k-1}H_{k-1}y_{k-1}^{gm}$.

Proof. We eliminate the superscripts that denote the GMRES method. Let $Q_m[R_m^T, 0]^T = H_m$ be the QR factorization of H_m, with

$$Q_m^T = \Omega_m \Omega_{m-1} \cdots \Omega_1 \in \mathbb{R}^{(m+1)\times(m+1)}, \tag{5.1}$$

$$\Omega_k = \begin{bmatrix} I_{k-1} & & & \\ & c_k & s_k & \\ & -s_k & c_k & \\ & & & I_{m-k} \end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}, \tag{5.2}$$

where c_k, s_k are the sines and cosines of the Givens rotations to annihilate the corresponding elements $h_{k+1,k}$ of $H_m = (h_{i,j})_{\substack{i=1,\ldots,m+1 \\ j=1\ldots,m}}$, and blanks in (5.2) indicate zero entries [29]. Let

$$g_m = [\gamma_1, \ldots, \gamma_m]^T \tag{5.3}$$

be the vector of the first m components of $Q_m^T(\beta e_1)$. We have $|\gamma_k| = |c_k s_1 s_2 \cdots s_{k-1}\beta|$; see, e.g., [29, eq. (6.42)]. Moreover, $\|\tilde{r}_{k-1}\| = \beta|s_1 s_2 \cdots s_{k-1}|$; see, [6, eq. (4.4)] or [29, eq. (6.48)].

Since $y_m = R_m^{-1}g_m$ [29, Proposition 6.9], and since R_m is upper triangular implying that R_m^{-1} is also upper triangular, then

$$|\eta_k| = |(R_m^{-1})_{k,1:m}g_m| \leq \|(R_m^{-1})_{k,k:m}\| \, \|(g_m)_{k:m}\| = \|e_k^T R_m^{-1}\| \, \|(g_m)_{k:m}\|.$$

Therefore

$$\begin{aligned} \|(g_m)_{k:m}\|^2 &= \gamma_k^2 + \gamma_{k+1}^2 + \cdots + \gamma_m^2 \\ &= \beta^2 \left(|c_k s_1 s_2 \cdots s_{k-1}|^2 + |c_{k+1} s_1 s_2 \cdots s_k|^2 + \cdots + |c_m s_1 s_2 \cdots s_{m-1}|^2\right) \\ &= \beta^2 |s_1 s_2 \cdots s_{k-1}|^2 \left(|c_k|^2 + |c_{k+1}s_k|^2 + \cdots + |c_m s_k s_{k+1} \cdots s_{m-1}|^2\right) \\ &= \|\tilde{r}_{k-1}\|^2 \left(|c_k|^2 + |c_{k+1}s_k|^2 + \cdots + |c_m s_k s_{k+1} \cdots s_{m-1}|^2\right). \end{aligned}$$

We note that $\|[c_k, c_{k+1}s_k, \ldots, c_m s_k s_{k+1} \cdots s_{m-1}]\| = \|[I_m, 0]\Omega_m \cdots \Omega_{k+1}\Omega_k e_k\|$, thus $|c_k|^2 + |c_{k+1}s_k|^2 + \cdots + |c_m s_k s_{k+1} \cdots s_{m-1}|^2 \leq 1$, from which

$$|\eta_k| \leq \|e_k^T R_m^{-1}\| \, \|\tilde{r}_{k-1}\| \leq \|R_m^{-1}\| \, \|\tilde{r}_{k-1}\|. \tag{5.4}$$

Noticing that $\|R_m^{-1}\| = (\sigma_m(H_m))^{-1}$, the result follows. □

An analogous relation holds for the FOM approximation.

LEMMA 5.2. *Assume that m iterations of the inexact Arnoldi method have been carried out and that the FOM solution $y_m^{fom} = \widehat{H}_m^{-1}e_1\beta$ exists. Then, for any $k = 1, \ldots, m$,*

$$|\eta_k^{fom}| = |(y_m^{fom})_k| \leq \frac{1}{\sigma_m(\widehat{H}_m)} \|\tilde{r}_{k-1}^{gm}\|, \tag{5.5}$$

where $\tilde r_{k-1}^{gm} = r_0 - V_{k-1}H_{k-1}y_{k-1}^{gm}$ *is the GMRES residual as in Lemma* 5.1*. Moreover, if for any* $k = 1,\ldots,m-1$ *the FOM residual* $\tilde r_{k-1}^{fom} = r_0 - V_{k-1}H_{k-1}y_{k-1}^{fom}$ *exists, then*

$$|\eta_k^{fom}| = |(y_m^{fom})_k| \le \frac{1}{\sigma_m(\widehat H_m)}\,\|\tilde r_{k-1}^{fom}\|. \tag{5.6}$$

Proof. We eliminate the superscripts that denote the FOM method. The proof is strictly related to that of Lemma 5.1; therefore, we use the same matrices and notation. Let $Q_m[R_m^T, 0]^T = H_m$ be the QR factorization of H_m with Q_m as in (5.1). Then $\widehat R_m = Q_{m-1}^T \widehat H_m$ is an $m\times m$ upper triangular matrix and $y_m = \widehat R_m^{-1}Q_{m-1}^T e_1\beta$ [29]. Let $\hat g_m = Q_{m-1}^T e_1\beta$. Then $\hat g_m = [g_{m-1}^T, \gamma_m/c_m]^T$ with g_{m-1} defined as in (5.3). Therefore

$$|(y_m)_k| = |(\widehat R_m^{-1})_{k,1:m}\hat g_m| \le \|(\widehat R_m^{-1})_{k,k:m}\|\,\|(\hat g_m)_{k:m}\| = \|e_k^T\widehat R_m^{-1}\|\,\|(\hat g_m)_{k:m}\|. \tag{5.7}$$

Moreover,

$$\|(\hat g_m)_{k:m}\|^2 = \beta^2|s_1\cdots s_{k-1}|^2\left(|c_k|^2 + |c_{k+1}s_k|^2 + \cdots + \frac{1}{c_m^2}|c_m s_k s_{k+1}\cdots s_{m-1}|^2\right).$$

We have $\|[c_k, c_{k+1}s_k, \ldots, s_k s_{k+1}\cdots s_{m-1}]\| = \|\Omega_{m-1}\cdots\Omega_{k+1}\Omega_k e_k\| = 1$. Moreover, $\beta|s_1\cdots s_{k-1}| = \|\tilde r_{k-1}^{gm}\|$. Substituting in (5.7) and using $\|e_k^T\widehat R_m^{-1}\| \le \|\widehat R_m^{-1}\| = 1/\sigma_m(\widehat H_m)$, we see that bound (5.5) follows. If $c_{k-1}\ne 0$, which holds when $\widehat H_{k-1}$ is nonsingular [29], then the FOM solution y_{k-1}^{fom} is defined, and using $\|\tilde r_{k-1}^{gm}\| = |c_{k-1}|\,\|\tilde r_{k-1}\| \le \|\tilde r_{k-1}\|$, we see that bound (5.6) follows. □

We note that the bound (5.5) is sharper than (5.6); moreover, the computed GMRES residual is always defined for any $k = 1,\ldots,m$, and its norm is cheaply available during the FOM computation if desired.

The estimates of Lemmas 5.1 and 5.2 provide a dynamic criterion for relaxing the accuracy in the application of the operator $\mathcal{A}v = (A+E_k)v$.

Theorem 5.3. *Let* $\varepsilon > 0$*. Let* $r_m^{gm} = r_0 - Az_m^{gm}$ *be the GMRES residual after* m *iterations of the inexact Arnoldi method. Under the same hypotheses and notation of Lemma* 5.1*, if for every* $k\le m$*,*

$$\|E_k\| \le \frac{\sigma_m(H_m)}{m}\frac{1}{\|\tilde r_{k-1}^{gm}\|}\varepsilon, \tag{5.8}$$

then $\|r_m^{gm} - \tilde r_m^{gm}\| \le \varepsilon$*. Moreover, if*

$$\|E_k\| \le \frac{1}{m\kappa(H_m)}\frac{1}{\|\tilde r_{k-1}^{gm}\|}\varepsilon,$$

then $\|(V_{m+1}H_m)^T r_m^{gm}\| \le \varepsilon$*.*

Proof. Let $z_m^{gm} = V_m y_m^{gm}$ with $y_m^{gm} = [\eta_1,\ldots,\eta_m]^T$. From Lemma 5.1, $|\eta_k| \le \sigma_m(H_m)^{-1}\,\|\tilde r_{k-1}^{gm}\|$, $k = 1,\ldots,m$. Therefore, if for each k,

$$\|E_k\| \le \frac{\sigma_m(H_m)\varepsilon}{m\|H_m\|\,\|\tilde r_{k-1}^{gm}\|},$$

then the bound $\sum_{k=1}^m |\eta_k^{(m)}|\,\|E_k\| \le \varepsilon/\|H_m\|$ holds. By Corollary 4.2, we then have $\|(V_{m+1}H_m)^T r_m^{gm}\| \le \varepsilon$. Similarly, if (5.8) holds, we have $|\eta_k|\|E_k\| \le \varepsilon/m$, $k = 1,\ldots,m$, and thus $\|r_m^{gm} - \tilde r_m^{gm}\| \le \varepsilon$. □

A corresponding result holds for the FOM approximate solution. The proof is analogous to that of the previous theorem and is therefore omitted.

THEOREM 5.4. *Let $r_m^{fom} = r_0 - Az_m^{fom}$ be the FOM residual after m iterations. Let $\varepsilon > 0$. Under the same hypotheses and notation of Lemma* 5.2*, if for every $k \le m$,*

$$\|E_k\| \le \frac{\sigma_m(\widehat{H}_m)}{m} \frac{1}{\|\tilde{r}_{k-1}^{gm}\|}\varepsilon, \tag{5.9}$$

then $\|r_m^{fom} - \tilde{r}_m^{fom}\| \le \varepsilon$ and $\|V_m^T r_m^{fom}\| \le \varepsilon$.

As noted in Lemma 5.2, if the FOM residuals exist for all $k \le m$, then we can replace $\|\tilde{r}_{k-1}^{gm}\|$ by $\|\tilde{r}_{k-1}^{fom}\|$ in (5.9). In the experiments reported in later sections we use this bound since the FOM residuals do exist and are directly available.

Observe that the magnitude of $\sigma_m(\widehat{H}_m)$ may vary considerably with m, showing that FOM may be very sensitive to the use of an inexact operator $\mathcal{A}$; see, e.g., [6, section 6], for an example where $\widehat{H}_m$ is singular at every other step, and also [23].

In both GMRES and FOM methods, we have determined a condition on $\|E_k\|$ of the type

$$\|E_k\| \le \ell_m \frac{1}{\|\tilde{r}_{k-1}\|}\varepsilon, \tag{5.10}$$

which guarantees overall convergence below the given tolerance. In [3] a similar condition is proposed, where the factor ℓ_m is given the value one. It was shown in [3] as well as in [5] that setting $\ell_m = 1$ is not harmful in many circumstances. Nonetheless, in some applied problems, setting $\ell_m = 1$ causes the true residual to reach its final attainable accuracy above the requested tolerance; cf. (4.7) and see [5], [41]. Several of our experiments below also indicate that a value of ℓ_m which takes into account the information on A or H_m is needed.

Theorems 5.3 and 5.4 are of important theoretical value, but since matrix H_m is not available after $k < m$ iterations, they fail to provide us with a truly computable criterion to bound $\|E_k\|$. Moreover, H_m differs depending on the (amount of) perturbation occurring in the operation with $\mathcal{A}v$ during the m steps. Depending on the problem at hand, one needs to provide estimates for quantities such as $\sigma_m(H_m)$. We do not have a full answer to the question of how to best obtain these estimates. We offer some possibilities below; see also comments in Examples 8.2 and 9.3.

First, a simple compromise consists of replacing m with the maximum number of iterations allowed, $m_\star$. Second, at least for exact GMRES, $\sigma_m(H_m)$ may be bounded by $\sigma_n(A)$. For inexact GMRES the bound would be $\sigma_n(A + \mathcal{E}_m)$, though this is hard to estimate. Some bounds can be obtained directly from (2.2), namely $\|H_m\| \le \|A\| + \|[E_1 v_1, E_2 v_2, \ldots, E_m v_m]\|$ and

$$\sigma_m(H_m) \ge \sigma_n(A) - \|[E_1 v_1, E_2 v_2, \ldots, E_m v_m]\|. \tag{5.11}$$

In most experiments reported here we have used $\sigma_n(A)$ to estimate $\sigma_{m_\star}(H_{m_\star})$, and this turned out to be sufficient for providing a practical bound. We have found that condition (5.10) is in fact stricter than necessary. This is a consequence of the nontightness of the bounds used in our derivation, for example, in (4.4) and (5.4). Therefore, estimates for $\sigma_m(H_m)$ that are not so sharp would still be very useful. The same applies to estimates of $\sigma_n(A)$ when this quantity is not readily available.

A further question is whether it is possible to prevent $\sigma_{m_\star}(H_{m_\star})$ from getting too small. Indeed, for large enough perturbations, $(A + \mathcal{E}_m)V_m$ may be (almost) rank

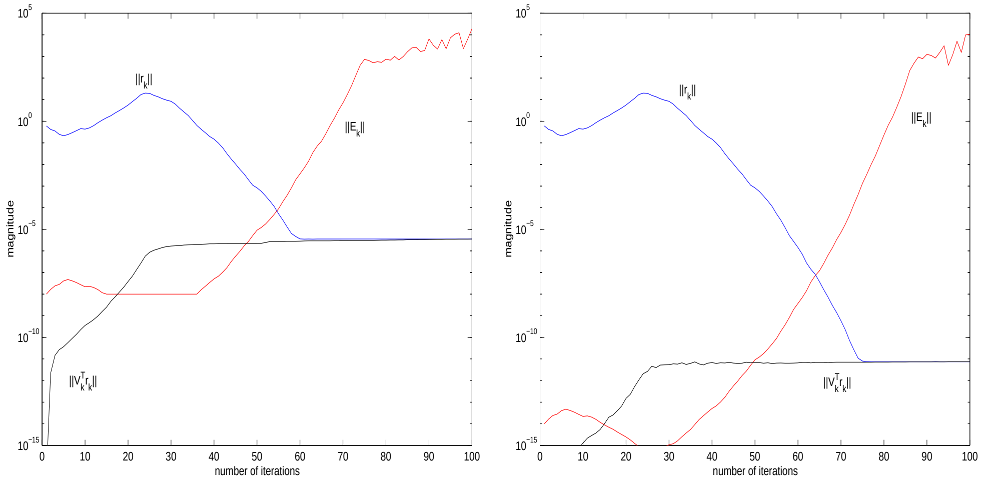

FIG. 5.1. *Example 5.5. Inexact FOM convergence history and other norms. Left: Tolerance with $\ell_{m_\star} = 1$. Right: Tolerance with $\ell_{m_\star} = \sigma_n(A)/m_\star$.*

deficient. In such a case, the proposed estimate $\sigma_n(A)$ becomes too loose, but most importantly the theoretical inexactness tolerance, i.e., the right-hand side in condition (5.10), becomes extremely small. A possible, though very stringent, strategy to limit the decrease of $\sigma_{m_\star}(H_{m_\star})$ is to impose a constraint on the size of $\|E_k\|$. Using (5.11), this is achieved by requiring that in addition to (5.10), E_k satisfy $\|E_k\| < \sigma_n(A)/m_\star$ for $k = 1, \ldots, m_\star$. Alternatively, for small space dimensions, the magnitude of $\sigma_m(H_m)$ could be monitored as m grows, yielding a nonmonotonic increase of $\|E_m\|$. Although the implementation consequences of the latter approach have not been fully explored, we believe it is a more feasible strategy than the very stringent one just discussed.

We point out that since the bounds of the form (5.10) are used to control the matrix-vector product, computationally one actually obtains relations of the form

$$||E_k|| \; ||\tilde{r}_{k-1}|| \approx \ell_m \, \varepsilon,$$

where the right-hand side is in fact a constant. This inverse relation can be clearly observed in the computational experiments, e.g., in Figure 5.1.

The following is an example where the use of $\ell_{m_\star} = 1$ is not satisfactory, while the use of $\ell_{m_\star} = \sigma_n(A)/m_\star$ provides an appropriate computable criterion to achieve the desired convergence.

Example 5.5. We consider the diagonal matrix $A = \text{diag}([10^{-4}, 2, 3, \ldots, 100])$. The perturbation matrix E_k is built up at each iteration $k > 0$ as a symmetrized random 100×100 matrix (with normally distributed values, Matlab function `randn`), while the right-hand side is $b = \texttt{randn}(100, 1)$, normalized so that $\|b\| = 1$. We considered $\varepsilon = 10^{-8}$ and $m_\star = 100$ iterations. In Figure 5.1 (left) we report the convergence history of FOM when using the condition (5.10) with $\ell_{m_\star} = 1$, as suggested in [3]. The right plot of Figure 5.1 displays the convergence history when $\ell_{m_\star} = \sigma_n(A)/m_\star$, which is 10^{-6} in this example, as an approximation to the condition in Theorem 5.4. The left plot shows that when using $\ell_{m_\star} = 1$, the final residual norm remains above the required tolerance ε. This is in general undesirable, since ε represents a user estimate of the required accuracy. The stricter condition with $\ell_{m_\star} = \sigma_n(A)/m_\star$ forces the quasi-orthogonality condition and thus the residual norm to be well below the expected tolerance.

The next example illustrates the fact that the use of inexact matrix-vector products may degrade the rate of convergence of the Krylov method. It also illustrates the effect of different values of $\ell_{m_\star}$.

Example 5.6. We consider the Grcar matrix of size $n = 100$, i.e., A is a Toeplitz matrix with minus ones on the subdiagonal, ones on the diagonal, and five superdiagonals of ones [18]. This matrix is known to have sensitive eigenvalues. Note that $\|A\| \approx 4.9985$ and $\sigma_{min}(A) \approx 0.7898$. The right-hand side is $b = e_1$. We compare solving with the exact GMRES method and with the inexact scheme with $\ell_{m_\star} = 1$ and with $\ell_{m_\star} = 10^{-2}$. The perturbation is as in Example 5.5 without symmetrization. In Figure 5.2 the convergence history (norm of true residual) is plotted for all methods. The results show that the rate of convergence in the inexact case degrades, yielding worse case performance when using a looser condition on $\|E_k\|$. Even more dramatic differences in the convergence rate can be observed; see, e.g., [36, Fig. 7.1].

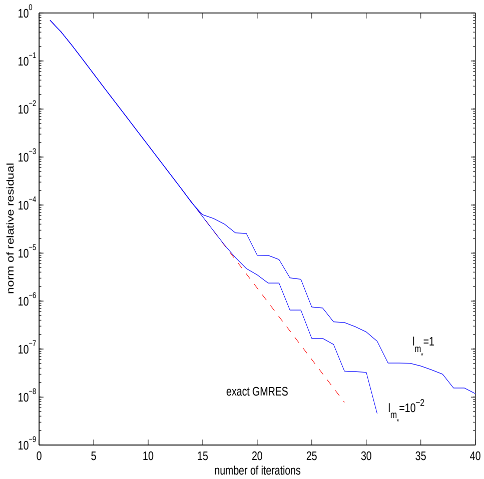

FIG. 5.2. *Example* 5.6. *Convergence of exact GMRES and inexact GMRES on the* 100×100 *Grcar matrix.*

6. Comments on the case of V_m nonorthogonal.

In this section, we relate our results to the cases when the columns of V_m in (2.2) are not orthogonal. This is the case, e.g., in inexact or flexible CG methods [14], [22]. In exact CG, the columns of V_m are (implicitly) generated by short recurrences. In inexact CG, short recurrences are used, but the global orthogonality is lost.

The assumption $V_m^T V_m = I$ is used in several places to obtain our results. For example, we use that $\|\tilde{r}_m\| = \|r_0 - V_{m+1} H_m y\| = \|e_1 - H_m y\|$. If this assumption were not to hold, then we would need to carry quantities such as $\|V_m\|$ in our bounds, and the bounds for $|\eta_k|$ in Lemmas 5.1 and 5.2 would become

$$|\eta_k| \leq \|R_m^{-1}\| \, \|\tilde{r}_{k-1}\| / \sigma_m(V_m).$$

Thus, for example, for inexact CG (or truncated FOM), to satisfy the requirement $\delta_m \leq \varepsilon$ we would have to replace (5.9) with

$$\|E_k\| \leq \frac{\sigma_m(V_m)\sigma_m(\widehat{H}_m)}{m} \frac{\varepsilon}{\|\tilde{r}_{k-1}\|}, \quad k \leq m.$$

If V_m is far from orthogonal, this condition is extremely stringent. It indicates that the loss of orthogonality implies a restriction to smaller perturbations, i.e., to more exact matrix-vector products. This observation is consistent with the comments in [22] (see also [36]) where truncation is used, i.e., by keeping only a certain number of vectors in V_m orthogonal. In summary, one has to pay somehow for the loss of orthogonality, either by allowing very small perturbations $\|E_k\|$ or by allowing for more reorthogonalization.

We note here that Sleijpen and van den Eshof [36] do study the case of nonorthogonal V_m and allow larger perturbations but only obtain bounds for what they term the "residual gap" $\delta_m = \|r_m - \tilde{r}_m\|$; i.e., this difference might be small, but in actual computations the residuals are not always small. In fact, in [36, section 6] it is assumed that the computed residual goes to zero. We are able to also bound the norm of the projected residuals $V_m^T r_m^{fom}$ and $W_m^T r_m^{gm}$ by assuming the extra orthogonality of V_m.

The approach in [36] is different from ours in other respects as well. In [36], the analysis starts with the relaxation strategy (5.10) from [3], i.e., $\ell_m = 1$, and then a bound for δ_m is obtained (involving $\kappa(A)$, cf. our discussion in section 5). Instead, we prescribe the desired level of δ_m (and the norm of the projected residual) and develop the necessary value of ℓ_m in (5.10) to reach that level.

7. Dynamic inner stopping criteria for inner-outer iterations.

This section serves as a brief introduction to the rest of the paper, where we describe a few application problems, each dealing with an inexact computation of the matrix-vector multiplication; see also, e.g., [36], [41], for applications different from those presented here. In each of the following four sections, we illustrate the use of the theory developed in the preceding sections to a different scientific computing application.

In general, for different problems, one would need a different monitoring strategy for $\|E_k\|$. However, for the applications we consider, all require the approximate solution of some linear system to perform the operation

$$\mathcal{A}v = Av + p(v). \tag{7.1}$$

If p_k is the residual in (7.1) corresponding to E_k, all monitoring strategies we use for our applications are of the form

$$\|p_k\| \le \ell \frac{1}{\|\tilde{r}_{k-1}\|}\varepsilon,$$

where ℓ depends on m (or $m_\star$, the maximum number of iterations) and on the exact operator associated with $\mathcal{A}$; cf. (5.10). In all cases studied, such a bound will imply that $\delta_m = \|r_m - \tilde{r}_m\| \le \varepsilon$ for both FOM and GMRES and that $\|V_m^T r_m\| \le \varepsilon$ in the case of FOM, or that $\|(V_m H_m)^T r_m\| \le \varepsilon$ for GMRES (for which a different value of ℓ is required).

We emphasize that in several cases, sharper stopping criteria can be devised by exploiting the structure of the problem; see, e.g., Example 9.3. Most plots report the magnitude of the variable inner stopping tolerance $\varepsilon_{\text{inner}} \equiv \ell \frac{1}{\|\tilde{r}_{k-1}\|}\varepsilon$, the convergence curve of the computed residual norm $\|\tilde{r}_m\|$ and of the true residual norm $\|r_m\|$, and finally, the magnitude of $\delta_m = \|r_m - \tilde{r}_m\|$; note that for FOM, $\|V_m^T r_m\| \le \delta_m$.

8. Schur complement systems.

Block structured linear systems can be solved by using the associated Schur complement; see, e.g., [8], [37]. For instance, if the system stems from a saddle point problem, then the algebraic equation has the following form:

$$\begin{bmatrix} S & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} w \\ x \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}, \tag{8.1}$$

with S symmetric. The corresponding (symmetric) Schur complement system is

$$B^T S^{-1} B x = B^T S^{-1} f, \tag{8.2}$$

whose solution x can be used to recover the unknown w. Let us set $A = B^T S^{-1} B$ and $b = B^T S^{-1} f$. Therefore, if a Krylov subspace method is used to solve $Ax = b$, operations with the coefficient matrix A are inexact unless S^{-1} is applied exactly. Thus, if systems with S are only solved approximately, or iteratively, at the kth iteration of the Krylov subspace method, the matrix-vector multiplication Av is replaced by $\mathcal{A}v = B^T z_j^{(k)}$, where $z_j^{(k)}$ is the approximation obtained at the jth (inner) iteration of the solution to the equation $Sz = Bv$. Let $p_j^{(k)} = S z_j^{(k)} - Bv$ be the associated residual, so that $z_j^{(k)} = z + S^{-1} p_j^{(k)}$. The question we address in this context is when to stop the inner iteration, i.e., how small should $\|p_j^{(k)}\|$ be? We have that

$$\mathcal{A}v = B^T z_j^{(k)} = B^T z + B^T S^{-1} p_j^{(k)} = \left(B^T S^{-1} B + B^T S^{-1} p_j^{(k)} \frac{v^T}{\|v\|^2} \right) v \equiv (A + E_k) v;$$

hence, assuming for simplicity[1] that $\|v\| = 1$, we have

$$\|E_k\| \le \|B^T S^{-1}\| \, \|p_j^{(k)}\|. \tag{8.3}$$

The bound above, together with the condition in Theorem 5.3 or in Theorem 5.4, provides a stopping criterion for the inner iteration involving S, which ensures that the orthogonality condition is satisfied with the desired tolerance. We state the result for the FOM method. An analogous result holds for GMRES.

PROPOSITION 8.1. *With the notation of Theorem* 5.4, *let* $\varepsilon > 0$. *Assume that at each outer iteration* $k \le m$ *of the FOM method, the inner residual* $p_j^{(k)}$ *after* j *inner iterations satisfies*

$$\|p_j^{(k)}\| \le \frac{\sigma_m(\widehat{H}_m)}{\|B^T S^{-1}\| m} \frac{1}{\|\tilde{r}_{k-1}^{fom}\|} \varepsilon \equiv \varepsilon_{\text{inner}}.$$

Then $\|r_m^{fom} - \tilde{r}_m^{fom}\| \le \varepsilon$ *and* $\|V_m^T r_m^{fom}\| \le \varepsilon$.

We see that the bound involves S^{-1}, and therefore the estimates as well as the stopping criterion threshold depend on the conditioning of the coefficient matrix S. We also recall that $\widehat{H}_m$ is no longer symmetric but upper Hessenberg, for significant nonzero entries appear in the upper triangular portion of the matrix to maintain orthogonality of V_m. An alternative strategy explored, for instance, in [14], [22] consists of maintaining *local* orthogonality among the basis vectors, so as to work with a banded $\widehat{H}_m$. See section 6 for comments on this case.

Example 8.2. We consider a problem of the form (8.1) derived from a two-dimensional saddle point magnetostatic problem described in [26, section 3]; see also

[1] This is always the case if the Arnoldi algorithm is explicitly used to construct the Krylov subspace. If $\|v\| \ne 1$, then the bound in (8.3) requires the additional factor $\|v\|^{-1}$.

[25] for additional experiments. In the data we employ, S has size 1272×1272 while B has dimension 1272×816. The Schur complement system is (8.2), with $\|B^T S^{-1}\| \approx 3.5792 \cdot 10^3$ and $\sigma_{min}(B^T S^{-1} B) \approx 3.2980 \cdot 10^2$. Considering $m_\star = 120$, we approximate the factor in the estimate for the stopping tolerance as

$$\frac{\sigma_{m_\star}(\widehat{H}_{m_\star})}{\|B^T S^{-1}\| m_\star} \approx \frac{\sigma_{min}(B^T S^{-1} B)}{\|B^T S^{-1}\| m_\star} \approx 7.6786 \cdot 10^{-4}.$$

A posteriori, we computed $\sigma_{m_\star}(\widehat{H}_{m_\star}) \approx 3.2973 \cdot 10^2$, which is clearly well approximated by $\sigma_{min}(B^T S^{-1} B)$.

When the saddle point problem (8.1) originates from certain classes of PDEs, it can be shown that $\sigma_{min}(B^T S^{-1} B) = \mathcal{O}(h^\alpha)$ for some known integer α, where h is a mesh-dependent parameter. Therefore, rough estimates of such quantity can be easily obtained; see, e.g., [20], [25], [31] and references therein.

Figure 8.1 displays the convergence history of the computed and true residuals of inexact FOM for $\varepsilon = 10^{-4}$. In the plot, $\tilde{r}_m$ stands for the computed inexact FOM residual $\tilde{r}_m^{fom}$. Note that $\|r_m\|$ remains larger than $\delta_m = \|r_m - \tilde{r}_m\|$, while the latter remains well below the requested tolerance $\varepsilon = 10^{-4}$. In the plot,

$$\varepsilon_{\text{inner}} = \frac{\sigma_{min}(B^T S^{-1} B)}{\|B^T S^{-1}\| m_\star} \frac{1}{\|\tilde{r}_{k-1}^{fom}\|} \varepsilon.$$

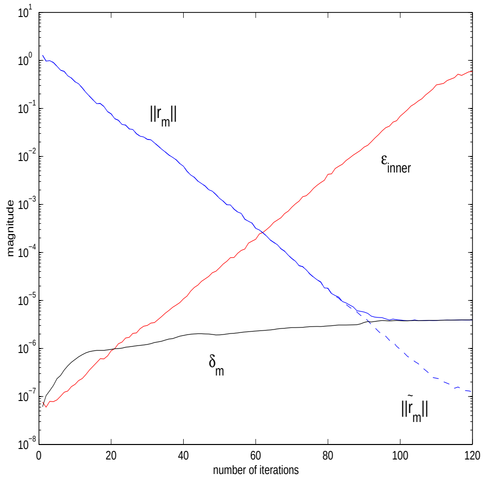

FIG. 8.1. *Example* 8.2. *Inexact FOM convergence history and other norms.*

When running the same example with $\varepsilon_{\text{inner}} = \varepsilon / \|\tilde{r}_{k-1}\|$, that is, $\ell_{m_\star} = 1$, we did not notice any change in the convergence behavior (the curves are the same as those in Figure 8.1), showing that our bound is very conservative on this problem.

9. Inexact preconditioning and flexible methods. In order to achieve better efficiency of the linear system solver, the preferred Krylov subspace method is commonly employed on the preconditioned system. Given a nonsingular matrix $\mathcal{P}$ or its inverse $\mathcal{P}^{-1}$, if right preconditioning is employed, the original system is transformed into

$$A\mathcal{P}^{-1}\bar{x} = b, \qquad x = \mathcal{P}^{-1}\bar{x}.$$

At each iteration of the solver the application of $\mathcal{P}^{-1}$ is thus required. In several cases, e.g., as in [10], neither $\mathcal{P}$ nor $\mathcal{P}^{-1}$ can be applied exactly but only through an operator, yielding a variable preconditioning procedure in which a different preconditioning operator is applied at each iteration. Often this is due to the approximation of an optimal, in some sense, but expensive preconditioner. More precisely, at each iteration the operation

$$A\tilde{z}_k, \qquad \tilde{z}_k \approx \mathcal{P}^{-1}v_k, \tag{9.1}$$

is carried out, where $\tilde{z}_k$ can be thought of as some approximation to the solution z_k of the linear system $\mathcal{P}z = v_k$, $k = 1, \dots, m$ [28], [34], [39]. The operation in (9.1) thus replaces the exact preconditioning product $A\mathcal{P}^{-1}v_k$. In the variable preconditioning case, the Arnoldi relation $A\mathcal{P}^{-1}V_m = V_{m+1}H_m$ is transformed into

$$A[\tilde{z}_1, \dots, \tilde{z}_m] = V_{m+1}H_m. \tag{9.2}$$

As a consequence of the relation (9.2), a Krylov subspace is no longer constructed, and standard convergence results do not apply; see [9] for a study of convergence which applies to this case. Let $\bar{x}_m = V_m y_m$ be an approximate computed solution, and let $y_m = [\eta_1, \dots, \eta_m]^T$, $Z_m = [z_1, \dots, z_m]$, $\tilde{Z}_m = [\tilde{z}_1, \dots, \tilde{z}_m]$. Observe that we can write $\tilde{Z}_m = Z_m + \mathcal{P}^{-1}P_m$, where P_m collects all residuals in the approximate application of $\mathcal{P}$. From (9.2) we clearly have $A\mathcal{P}^{-1}V_m + A\mathcal{P}^{-1}P_m = V_{m+1}H_m$. Then the true residual satisfies

$$r_m = r_0 - A\mathcal{P}^{-1}\bar{x}_m = (r_0 - V_{m+1}H_m y_m) + A\mathcal{P}^{-1}P_m y_m. \tag{9.3}$$

Note that the term in parentheses on the right-hand side is the computed residual. The relation (9.3) is completely analogous to (4.1). Therefore, setting $\tilde{r}_m := b - V_{m+1}H_m y_m$ and using the same argument as in Proposition 4.1, we obtain

$$\|r_m - \tilde{r}_m\| \le \sum_{k=1}^{m} |\eta_k| \, \|A\mathcal{P}^{-1}p_k\| \le \|A\mathcal{P}^{-1}\| \sum_{k=1}^{m} |\eta_k| \, \|p_k\| \tag{9.4}$$

and also

$$\|V_m^T r_m^{fom}\| \le \|A\mathcal{P}^{-1}\| \sum_{k=1}^{m} |\eta_k| \, \|p_k\|, \qquad \|(V_m H_m)^T r_m^{gm}\| \le \|H_m\| \|A\mathcal{P}^{-1}\| \sum_{k=1}^{m} |\eta_k| \, \|p_k\|.$$

Using these relations one obtains results analogous to Theorems 5.3 and 5.4.

PROPOSITION 9.1. *Let $\varepsilon > 0$. Assume that m iterations of the inexact GMRES method have been carried out. Let $\bar{x}_m^{gm} = V_m y_m^{gm}$, the true residual $r_m^{gm} = b - A\mathcal{P}^{-1}\bar{x}_m^{gm}$, and the computed residual $\tilde{r}_m^{gm} = r_0 - V_{m+1}H_m y_m^{gm}$. If at each iteration $k \le m$ the inner residual $p_k = v_k - \mathcal{P}\tilde{z}_k$ satisfies*

$$\|p_k\| \le \frac{\sigma_m(H_m)}{\|A\mathcal{P}^{-1}\| m} \frac{1}{\|\tilde{r}_{k-1}^{gm}\|} \varepsilon,$$

then $\|r_m^{gm} - \tilde{r}_m^{gm}\| \le \varepsilon$. Moreover, if

$$\|p_k\| \le \frac{1}{\|A\mathcal{P}^{-1}\| m \kappa(H_m)} \frac{1}{\|\tilde{r}_{k-1}^{gm}\|} \varepsilon,$$

then $\|(V_{m+1}H_m)^T r_m^{gm}\| \leq \varepsilon$.

PROPOSITION 9.2. *Let* $\varepsilon > 0$. *Assume that* m *iterations of the inexact FOM method have been carried out. Let* $\bar{x}_m^{fom} = V_m y_m^{fom}$, *the true residual* $r_m^{fom} = b - A\mathcal{P}^{-1}\bar{x}_m^{fom}$, *and the computed residual* $\tilde{r}_m^{fom} = b - V_{m+1}H_m y_m^{fom}$. *If at each iteration* $k \leq m$ *the inner residual* $p_k = v_k - \mathcal{P}\tilde{z}_k$ *satisfies*

$$\|p_k\| \leq \frac{\sigma_m(\widehat{H}_m)}{\|A\mathcal{P}^{-1}\|m} \frac{1}{\|\tilde{r}_{k-1}^{fom}\|}\varepsilon,$$

then $\|r_m^{fom} - \tilde{r}_m^{fom}\| \leq \varepsilon$ *and* $\|V_m^T r_m^{fom}\| \leq \varepsilon$.

Example 9.3. We work on the same saddle point problem as in Example 8.2, and we consider a now well-established technique which consists of preconditioning the original structured problem (8.1) with

$$\mathcal{P} = \begin{bmatrix} D & 0 \\ 0 & B^T B \end{bmatrix}, \tag{9.5}$$

where D is an approximation to S; see, e.g., [31] and the references given therein. Here we use $D = I$ as in [25]. At each iteration, the application of the exact preconditioner $\mathcal{P}$ requires solving a system with coefficient matrix $B^T B$, with sparse B, which is efficiently carried out by sparse direct methods on small problems. On large and denser problems (e.g., in three-dimensional applications), direct solution of systems with $B^T B$ is very time consuming, and approximations to $\mathcal{P}$ need to be considered [25], [26]. Here we consider solving with $B^T B$ by an iterative method, relaxing the accuracy of the solution as the outer process converges.

It is important to notice that for this structured problem, the conditions in Proposition 9.1 can be sharpened. We work with the first condition, since we want our computation to guarantee that $\delta_m = \|r_m - \tilde{r}_m\| \leq \varepsilon$. Let $p_k^T = [0, \hat{p}_k^T]$ be the residual[2] of the system with $\mathcal{P}$ at the kth iteration. Using the first bound in (9.4), we see that

$$A\mathcal{P}^{-1}p_k = A\mathcal{P}^{-1}\begin{bmatrix} 0 \\ \hat{p}_k \end{bmatrix} = \begin{bmatrix} B(B^T B)^{-1}\hat{p}_k \\ 0 \end{bmatrix},$$

therefore, $\|A\mathcal{P}^{-1}p_k\| \leq \|\hat{p}_k\|/\sigma_{n_B}(B)$, where $\sigma_{n_B}(B)$ is the smallest singular value of B. Hence, the condition in Proposition 9.1 can be relaxed to be

$$\|p_k\| \leq \frac{\sigma_{m_\star}(H_{m_\star})\sigma_{n_B}(B)}{m_\star} \frac{1}{\|\tilde{r}_{k-1}\|}\varepsilon. \tag{9.6}$$

In Figure 9.1 (right) we show the convergence history of GMRES applied to (8.1) with preconditioner (9.5) using condition (9.6) for the inner residual. The quantity δ_m is computed as $\delta_m = \|A\mathcal{P}^{-1}P_m y_m^{gm}\|$; cf. (9.3). A preconditioned conjugate gradient method is used as an inner iterative solver, where at least two inner iterations with $B^T B$ are performed. We consider $\varepsilon = 10^{-9}$, and we exploit the approximation $\sigma_{m_\star}(H_{m_\star}) \approx \sigma_n(A\mathcal{P}^{-1})$, which we use to explicitly estimate the approximate factor

$$\sigma_{m_\star}(H_{m_\star})\sigma_{n_B}(B) \approx \sigma_n(A\mathcal{P}^{-1})\sigma_{n_B}(B) \approx 1.$$

The same argument given in Example 8.2 applies here for estimating $\sigma_n(A\mathcal{P}^{-1})$ and $\sigma_{n_B}(B)$ when a PDE problem is being solved.

For a comparison with the criterion in [3], the left plot in Figure 9.1 considers the stopping criterion with $\ell_{m_\star} = 1$. One can appreciate the difference between the two criteria and in particular the difference of the final attainable accuracy.

[2] The first block of the residual vector is zero because the (1,1) block of $\mathcal{P}^{-1}$ is applied exactly.

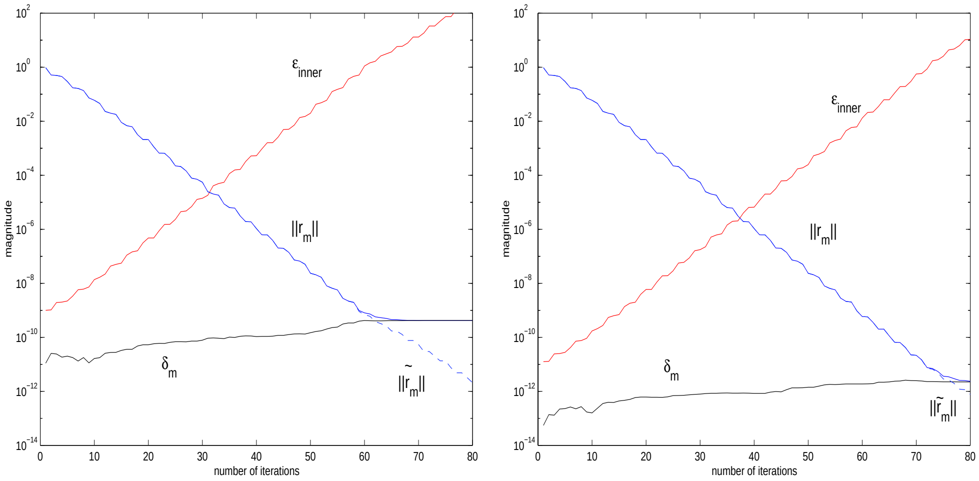

Fig. 9.1. *Example* 9.3. *Inexact GMRES with a minimum of two inner iterations, convergence history and other norms. Left: Tolerance with $\ell_{m_\star} = 1$. Right: Tolerance as in* (9.6).

9.1. Comments on flexible Krylov subspace methods. In this section we make some observations on flexible and inexact Krylov subspace methods.

As already illustrated earlier, variable preconditioning for a Krylov subspace method can be interpreted as an inexact subspace method, and this is the approach, e.g., in [14]. Flexible Krylov subspace methods such as FGMRES [28] or FQMR [39] are defined not only by the fact that the preconditioner changes from one iteration to the next, but also from the fact that the solution is obtained directly from the new preconditioned basis $\{\tilde{z}_1, \ldots, \tilde{z}_m\}$ that generates the approximation space $\mathcal{R}(\tilde{Z}_m)$; see (9.2). In other words, in flexible Krylov subspace methods the dependence on the approximation basis V_m is eliminated. In the present context, this allows us to directly form the approximate solution, bypassing the unpreconditioned space, which caused the inexactness of the true residual $r_m = b - A\mathcal{P}^{-1}\bar{x}_m$. Indeed, r_m is available, since x_m is computed as

$$x_m = [\tilde{z}_1, \ldots, \tilde{z}_m]y_m = \tilde{Z}_m y_m \tag{9.7}$$

for some y_m; see [28], [39].

Let $\tilde{z}_k = z_k + \mathcal{P}^{-1}p_k$, $k \leq m$, where p_k is the (inner) residual of the system with $\mathcal{P}$. The perturbation in the matrix-vector product with A can be written explicitly as

$$\mathcal{A}v_k = A\tilde{z}_k = A\mathcal{P}^{-1}v_k - A\mathcal{P}^{-1}p_k \frac{v_k^T}{\|v_k\|^2}v_k \equiv (A\mathcal{P}^{-1} + E_k)v_k,$$

implying that the error computed at each matrix-vector multiplication has the bound

$$\|E_k\| \leq \|A\mathcal{P}^{-1}\| \frac{\|p_k\|}{\|v_k\|}. \tag{9.8}$$

In the case of FOM (and CG), the expression (9.7) indicates that the perturbation E_k does not affect the computation of the true residual, while full orthogonality is preserved with respect to the generated space. In other words, in exact arithmetic,

$r_m = \tilde{r}_m$ for $m > 0$ and r_m is orthogonal to $\mathcal{R}(V_m)$ to full precision, where V_m satisfies (9.2). As a consequence, the inexactness of the matrix-vector multiplication only affects the goodness of the approximation space $\mathcal{R}(\tilde{Z}_m)$, and thus the accuracy of the obtained approximate solution is $x_m = \tilde{Z}_m y_m$. From this discussion we deduce that in the flexible context, a dynamic stopping criterion can be employed to save computational effort by limiting the work of the preconditioner (so that the bound on the inner residual norm $\|p_k\|$ is not too stringent) without sacrificing the closeness of the computed and true residuals.

A particular case of the flexible Krylov subspace methods is when the preconditioner $\mathcal{P}$ is the inexact application of matrix A itself, e.g., when an inner Krylov subspace iteration is used to approximate the solution of $A\tilde{z}_k = v_k$ in (9.1) (possibly with a much looser tolerance) [28], [34], [39]. In such a setting, if $p_j^{(k)}$ is the inner residual after j iterations of the inner Krylov subspace method, we have

$$\mathcal{A}v_k = \left(I + p_j^{(k)} \frac{v_k^T}{\|v_k\|^2} \right) v_k,$$

which obviously shows that the distance from the perfectly preconditioned matrix (the identity matrix) is directly monitored through the inner residual, i.e., (9.8) becomes $\|E_k\| \le \|p_j^{(k)}\|/\|v_k\|$.

10. Linear systems with a parameter.

Consider the linear system

$$(M + \omega K)x = b, \tag{10.1}$$

which needs to be solved for several values of the parameter ω; see, e.g., [21], [33]. A similar problem arises in the context of eigenvalue computation: when using certain formulations of the inexact shift-and-invert Arnoldi method, a system of the form

$$(M + \omega K)K^{-1}x = b \tag{10.2}$$

needs to be solved at each iteration of the method [1].

If K is nonsingular, the system (10.1) can be transformed into

$$(A + \omega I)\tilde{x} = b, \qquad A = MK^{-1}, \quad \tilde{x} = Kx, \tag{10.3}$$

yielding a shifted system with coefficient matrix $A + \omega I$ that can be solved efficiently with a Krylov subspace method for several values of ω simultaneously; see [33] and references therein. Each iteration involves solving a system with coefficient matrix K, and if the problem has large dimension, such a solution is carried out iteratively. The fact that K^{-1} is only applied approximately destroys the equivalence of the two formulations (10.1) and (10.3), even in exact arithmetic. In other words, the coefficient matrix in (10.2) is perturbed, and thus the analysis of the previous sections applies. While the system which is approximately solved is (10.3), the equation of interest is in fact (10.1). It was observed in [33] that the inner tolerance greatly influences the performance of the overall method, and as a consequence the accuracy in the solution of (10.1) may deteriorate a great deal. It is the aim of this section to provide a better understanding of this phenomenon.

Assume that the operation $\mathcal{A}v_k$ replaces Av_k, and let $\tilde{z}_k$ be the approximation to $z_k = K^{-1}v_k$. As in the previous section, we can write $\tilde{z}_k = z_k + K^{-1}p_k$, using which we obtain

$$\mathcal{A}v_k = M\tilde{z}_k = MK^{-1}v_k + MK^{-1}p_k \frac{v_k^T}{\|v_k\|^2} v_k \equiv (A + E_k)v_k,$$

implying that the error computed at each matrix-vector multiplication has the bound $\|E_k\| \le \|MK^{-1}\| \, \|p_k\|/\|v_k\|$. We can be more precise on the effect on the whole procedure. If systems with K were solved exactly, then the Arnoldi relation $MK^{-1}V_m = V_{m+1}H_m$ would hold, which, after shifting, would yield

$$(MK^{-1} + \omega I)V_m = V_{m+1}(H_m + \omega \tilde{I}_m) \qquad \tilde{I}_m = [I_m, 0]^T.$$

Let $\tilde{Z}_m = [\tilde{z}_1, \ldots, \tilde{z}_m]$ and $Z_m = [z_1, \ldots, z_m]$. In the inexact case, the shifted inexact Arnoldi relation is

$$M\tilde{Z}_m + \omega V_m = V_{m+1}(H_m + \omega \tilde{I}_m). \tag{10.4}$$

Let $x_m = \tilde{Z}_m y_m$ be the approximate solution to x with $y_m \in \mathbb{R}^m$, $x_0 = 0$, and let $P_m = [p_1, \ldots, p_m]$. Note that x_m is computed by means of the flexible strategy described in the previous section, that is, $x_m \in \mathcal{R}(\tilde{Z}_m)$. From $\tilde{Z}_m = Z_m + K^{-1}P_m$ we see that $K\tilde{Z}_m = V_m + P_m$. Therefore, using (10.4) the exact (and computable) residual can be written as

$$\begin{aligned} r_m &= b - (M + \omega K)x_m = b - (M + \omega K)\tilde{Z}_m y_m \\ &= b - M\tilde{Z}_m y_m - \omega K\tilde{Z}_m y_m = b - V_{m+1}(H_m + \omega \tilde{I}_m)y_m - \omega P_m y_m \\ &= V_{m+1}\left(e_1\beta - (H_m + \omega \tilde{I}_m)y_m\right) - \omega P_m y_m. \end{aligned}$$

The relation above is completely analogous to that in (4.1), and

$$\delta_m := \|\omega P_m y_m\| \tag{10.5}$$

measures the distance between the true residual and the residual of the underlying method. Moreover, if $y_m = [\eta_1, \ldots, \eta_m]^T$ is determined as $y_m^{fom} = (\widehat{H}_m + \omega I_m)^{-1}e_1\beta$, then the associated residual $r_m^{fom} = b - (M + \omega K)\tilde{Z}_m y_m^{fom}$ satisfies

$$\|V_m^T r_m^{fom}\| \le |\omega| \sum_{k=1}^{m} |\eta_k| \, \|p_k\|. \tag{10.6}$$

An analogous relation holds for $\|(V_{m+1}(H_m + \omega \tilde{I}_m))^T r_m^{gm}\|$.

We can thus derive a dynamic stopping criterion for the inner iteration that solves a system with K, so that, for instance, the bound (10.6) can be used to obtain $\|V^T r_m^{fom}\| < \varepsilon$. The proof follows the same argument as that of Theorem 5.4.

PROPOSITION 10.1. *Let $\varepsilon > 0$. Let the residual of the inexact FOM applied to* (10.3) *be $r_m^{fom} = b - (M + \omega K)\tilde{Z}_m y_m^{fom}$ and the computed residual be $\tilde{r}_k^{fom} = V_{k+1}(e_1\beta - (H_k + \omega \tilde{I}_k)y_k^{fom})$ for $k \le m$. If at each iteration $k \le m$ the inner residual $p_k = v_k - K\tilde{z}_k$ satisfies*

$$\|p_k\| \le \frac{\sigma_m(\widehat{H}_m + \omega I_m)}{|\omega| m} \frac{1}{\|\tilde{r}_{k-1}^{fom}\|}\varepsilon \equiv \varepsilon_{\text{inner}}, \tag{10.7}$$

then $\|r_m^{fom} - \tilde{r}_m^{fom}\| \le \varepsilon$ and $\|V_m^T r_m^{fom}\| \le \varepsilon$.

A similar relation holds for the GMRES solution, using Lemma 5.1.

Example 10.2. We consider the system

$$(M - \lambda^2 K)x = b$$

arising in the direct frequency analysis of an n-DOF discretized linear system in the absence of viscous damping, studied in [33]. The parameter λ represents the inverse of the frequency, M is the kinetic matrix, and $K = K_0 + \imath D_H$ is given by the potential energy matrix K_0 and the hysteretic damping matrix D_H. For more details on the problem we refer to [33] and references therein. The problem (test case **B** in [33]) has size $n = 3627$. We consider for the illustrative experiment here only one value of λ, namely $\lambda = (30 \cdot 2\pi)^{-1}$. The matrix K is complex symmetric and M is real with all positive diagonal entries. We ran inexact FOM with $m_\star = 30$ and $\varepsilon = 10^{-8}$. Inner systems with K were solved with the complex symmetric version of the preconditioned conjugate gradients. For a comparison with the criterion in [3], the left plot in Figure 10.1 considers the stopping criterion with $\ell = 1$, i.e.,

$$\|p_k\| \le \frac{1}{\|\tilde{r}_{k-1}\|}\varepsilon. \tag{10.8}$$

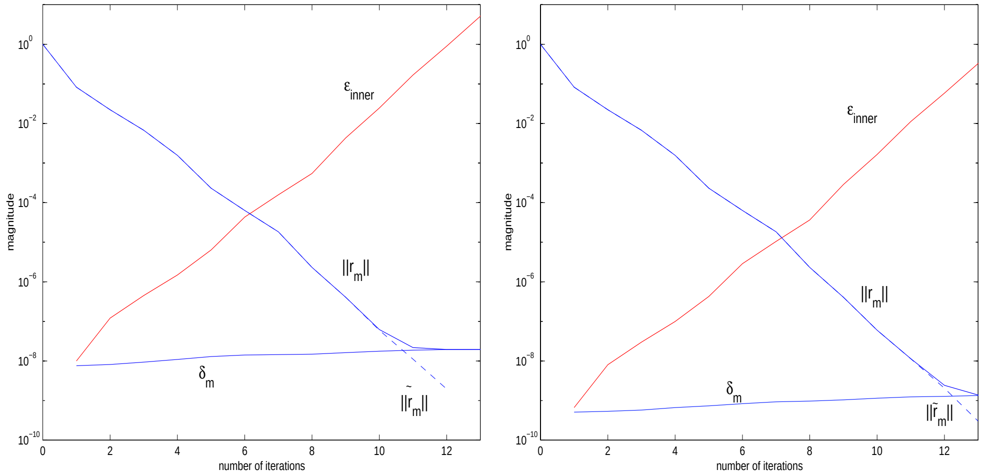

FIG. 10.1. *Example* 10.2. *Inexact FOM convergence history and other norms. Left: Inner stopping criterion with $\ell = 1$. Right: Inner stopping criterion with $\ell = 1/m_\star$.*

We observed that for this problem $\sigma_{m_\star}(\widehat{H}_{m_\star} - \lambda^2 I) \approx \lambda^2$, therefore we replaced the inner stopping criterion (10.7) with

$$\|p_k\| \le \frac{1}{m_\star}\frac{1}{\|\tilde{r}_{k-1}\|}\varepsilon. \tag{10.9}$$

In both tests we show δ_m computed as in (10.5). The left plot shows that when using test (10.8), the condition $\delta_m \le \varepsilon$ fails to be satisfied. Such a condition is instead satisfied when using the stricter inner stopping test (10.9).

11. Application to eigenvalue computations. Following the empirical results in [4], in this section we show that inexactness of the computed matrix-vector multiplication can be monitored also in the eigenvalue context. The results here though are less general than in the linear system setting. If the exact Arnoldi method is employed to approximate the eigenpairs of A (see, e.g., [1], [27]), then starting with a unit norm vector v_1, the Krylov subspace $\mathcal{K}_m(A, v_1)$ is constructed and $(\tilde{\lambda}_i, V_m y_i)$,

$i = 1, \ldots, m$, are approximate (Ritz) eigenpairs to some of the (exact) eigenpairs of A, where $\tilde{\lambda}_i, y_i$, $i = 1, \ldots, m$, satisfy $\widehat{H}_m y_i = \tilde{\lambda}_i y_i$, $i = 1, \ldots, m$.

In the inexact case, i.e., when the subspace (2.4) is used instead, the inexact Arnoldi relation (2.2) holds, and let $\tilde{\lambda}_i, y_i$, $i = 1, \ldots, m$, now be the eigenpairs of $\widehat{H}_m$. Then for each $i = 1, \ldots, m$, we can write

$$\begin{aligned} AV_m y_i - \tilde{\lambda}_i V_m y_i &= V_m \widehat{H}_m y_i + h_{m+1,m} v_{m+1} e_m^T y_i \\ &\qquad - [E_1 v_1, E_2 v_2, \ldots, E_m v_m] y_i - \tilde{\lambda}_i V_m y_i \\ &= h_{m+1,m} v_{m+1} e_m^T y_i - [E_1 v_1, E_2 v_2, \ldots, E_m v_m] y_i. \end{aligned}$$

Therefore, the norm of the difference between the true and computable eigenvalue residuals is

$$\|(AV_m y_i - \tilde{\lambda}_i V_m y_i) - h_{m+1,m} v_{m+1} e_m^T y_i\| = \|[E_1 v_1, E_2 v_2, \ldots, E_m v_m] y_i\|.$$

Moreover,

$$\begin{aligned} \|V_m^T (AV_m y_i - \tilde{\lambda}_i V_m y_i)\| &\le \|[E_1 v_1, E_2 v_2, \ldots, E_m v_m] y_i\| \\ &\le \sum_{k=1}^{m} |\eta_k^{(i)}| \, \|E_k\|, \end{aligned}$$

where $y_i^T = [\eta_1^{(i)}, \ldots, \eta_m^{(i)}]$. The bound above is completely analogous to those in Proposition 4.2 and Proposition 4.3. This allows us to accept large values of $\|E_k\|$ for k large enough such that $|\eta_k^{(i)}|$ is small. The components of the eigenvectors corresponding to the extreme eigenvalues do have a decreasing pattern, at least in the symmetric case [16], [24], and this allows us to effectively relax the accuracy in the computation of the matrix-vector products with A [4]. We note that this fact was already noticed and exploited in the symmetric case in [16].

We should keep in mind that the relation above holds for any $i = 1, \ldots, m$. In particular, different approximate eigenvectors $V_m y_i$, $i = 1, \ldots, m$, in general converge with a different convergence rate as m increases, implying that the magnitude of the last components of two eigenvectors y_i, y_j, $i \neq j$, will in general decrease with a possibly very different pattern as m increases. As a consequence, depending on the approximate eigenvector convergence rate, the relaxation strategy for $\|E_k\|$ may substantially differ. Therefore, in practical implementations, it would be possible to relax the accuracy in the computation of A only when approximating leading groups of eigenpairs for which the iterates show a similar convergence rate.

12. Conclusions.

We have analyzed a class of inexact Krylov subspace methods and answered several outstanding questions on their convergence properties. In particular, we showed why the norm of the matrix-product perturbation can grow as the iteration progresses. This is due to the combination of two factors. First, the inexact method approximates the initial residual r_0 in a certain subspace (see (3.1) or (3.2)), and this implies that the quantities η_k converge to zero as $k \to \infty$ if no breakdown occurs; see the discussion after Proposition 3.3. Second, the quasi orthogonality of the true residual r_m (e.g., of the form $\|V^T r_m\| \le \varepsilon$) is maintained if all products $|\eta_k| \, \|E_k\|$ are kept small ($k \le m$).

In fact, the theory presented here not only explains the empirical behavior noted by other researchers, but also provides the basis for practical dynamic strategies for the relaxation of the matrix-vector products. These strategies translate into stopping criteria for the inner solver in inner-outer procedures. Furthermore, these criteria may be efficiently sharpened for specific problems, as illustrated, e.g., in Example 9.3.

We also made some observations on flexible Krylov subspace methods, indicating that, unlike the inexact methods, the true residual is available. Nevertheless a dynamic stopping criterion can still be useful for computational purposes.

Acknowledgments. We thank the referees for their suggestions, which were adopted in the final manuscript. We also thank Miro Rozložník for helpful discussions.

REFERENCES

[1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

[2] W. Bietenholz, N. Eicker, A. Frommer, Th. Lippert, B. Medeke, K. Schilling, and G. Weuffen, *Preconditioning of improved and 'perfect' actions*, Comput. Phys. Comm., 119 (1999), pp. 1–18.

[3] A. Bouras and V. Frayssé, *A Relaxation Strategy for Inexact Matrix–Vector Products for Krylov Methods*, Technical Report TR/PA/00/15, CERFACS, Toulouse, France, 2000.

[4] A. Bouras and V. Frayssé, *A Relaxation Strategy for the Arnoldi Method in Eigenproblems*, Technical Report TR/PA/00/16, CERFACS, Toulouse, France, 2000.

[5] A. Bouras, V. Frayssé, and L. Giraud, *A Relaxation Strategy for Inner–Outer Linear Solvers in Domain Decomposition Methods*, Technical report TR/PA/00/17, CERFACS, Toulouse, France, 2000.

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

[7] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[8] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Oxford University Press, Oxford, 1989.

[9] M. Eiermann and O. Ernst, *Geometric aspects in the theory of Krylov subspace methods*, in Acta Numer. 10, Cambridge University Press, Cambridge, UK, 2001, pp. 251–312.

[10] H. C. Elman, O. G. Ernst, and D. P. O'Leary, *A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations*, SIAM J. Sci. Comput., 23 (2001), pp. 1291–1315.

[11] J. van den Eshof, A. Frommer, Th. Lippert, K. Schilling, and H. van der Vorst, *Numerical methods for the QCD overlap operator:* I. *Sign-function and error bounds*, Comput. Phys. Comm., 146 (2002), pp. 203–244.

[12] A. Frommer and D. B. Szyld, *H-splittings and two-stage iterative methods*, Numer. Math., 63 (1992), pp. 345–356.

[13] G. H. Golub and M. L. Overton, *The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems*, Numer. Math., 53 (1988), pp. 571–593.

[14] G. H. Golub and Q. Ye, *Inexact preconditioned conjugate gradient method with inner-outer iteration*, SIAM J. Sci. Comput., 21 (1999), pp. 1305–1320.

[15] G. H. Golub and Q. Ye, *Inexact inverse iterations for the generalized eigenvalue problems*, BIT, 40 (2000), pp. 671–684.

[16] G. H. Golub, Z. Zhang, and H. Zha, *Large sparse symmetric eigenvalue problems with homogeneous linear constraints: The Lanczos process with inner–outer iterations*, Linear Algebra Appl., 309 (2000), pp. 289–306.

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

[18] N. J. Higham, *The Test Matrix Toolbox for Matlab (Version* 3.0*)*, Technical Report 276, Manchester Centre for Computational Mathematics, University of Manchester, Manchester, UK, 1995.

[19] J. Mandel, *On block diagonal and Schur complement preconditioning*, Numer. Math., 58 (1990), pp. 79–93.

[20] J. Maryška, M. Rozložník, and M. Tůma, *Schur complement systems in the mixed-hybrid finite element approximation of the potential fluid flow problem*, SIAM J. Sci. Comput., 22 (2000), pp. 704–723.

[21] K. Meerbergen, *The solution of parametrized symmetric linear systems*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 1038–1059.

[22] Y. Notay, *Flexible conjugate gradients*, SIAM J. Sci. Comput., 22 (2000), pp. 1444–1460.

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

[24] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[25] I. Perugia and V. Simoncini, *Block–diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear Algebra Appl., 7 (2000), pp. 585–616.

[26] I. Perugia, V. Simoncini, and M. Arioli, *Linear algebra methods in a mixed approximation of magnetostatic problems*, SIAM J. Sci. Comput., 21 (1999), pp. 1085–1101.

[27] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halstead Press, New York, 1992.

[28] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.

[29] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, MA, 1996.

[30] A. H. Sherman, *On Newton-iterative methods for the solution of systems of nonlinear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 755–771.

[31] D. Silvester and A. Wathen, *Fast iterative solution of stabilized Stokes systems part* II: *Using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1367.

[32] V. Simoncini and L. Eldèn, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.

[33] V. Simoncini and F. Perotti, *On the numerical solution of* $(\lambda^2 A + \lambda B + C)x = b$ *and application to structural dynamics*, SIAM J. Sci. Comput., 23 (2002), pp. 1875–1897.

[34] V. Simoncini and D. B. Szyld, *Flexible inner-outer Krylov subspace methods*, SIAM J. Numer. Anal., 40 (2003), pp. 2219–2239.

[35] V. Simoncini and D. Szyld, *Flexible Inner-Outer Krylov Methods (and Inexact Krylov Methods)*, presentation at the Latsis Symposium on Iterative Solvers for Large Linear Systems, ETH, Zurich, Switzerland, 2002.

[36] G. L. G. Sleijpen and J. van den Eshof, *Inexact Krylov Subspace Methods for Linear Systems*, Preprint 1224, Department of Mathematics, Universiteit Utrecht, Utrecht, The Netherlands, 2002.

[37] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, New York, Melbourne, 1996.

[38] G. W. Stewart, *Backward error bounds for approximate Krylov subspaces*, Linear Algebra Appl., 340 (2002), pp. 81–86.

[39] D. B. Szyld and J. A. Vogel, *A flexible quasi-minimal residual method with inexact preconditioning*, SIAM J. Sci. Comput., 23 (2001), pp. 363–380.

[40] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.

[41] J. S. Warsa, M. Benzi, T. A. Wareing, and J. E. Morel, *Preconditioning a mixed discontinuous finite element method for radiation diffusion*, Numer. Linear Algebra Appl., to appear.