Noname manuscript No.
(will be inserted by the editor)

Backward error analysis of an inexact Arnoldi method using a certain Gram Schmidt variant

Ute Kandler · Christian Schröder

Received: date / Accepted: date

Abstract In numerous recent applications including tensor computations, compressed sensing and mixed precision arithmetics vector operations like summing, scaling, or matrix-vector multiplication are subject to inaccuracies whereas inner products are exact. We investigate the behavior of Arnoldi's method for Hermitian matrices under these circumstances. We introduce a special purpose variant of Gram Schmidt orthogonalization and prove bounds on the distance to orthogonality of the now-not-anymore orthogonal Krylov subspace basis. This Gram Schmidt variant additionally implicitly provides an exactly orthogonal basis. In the second part we perform a backward error analysis and show that this exactly orthogonal basis satisfies a Krylov relation for a perturbed system matrix – even in the Hermitian case. We prove bounds for the norm of the backward error which is shown to be on the level of the accuracy of the vector operations. Care is taken to avoid problems in case of near breakdowns. Finally, numerical experiments confirm the applicability of the method and of the proven bounds.

Keywords inexact matrix-vector operations · Gram Schmidt orthogonalization · loss of orthogonality · Arnoldi's method · Krylov relation · backward error bounds

Mathematics Subject Classification (2000) 65F15 · 65F25

1 Introduction

In the identification of ground states of quantum systems one has to solve eigenvalue problems of extremely large dimension ($n = 2^{100}$ is not uncommon) [20,34,35]. In particular, one is interested in the smallest eigenvalues (in physics terms, the ground state energies) and their distance to one another.

The matrix eigenvalue problem, i.e., obtaining eigenvalues, eigenvectors and/or invariant subspaces of a matrix $A \in \mathbb{C}^{n\times n}$, i.e., solving the equation

$$Ax = \lambda x$$

work supported by German research council, DFG under project "Scalable Numerical Methods for Adiabatic Quantum Preparation"

Ute Kandler
TU Berlin, Str. des 17. Juni 136, 10623 Berlin, GERMANY
Tel.: +49 (0)30 314 - 79177
Fax: +49 (0)30 314 - 79706
E-mail: kandler@math.tu-berlin.de

Christian Schröder
TU Berlin, Str. des 17. Juni 136, 10623 Berlin, GERMANY
Tel.: +49 (0)30 314 - 24767
Fax: +49 (0)30 314 - 79706
E-mail: schroed@math.tu-berlin.de

for $\lambda \in \mathbb{C}$, $x \in \mathbb{C}^n \setminus \{0\}$ is among the best studied problems in numerical linear algebra. For the case of large sparse matrices A, most prominent are iterative methods that search in a Krylov subspace

$$\mathcal{K}_k := \mathcal{K}_k(A, v_1) := \operatorname{span}(v_1, Av_1, A^2v_1, \ldots, A^{k-1}v_1)$$

for approximations of eigenvectors. An orthonormal basis $V_k = [v_1, v_2, \ldots, v_k] \in \mathbb{C}^{n\times k}$ of $\mathcal{K}_k$ may be constructed by Arnoldi's method [1,2,25], which computes v_{i+1} by orthonormalizing Av_i against the previous basis vectors $v_1, \ldots, v_i$, i.e., $v_{i+1} = \alpha(I - V_iV_i^H)Av_i$ with α chosen such that v_{i+1} has unit norm. The dominant operations are thus matrix-vector products, weighted vector sums, vector scalings, and scalar products.

Because of the high problem dimension occurring in quantum system computations the memory capacity of even a large computing cluster is not sufficient to store even a single vector in standard format. For this reason, the vectors are usually stored in a data sparse tensor format, like the tensor train [19, 28, 29] or the hierarchical tensor formats [15, 16, 24]. While this makes storing vectors possible in the first place for these applications it entails the drawback that vector operations cannot be carried out exactly. Instead only approximations of the intended quantities are available, e.g., in case of the matrix-vector multiplication, instead of Av_i we obtain $Av_i + f_i$ where f_i is some small unknown vector.

In this paper we consider an inexact Arnoldi method, where matrix-vector multiplication, vector addition and vector scaling are inaccurate. On the other hand we do assume that scalar products can be evaluated exactly, as is the case for the mentioned tensor formats.

Apart from quantum system computations inexact vector operations occur in further scenarios of practical interest. Our analysis is independent of the actual source of perturbation and thus applies to all these situations alike. Other applications include

- *mixed precision arithmetic:* Consider an Arnoldi method where the basis vectors of the Krylov space are stored in single precision whereas all remaining quantities and computations are handled in double precision. (In our analysis double precision is approximated by infinite precision.) This approach effectively halves the memory requirements of Arnoldi's method which are dominated by the need to store the basis vectors. However every computed vector has to be rounded to single precision – introducing a perturbation to these operations;
- *sparse eigenvectors:* If the eigenvectors of interest are known to be well approximable by sparse vectors (i.e., vectors with only a few non-zeros) then it is natural to desire sparse Arnoldi basis vectors. This could be achieved by thresholding, i.e., neglecting small elements in the computed basis vectors - constituting a perturbation. In this scenario, see, e.g., [4,5,7], matrix-vector-multiplication and vector addition (including sparsification) are inexact whereas vector scaling and scalar products are exact;
- *Lyapunov equations:* A GMRES-like method to solve Lyapunov equations $\mathcal{A}X + X\mathcal{A}^T = -BB^T$ (cf. [10,21,22]) builds a search space using Arnoldi's method. Here the basis vectors are vectorizations of matrices and the system matrix $\mathcal{A}$ consists of a sum of Kronecker products. Since the solution X is often known to be well approximable by a low rank matrix, each basis vector is the vectorization of a matrix truncated to low rank. These truncations can be interpreted as perturbation in the vector operations.

The matrix A as it arises in quantum system computation is Hermitian. Thus later in this paper we will assume that A is Hermitian. In this case the Arnoldi method reduces to the Lanczos method which features a short recurrence relation. Unfortunately this short recurrence is very sensitive to perturbations and their presence leads to a rapid loss of orthogonality and to spurious eigenvalues. There is a variant of the Lanczos methods without these drawbacks using full recurrence, which is used, e.g., in the popular ARPACK package [25]. We restrict the scope to this Lanczos method with full recurrence. To unify notation we speak of Arnoldi's method also in the case of a Hermitian A.

Classically, i.e., without perturbations, Arnoldi's method constructs an orthonormal basis matrix $V_k = [v_1, \ldots, v_k]$ and a Hessenberg matrix $H_k \in \mathbb{C}^{k\times k}$ such that the so-called *Arnoldi relation*

$$AV_k = V_kH_k + v_{k+1}h_{k+1,k}e_k^T \tag{1}$$

holds. Here e_k denotes the k-th column of the identity matrix. The purpose of this paper is to analyze what happens to orthonormality and the Arnoldi relation in the presence of perturbations. More precisely, we consider three subproblems. a) We provide bounds on the distance from orthonormality of the now-not-anymore orthogonal basis vectors $\tilde{V}_k$ obtained using a certain variant of Gram-Schmidt orthogonalization. b) The Arnoldi relation (1) does not hold anymore. However, we will prove that its residual norm, $\|AV_k - V_kH_k - v_{k+1}h_{k+1,k}e_k^T\|$, is small. In the spirit of a backward error analysis we then show that a relation of the form (1) holds, where V_k, H_k are replaced by the computed counterparts $\tilde{V}_k, \tilde{H}_k$ and A is replaced by $A+E$, where E is small. c) If A is Hermitian, it is natural to restrict E to be Hermitian as well. Unfortunately, it turns out that in general there is no Hermitian E such that the Arnoldi relation holds for $A+E$. In order to rescue Relation (1) we have to give up the Hessenberg structure of H_k. We will show that upon replacement of H_k by any Hermitian $k \times k$ matrix B_k, there exists a Hermitian E such that (1) holds for $A+E$. Moreover, we provide bounds on the norm of E for suitable choices of B_k.

These results complement earlier work in this field. For background on the Arnoldi process without perturbations see, e.g., [32, section 6.5] or [8, 23, 26, 31, 33]. Arnoldi's method with perturbations was considered in [11, 36–38]. In every case only inexact matrix-vector products were assumed. Consequences of perturbations to the Gram-Schmidt orthogonalization process were analyzed in [3, 13, 18, 31].

The paper is structured as follows: We state our inexact Arnoldi algorithm in Section 2 and discuss its differences to the classical method. Then, in Section 3 we analyze the distance of the obtained basis from orthonormality and give bounds for different implementations of the orthogonalization step. We then show that the obtained subspace can be interpreted as an exact Krylov subspace of a matrix close to A and establish bounds for the backward error. Numerical examples illustrate the theoretical results in Section 5. Finally we offer some concluding remarks in Section 6.

2 The algorithm and notation

We will analyze the following method. Initialized with a matrix A and a normalized vector $\tilde{v}_1$ it constructs a search space basis $\tilde{V}_k$ and a Hessenberg matrix $\tilde{H}_k$ consisting of orthogonalization coefficients. In order to emphasize that their computation entailed perturbations we named the variables $\tilde{V}_k$ and $\tilde{H}_k$ instead of V_k and H_k, i.e., with tildes. All other variables that also appear in the standard Arnoldi method have a tilde, too. The matrices $D_k := \tilde{V}_k^H \tilde{V}_k$ are new; they do not appear in the standard Arnoldi method. The vectors $f_k^{(M)}$, $f_k^{(0)}$, and $f_k^{(S)}$ model the perturbations in matrix-vector multiplication, in orthogonalization and in vector scaling, respectively, in the kth step.

Algorithm 1 *Inexact Arnoldi Method*
Input: $A \in \mathbb{C}^{n\times n}$, $\tilde{v}_1 \in \mathbb{C}^n$ normalized, $m \in \mathbb{N}$
Output: $\tilde{V}_{m+1} \in \mathbb{C}^{n\times m+1}$, $\tilde{H}_m \in \mathbb{C}^{m\times m}$, $\tilde{h}_{m+1,m}$, $D_{m+1} \in \mathbb{C}^{m+1\times m+1}$,

1: $\tilde{V}_1 = \tilde{v}_1$, $D_1 = 1$, $\tilde{H}_0 = [\,] \in \mathbb{C}^{0\times 0}$ (initialization)
2: **for** $k = 1, 2, 3, \ldots, m$ **do**
3: $\tilde{w}_{k+1} = A\tilde{v}_k - f_{k+1}^{(M)}$ (matrix multiplication, perturbed)
4: $[\tilde{v}_{k+1}, \tilde{h}_{1:k,k}, \tilde{h}_{k+1,k}, D_{k+1}] = \mathrm{ComGS}(\tilde{w}_{k+1}, \tilde{V}_k, D_k)$ (orthogonalzation)
5: $\tilde{H}_k = \begin{bmatrix} \tilde{H}_{k-1} & \tilde{h}_{1:k-1,k} \\ \tilde{h}_{k,k-1}e_{k-1}^T & \tilde{h}_{k,k} \end{bmatrix}$ (update $\tilde{H}_k$)
6: $\tilde{V}_{k+1} = [\tilde{V}_k, \tilde{v}_{k+1}]$ (update $\tilde{V}_k$)
7: **end for**

Algorithm 2 *ComGS*

Input: $\tilde{w}_{k+1} \in \mathbb{C}^n$, $\tilde{V}_k \in \mathbb{C}^{n\times k}$, $D_k \in \mathbb{C}^{k\times k}$

Output: $\tilde{v}_{k+1} \in \mathbb{C}^n, \tilde{h}_{1:k,k} \in \mathbb{C}^k, \tilde{h}_{k+1,k} \in \mathbb{C}, D_{k+1} \in \mathbb{C}^{k+1\times k+1}$

1: $\tilde{h}_{1:k,k} = D_k^{-1}\tilde{V}_k^H\tilde{w}_{k+1}$ (orthogonalization coefficients)

2: $\tilde{l}_{k+1} = \tilde{w}_{k+1} - \tilde{V}_k\tilde{h}_{1:k,k} - f_{k+1}^{(0)}$ (orthogonalization, perturbed)

3: $\tilde{h}_{k+1,k} = \|\tilde{l}_{k+1}\|_2$

4: $\tilde{v}_{k+1} = (\tilde{l}_{k+1} - f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$ (normalization, perturbed)

5: $D_{k+1} = \begin{bmatrix} D_k & \tilde{V}_k^H\tilde{v}_{k+1} \\ \tilde{v}_{k+1}^H\tilde{V}_k & \tilde{v}_{k+1}^H\tilde{v}_{k+1} \end{bmatrix}$ (update D_k)

We have the following remarks. First, the standard Arnoldi method is obtained upon omitting the perturbation vectors $f_{k+1}^{(*)}$ and the matrix D_k in algorithms 1 and 2.

Second, we use an unusual kind of projection to orthogonalize $\tilde{w}_k$ against $\tilde{V}_k$. Since we know that the basis is not orthonormal, we use the projector $I - \tilde{V}_k(\tilde{V}_k^H\tilde{V}_k)^{-1}\tilde{V}_k^H$ for non-orthogonal bases (instead of $I-\tilde{V}_k\tilde{V}_k^H$ for orthogonal bases). For this purpose we need to construct the cross product matrix $D_k = \tilde{V}_k^H\tilde{V}_k$ which can be updated during the iteration (Step 5 of Algorithm 2). The more complicated projector is chosen in order to alleviate the damage, the perturbations inflict on the orthogonality of $\tilde{V}_k$. Since the use of D_k^{-1} in Step 1 of Algorithm 2 compensates for the non-orthogonality in $\tilde{V}_k$, we speak of the Compensated Gram-Schmidt process (ComGS).

ComGS is a slight modification of Classical Gram Schmidt (CGS) to work with non-orthogonal bases. It inherits the property of CGS that all scalar products $\tilde{v}_1^H\tilde{w}_{k+1}, \ldots, \tilde{v}_k^H\tilde{w}_{k+1}$ may be computed in parallel (whereas in modified Gram Schmidt (MGS) they have do be evaluated sequentially). We note that also MGS can deal to some extend with non-orthogonal bases, as it implicitly solves a linear system with the matrix D_k (as ComGS does in Step 1), see [3, page 308 f.].

Our numerical tests will show that ComGS retains orthogonality much better than CGS and is comparable to MGS. In contrast to those standard schemes, ComGS additionally provides implicit access to a second basis that is even closer to orthogonality, see below. The price to pay is that ComGS needs $6nk + \mathcal{O}(k^3)$ floating point operations (flops) in the kth iteration whereas the CGS and MGS schemes require only $4nk$ flops.

Finally we note that in the unperturbed case, i.e., when all perturbations are zero, D_k becomes the identity matrix, and thus Algorithm 2 reduces to CGS and Algorithm 1 to the standard Arnoldi method. Note that we exclude the case that the algorithm breaks down, i.e., that $\tilde{h}_{k+1,k}$ is zero or that the matrix D_k is singular.

Notation We define

$$\hat{l}_{k+1} := \tilde{w}_{k+1} - \tilde{V}_k\tilde{h}_{1:k,k} = \tilde{l}_{k+1} + f_{k+1}^{(0)}. \tag{2}$$

Note that these exactly orthogonalized vectors are not available in practice, but we will use them in our analysis below. Further useful quantities include

$$D_k =: C_k^H C_k, \quad \tilde{\mathcal{K}}_k := \mathrm{img}(\tilde{V}_k) = \mathrm{img}(\hat{V}_k), \quad \hat{V}_k := \tilde{V}_k C_k^{-1}, \quad P_{\tilde{\mathcal{K}}_k} := \tilde{V}_k D_k^{-1}\tilde{V}_k^H = \hat{V}_k\hat{V}_k^H. \tag{3}$$

The Cholesky factor C_k of D_k, is useful for solving the linear systems with D_k that occur in Step 1 of Algorithm 2. It can be updated along with D_k, via [17, Proof of Theorem 10.1],

$$C_{k+1} = \begin{bmatrix} C_k & c_{k+1} \\ 0 & \gamma_{k+1} \end{bmatrix} \quad \text{with} \quad c_{k+1} = C_k^{-H}(\tilde{V}_k^H\tilde{v}_{k+1}), \quad \gamma_{k+1} = \sqrt{\|\tilde{v}_{k+1}\|_2^2 - \|c_{k+1}\|_2^2}.$$

And doing so lowers the cost of ComGS from $6nk+\mathcal{O}(k^3)$ to $6nk+\mathcal{O}(k^2)$. Another useful property of C_k is that $\hat{V}_k = \tilde{V}_k C_k^{-1}$ has orthonormal columns. $\tilde{\mathcal{K}}_k$ is the search space used by our method. Note that $\tilde{\mathcal{K}}_k$ is in general not a Krylov subspace for A. However, since $\mathrm{img}(\tilde{V}_k) = \mathrm{img}(\hat{V}_k)$ an orthonormal basis of $\tilde{\mathcal{K}}_k$ is implicitly available. Finally, $P_{\tilde{\mathcal{K}}_k}$ denotes the orthogonal projector onto the search space $\tilde{\mathcal{K}}_k$. Note that $\hat{V}_k$ cannot be formed explicitly as its computation involves vector sums which are inexact.

3 Distance to orthogonality

In this section we analyze the distance to orthogonality of the basis $\tilde{V}_k$ produced by the inexact Arnoldi method, Algorithm 1. Note that only the perturbations $f_k^{(0)}$ and $f_k^{(S)}$ during orthonormalization in Algorithm 2 play a role here; those occurring during the matrix-vector multiplication are insignificant for the deviation from orthogonality. Indeed, in the special case with perturbation only in Algorithm 1, but not in Algorithm 2, the columns of $\tilde{V}_k$ would be orthonormal. (However, the spanned space would nevertheless cease to be a Krylov subspace for A.)

As measures for the distance to orthonormality we will use the quantities $\|\tilde{V}_k^H \tilde{v}_{k+1}\|_2$, $\|D_k - I\|_* = \|\tilde{V}_k^H \tilde{V}_k - I\|_*$, and $\|C_k - I\|_*$, where $\|.\|_*$ denote either the spectral or the Frobenius norm. Especially $\|D_k - I\|_2$ is commonly used (e.g., in [14]) and is a good estimator for the canonic distance $\delta_{\text{orth}}(\tilde{V}_k) := \min_{U \in \mathbb{C}^{n\times k}} \{\|\tilde{V}_k - U\|_2 : U^H U = I_k\}$. More precisely, we have [17, Problem 19.14] $\|D - I_k\|_2/(\|\tilde{V}_k\|_2 + 1) \le \delta_{\text{orth}}(\tilde{V}_k) \le \|D - I_k\|_2$.

Our bounds depend on the relative norm of the perturbations $f_k^{(0)}$, $f_k^{(S)}$. We assume that $\|f_k^{(0)}\|$ and $\|f_k^{(S)}\|$ are small compared to $\|\tilde{w}_k\|$ and $\|\tilde{l}_{k+1}\|$, respectively. This is the case, whenever $f_k^{(0)}$ may be interpreted as the error which arises in the vector sum in Step 2 of Algorithm 2, because $\|w_k\|_2$ is the largest summand within this sum. More precisely, we will assume the bounds $\|f_k^{(0)}\|_2 < k\varepsilon\|w_k\|_2$ and $\|f_k^{(S)}\|_2 < \varepsilon\|\tilde{l}_k\|_2$. The parameter ε depends on the actual perturbation source; in case of tensor approximation ε denotes the truncation threshold; whereas in case of mixed precision arithmetic ε can be interpreted as the single machine precision $\varepsilon_s \approx 6 \cdot 10^{-8}$. Clearly, the best we can hope for is that the distance form orthogonality is on the order of ε. This turns out to be the case – up to an unpleasant constant.

Theorem 1 *Let $A \in \mathbb{C}^{n\times n}$ and $\tilde{v}_1 \in \mathbb{C}^n$ with $\|\tilde{v}_1\|_2 = 1$. Let $\tilde{H}_k$ and $\tilde{V}_{k+1} = [\tilde{V}_k, \tilde{v}_{k+1}]$ be as in Algorithm 1 after k iterations. Assume that the perturbations in steps 2 and 4 of Algorithm 2 are bounded by $\|f_{k+1}^{(0)}\|_2 \le k\varepsilon\|\tilde{w}_{k+1}\|_2$ and $\|f_{k+1}^{(S)}\|_2 < \varepsilon\|\tilde{l}_{k+1}\|_2$, for some $\varepsilon < 1/(k+2)$. Then*

$$\|\tilde{V}_k^H \tilde{v}_{k+1}\|_2 \le \|\tilde{V}_k\|_2 (k + 1 + k\|\tilde{V}_k\|_2 \kappa_k) \frac{\varepsilon}{1 - k\varepsilon} \le k^2 (2 + \kappa_k) \frac{\varepsilon}{1 - (k+2)\varepsilon}. \tag{4}$$

where $\kappa_k := \|\tilde{h}_{1:k,k}\|_2 / \tilde{h}_{k+1,k}$.

The proof will require the following lemma for support.

Lemma 1 *Let $V = [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$, with $\|v_i\|_2 \le 1 + \varepsilon$ for all $i = 1, \ldots, k$ for some $\varepsilon > -1$, then $\|V\|_2 \le \sqrt{k}(1 + \varepsilon)$.*

Proof This follows from $\|V\|_2^2 \le \|V\|_F^2 \le k(1 + \varepsilon)^2$. □

Proof (of Theorem 1) From Algorithm 2, equation (2) and the assumed bound on $\|f_{k+1}^{(S)}\|_2$ we have $\|\tilde{v}_i\|_2 \le 1+\varepsilon$, $\tilde{h}_{k+1,k} > 0$, and $\tilde{v}_{k+1} = (\hat{l}_{k+1} + f_{k+1}^{(0)} + f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$. Hence, using orthogonality of $\hat{l}_{k+1}$ to $\tilde{V}_k$, we have

$$\begin{aligned} \|\tilde{V}_k^H \tilde{v}_{k+1}\|_2 = \frac{\|\tilde{V}_k^H (\hat{l}_{k+1} + f_{k+1}^{(0)} + f_{k+1}^{(S)})\|_2}{\tilde{h}_{k+1,k}} &= \frac{\|\tilde{V}_k^H (f_{k+1}^{(0)} + f_{k+1}^{(S)})\|_2}{\tilde{h}_{k+1,k}} \\ &\le \frac{\|\tilde{V}_k\|_2}{\tilde{h}_{k+1,k}} (\|f_{k+1}^{(0)}\|_2 + \|f_{k+1}^{(S)}\|_2). \end{aligned} \tag{5}$$

For $\|f_{k+1}^{(0)}\|_2$ we obtain

$$\begin{aligned} \|f_{k+1}^{(0)}\|_2 &\le k\varepsilon\|\tilde{w}_{k+1}\|_2 \\ &= k\varepsilon\|\tilde{V}_k \tilde{h}_{1:k,k} + f_{k+1}^{(0)} + \tilde{l}_{k+1}\|_2 \\ &\le k\varepsilon\|f_{k+1}^{(0)}\|_2 + k\varepsilon\tilde{h}_{k+1,k} + k\varepsilon\|\tilde{V}_k \tilde{h}_{1:k,k}\|_2. \end{aligned}$$

Solving for $\|f_{k+1}^{(0)}\|_2$ results in

$$\|f_{k+1}^{(0)}\|_2 \leq \left(\tilde{h}_{k+1,k} + \|\tilde{V}_k\|_2 \|\tilde{h}_{1:k,k}\|_2\right) \frac{k\varepsilon}{1-k\varepsilon} = \tilde{h}_{k+1,k}(1 + \|\tilde{V}_k\|_2 \kappa_k) \frac{k\varepsilon}{1-k\varepsilon}$$

Using $\|f_{k+1}^{(S)}\|_2 \leq \tilde{h}_{k+1,k}\varepsilon \leq \tilde{h}_{k+1,k}\varepsilon/(1-k\varepsilon)$ yields

$$\|f_{k+1}^{(0)}\|_2 + \|f_{k+1}^{(S)}\|_2 \leq \tilde{h}_{k+1,k} \left((1 + \|\tilde{V}_k\|_2 \kappa_k) \frac{k\varepsilon}{1-k\varepsilon} + \frac{\varepsilon}{1-k\varepsilon}\right) = \tilde{h}_{k+1,k}(k+1+k\|\tilde{V}_k\|_2 \kappa_k) \frac{\varepsilon}{1-k\varepsilon}.$$

Plugging this into (5) proves the first inequality. Finally, with Lemma 1 we have

$$\begin{aligned} &\|\tilde{V}_k\|_2 (k+1+k\|\tilde{V}_k\|_2 \kappa_k) \frac{\varepsilon}{1-k\varepsilon} \leq \sqrt{k}(1+\varepsilon) \left((k+1)(1+\varepsilon) + k\sqrt{k}(1+\varepsilon)\kappa_k\right) \frac{\varepsilon}{1-k\varepsilon} \\ &\leq \sqrt{k}(1+\varepsilon)^2 (k+1+k\sqrt{k}\kappa_k) \frac{\varepsilon}{1-k\varepsilon} \leq \sqrt{k}(k+1+k\sqrt{k}\kappa_k) \frac{\varepsilon}{1-(k+2)\varepsilon} \leq k^2(2+\kappa_k) \frac{\varepsilon}{1-(k+2)\varepsilon}. \end{aligned}$$

□

The bound of Theorem 1 describes the degree of orthogonality of $\tilde{v}_{k+1}$ to its predecessors $[\tilde{v}_1, \ldots, \tilde{v}_k]$. Two bounds are given. The one involving $\|\tilde{V}_k\|_2$ is more complicated, but also sharper than the other one. We deduct that $\|\tilde{V}_k^H \tilde{v}_{k+1}\|_2$ is small, i.e., on the order of ε, whenever the factor $(2+\kappa_k)$ is not too large. In this sense $(2+\kappa_k)$ can be interpreted as a condition number, with $(2+\kappa_k) \in [2, \infty)$. κ_k gets large only if the subdiagonal element $\tilde{h}_{k+1,k}$ of the Hessenberg matrix is tiny compared to the remaining elements in the k-th column of $\tilde{H}_k$. Hence as long as $A\tilde{v}_k$ is securely linearly independent of the previous basis vectors, κ_k will be moderate and the upper bound (4) will be on the order of ε. However, $(2+\kappa_k)$ can become arbitrarily large, if the subdiagonal element $\tilde{h}_{k+1,k}$ is very small. Usually, this situation is called a "lucky breakdown" as it indicates convergence of the search space to an invariant subspace of A. Here, tiny subdiagonal elements are "unlucky", because the bounds (4) get large and we loose orthogonality of $\tilde{V}_{k+1}$.

Note that in this case only $\tilde{v}_{k+1}$ is far from being orthogonal to $\tilde{V}_k$. However, $\tilde{V}_k$ itself will be an almost orthogonal basis of the almost invariant subspace. Moreover, note that the bound (4) is not recursive, i.e., even if $\tilde{V}_k$ is highly non-orthogonal, the bound for the next vector $\tilde{v}_{k+1}$ may be small nevertheless (provided that κ_k is moderate). This fortunate behavior is a huge advantage of the ComGS method over CGS. Indeed, CGS does not recover once orthogonality is lost, but on the contrary only loses orthogonality even faster [12].

We note two relations to other methods: a) if κ_k is large then $[\tilde{H}_k^T, \tilde{h}_{k+1,k} e_k]^T$ is ill-conditioned, i.e., the R-factor of the QR-factorization of the matrix $[\tilde{v}_1, A\tilde{V}_k]$ is ill-conditioned. This implies that the matrix $[\tilde{v}_1, A\tilde{V}_k]$ itself is ill-conditioned - in which case Gram Schmidt methods are known to loose orthogonality. So the occurrence of the factor κ should not surprise. b) Loss of orthogonality in case of convergence was also described for the Lanczos algorithm in [30]. However, there the convergence of a single eigenvector in the Krylov subspace is enough to trigger the loss of orthogonality, whereas in our method the whole Krylov subspace has to be a converged invariant subspace (indicated by the tiny subdiagonal element $h_{k+1,k}$). This makes our method much more stable than the Lanczos algorithm.

However, the situation is not entirely satisfying. In the case when $A\tilde{v}_k$ is nearly linear dependent of $\tilde{V}_k$, i.e., when $\|\tilde{l}_{k+1}\| = \mathcal{O}(\varepsilon)$, then $\kappa_k = \mathcal{O}(\varepsilon^{-1})$. In this case the bound (4) is of order one and orthogonality is lost. Hence, we look for an improved orthogonalization scheme. We will add a reorthogonalization step (analyzed for QR factorizations in [6]) amounting to the following modification of Algorithm 2.

Algorithm 3 *ComGSre (with reorthogonalization)*
Input: $\tilde{w}_{k+1} \in \mathbb{C}^n$, $\tilde{V}_k \in \mathbb{C}^{n \times k}$, $D_k \in \mathbb{C}^{k \times k}$
Output: $\tilde{v}_{k+1} \in \mathbb{C}^n, \tilde{h}_{1:k,k} \in \mathbb{C}^k, \tilde{h}_{k+1,k} \in \mathbb{C}, D_{k+1} \in \mathbb{C}^{k+1 \times k+1}$
1: $s^{(0)} = D_k^{-1} \tilde{V}_k^H \tilde{w}_{k+1}$

2: $\tilde{l}_{k+1}^{(0)} = \tilde{w}_{k+1} - \tilde{V}_k s^{(0)} - f_{k+1}^{(0)}$ (orthogonalization, perturbed)

3: $s^{(1)} = D_k^{-1} \tilde{V}_k^H \tilde{l}_{k+1}^{(0)}$

4: $\tilde{l}_{k+1}^{(1)} = \tilde{l}_{k+1}^{(0)} - \tilde{V}_k s^{(1)} - f_{k+1}^{(1)}$ (reorthogonalization, perturbed)

5: $\tilde{h}_{k+1,k} = \|\tilde{l}_{k+1}^{(1)}\|_2$, $\tilde{h}_{1:k,k} = s^{(0)} + s^{(1)}$

6: $\tilde{v}_{k+1} = (\tilde{l}_{k+1}^{(1)} - f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$ (normalization, perturbed)

7: $D_{k+1} = \begin{bmatrix} D_k & \tilde{V}_k^H \tilde{v}_{k+1} \\ \tilde{v}_{k+1}^H \tilde{V}_k & \tilde{v}_{k+1}^H \tilde{v}_{k+1} \end{bmatrix}$ (update D)

ComGSre needs $10nk + \mathcal{O}(k^3)$ flops, slightly more than the $8nk$ flops of CGS and MGS with reorthogonalization.

As before we define the exactly orthogonalized vectors

$$\hat{l}_{k+1}^{(0)} := \tilde{w}_{k+1} - \tilde{V}_k s^{(0)} = \tilde{l}_{k+1}^{(0)} + f_{k+1}^{(0)} \quad \text{and} \quad \hat{l}_{k+1}^{(1)} := \tilde{l}_{k+1}^{(0)} - \tilde{V}_k s^{(1)} = \tilde{l}_{k+1}^{(1)} + f_{k+1}^{(1)}.$$

Reorthogonalization aims to improve orthogonality of $\tilde{v}_{k+1}$ by projecting the result of the first orthogonalization step $\tilde{l}_{k+1}^{(0)}$ a second time. In doing so, $f_{k+1}^{(0)}$ is orthogonalized to $\tilde{V}_k$. Unfortunately, also the vector sum in the reorthogonalization step is inexact which introduces a further perturbation $f_{k+1}^{(1)}$. However, while $f_{k+1}^{(0)}$ is small compared to w_k, we can assume $f_{k+1}^{(1)}$ to be small compared to $\tilde{l}_{k+1}^{(0)}$ (the largest term in the sum of Step 4), i.e., $\|f_{k+1}^{(1)}\|_2 \leq k\varepsilon \|\tilde{l}_{k+1}^{(0)}\|_2$. We have the following result for Algorithm 3 which is analogous to Theorem 1.

Theorem 2 *Let $A \in \mathbb{C}^{n\times n}$ and $\tilde{v}_1 \in \mathbb{C}^n$ with $\|\tilde{v}_1\|_2 = 1$. Let $\tilde{H}_k$ and $\tilde{V}_{k+1} = [\tilde{V}_k, \tilde{v}_{k+1}]$ be as in Algorithm 1 after k iterations. Assume that the perturbations in steps 2, 4, and 6 of Algorithm 3 are bounded by $\|f_{k+1}^{(0)}\|_2 \leq k\varepsilon\|\tilde{w}_{k+1}\|_2$, $\|f_{k+1}^{(1)}\|_2 \leq k\varepsilon\|\tilde{l}_{k+1}^{(0)}\|_2$, and $\|f_{k+1}^{(S)}\|_2 \leq \varepsilon\|\tilde{l}_{k+1}^{(1)}\|_2$, respectively, for some $\varepsilon < 1/(2k+2)$. Then*

$$\|\tilde{V}_k^H \tilde{v}_{k+1}\|_2 \leq \|\tilde{V}_k\|_2 \left(k + 1 + k^2\varepsilon \|\tilde{V}_k\|_2 \kappa_k\right) \frac{\varepsilon}{1-2k\varepsilon} \leq k^2 \left(2 + k\varepsilon\kappa_k\right) \frac{\varepsilon}{1-(2k+2)\varepsilon}, \tag{6}$$

where $\kappa_k := \|\tilde{h}_{1:k,k}\|_2/\tilde{h}_{k+1,k}$.

Proof From steps 2–5 of Algorithm 3 we have

$$\begin{aligned} \|f_{k+1}^{(0)}\|_2 &\leq k\varepsilon\|\tilde{w}_{k+1}\|_2 = k\varepsilon\|\tilde{l}_{k+1}^{(1)} + f_{k+1}^{(1)} + \tilde{V}_k s^{(1)} + f_{k+1}^{(0)} + \tilde{V}_k s^{(0)}\|_2 \\ &= k\varepsilon\|f_{k+1}^{(0)} + f_{k+1}^{(1)} + \tilde{l}_{k+1}^{(1)} + \tilde{V}_k \tilde{h}_{1:k,k}\|_2 \\ &\leq k\varepsilon\|f_{k+1}^{(0)}\|_2 + k\varepsilon\|f_{k+1}^{(1)}\|_2 + k\varepsilon\,\tilde{h}_{k+1,k} + k\varepsilon\|\tilde{V}_k\|_2\|\tilde{h}_{1:k,k}\|_2. \end{aligned}$$

Solving for $\|f_{k+1}^{(0)}\|_2$ we obtain

$$\|f_{k+1}^{(0)}\|_2 \leq \frac{k\varepsilon}{1-k\varepsilon}\left(\|f_{k+1}^{(1)}\|_2 + \tilde{h}_{k+1,k} + \|\tilde{V}_k\|_2\|\tilde{h}_{1:k,k}\|_2\right). \tag{7}$$

For $\|f_{k+1}^{(1)}\|_2$ we obtain

$$\begin{aligned} \|f_{k+1}^{(1)}\|_2 &\leq k\varepsilon\|\tilde{l}_{k+1}^{(0)}\|_2 = k\varepsilon\|f_{k+1}^{(1)} + \tilde{l}_{k+1}^{(1)} + \tilde{V}_k s^{(1)}\|_2 \\ &\leq k\varepsilon\|f_{k+1}^{(1)}\|_2 + k\varepsilon\tilde{h}_{k+1,k} + k\varepsilon\|\tilde{V}_k s^{(1)}\|_2. \end{aligned} \tag{8}$$

Since $\hat{l}_{k+1}^{(0)}$ is orthogonal to $\tilde{V}_k$, $\|\tilde{V}_k s^{(1)}\|_2$ can be bounded by

$$\|\tilde{V}_k s^{(1)}\|_2 = \|\tilde{V}_k D_k^{-1} \tilde{V}_k^H \tilde{l}_{k+1}^{(0)}\|_2 = \|P_{\tilde{\mathcal{K}}_k}(\hat{l}_{k+1}^{(0)} - f_{k+1}^{(0)})\|_2 = \|P_{\tilde{\mathcal{K}}_k} f_{k+1}^{(0)}\|_2 \leq \|f_{k+1}^{(0)}\|_2. \tag{9}$$

Combining (7), (8), and (9) yields

$$\begin{aligned}\|f_{k+1}^{(1)}\|_2 &\le k\varepsilon\|f_{k+1}^{(1)}\|_2 + k\varepsilon\tilde{h}_{k+1,k} + k\varepsilon\frac{k\varepsilon}{1-k\varepsilon}\left(\|f_{k+1}^{(1)}\| + \tilde{h}_{k+1,k} + \|\tilde{V}_k\|_2\|\tilde{h}_{1:k,k}\|_2\right)\\ &\le \frac{k\varepsilon}{1-k\varepsilon}\|f_{k+1}^{(1)}\|_2 + \frac{k\varepsilon}{1-k\varepsilon}\tilde{h}_{k+1,k} + \frac{(k\varepsilon)^2}{1-k\varepsilon}\|\tilde{V}_k\|_2\|\tilde{h}_{1:k,k}\|_2.\end{aligned}$$

Solving for $\|f_{k+1}^{(1)}\|_2$ results in

$$\|f_{k+1}^{(1)}\|_2 \le \left(\tilde{h}_{k+1,k} + k\varepsilon\|\tilde{V}_k\|_2\|\tilde{h}_{1:k,k}\|_2\right)\frac{k\varepsilon}{1-2k\varepsilon} = \tilde{h}_{k+1,k}\left(1 + k\varepsilon\|\tilde{V}_k\|_2\kappa_k\right)\frac{k\varepsilon}{1-2k\varepsilon}.$$

Consequently, using $\tilde{v}_{k+1} = (\hat{l}_{k+1}^{(1)} + f_{k+1}^{(1)} + f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$, orthogonality of $\hat{l}_{k+1}^{(1)}$ to $\tilde{V}_k^H$, and $\|f_{k+1}^{(S)}\|_2 \le \tilde{h}_{k+1,k}\varepsilon \le \tilde{h}_{k+1,k}\varepsilon/(1-2k\varepsilon)$, we have

$$\begin{aligned}\|\tilde{V}_k^H\tilde{v}_{k+1}\|_2 &= \frac{\|\tilde{V}_k^H(f_{k+1}^{(1)} + f_{k+1}^{(S)})\|_2}{\tilde{h}_{k+1,k}} \le \frac{\|\tilde{V}_k\|_2}{\tilde{h}_{k+1,k}}(\|f_{k+1}^{(1)}\|_2 + \|f_{k+1}^{(S)}\|_2)\\ &\le \|\tilde{V}_k\|_2\Big(\left(1 + k\varepsilon\,\|\tilde{V}_k\|_2\kappa_k\right)\frac{k\varepsilon}{1-2k\varepsilon} + \frac{\varepsilon}{1-2k\varepsilon}\Big) \le \|\tilde{V}_k\|_2\left(k + 1 + k^2\varepsilon\,\|\tilde{V}_k\|_2\kappa_k\right)\frac{\varepsilon}{1-2k\varepsilon},\end{aligned}$$

which proves the first inequality. Finally, with Lemma 1 we have

$$\begin{aligned}\|\tilde{V}_k\|_2\left(k + 1 + k^2\varepsilon\,\|\tilde{V}_k\|_2\kappa_k\right)\frac{\varepsilon}{1-2k\varepsilon} &\le \sqrt{k}(1+\varepsilon)\left((k+1)(1+\varepsilon) + \sqrt{k^5}\varepsilon(1+\varepsilon)\kappa_k\right)\frac{\varepsilon}{1-2k\varepsilon}\\ &\le \sqrt{k}(1+\varepsilon)^2\left(k + 1 + \sqrt{k^5}\varepsilon\kappa_k\right)\frac{\varepsilon}{1-2k\varepsilon} \le \sqrt{k}\left(k + 1 + \sqrt{k^5}\varepsilon\kappa_k\right)\frac{\varepsilon}{1-(2k+2)\varepsilon}\\ &\le k^2\left(2 + k\varepsilon\kappa_k\right)\frac{\varepsilon}{1-(2k+2)\varepsilon}.\end{aligned}$$

□

Comparing bounds (4) and (6), we see that the reorthogonalization step improves things. The dominant difference is the replacement of the term $(2 + \kappa_k)$ by $(2 + k\varepsilon\kappa_k)$. This implies little change for moderate values of κ_k (distance from orthogonality is still on the order of ε). For large κ_k, on the other hand, we observe a huge improvement. Simply speaking, $\tilde{v}_{k+1}$ will be almost orthogonal for values of κ_k up to the order of ε^{-1}.

One could ask the question if we could get even better results by reorthogonalizing repeatedly. This situation is completely analyzed in [31, pp. 115–117], where the "twice is enough" algorithm is presented for the two vectors case and it is referred to Kahan for a corresponding analysis. An extension to several nearly orthonormal vectors is given in [6]. This "twice is enough" rule of thumb also holds in our situation under mild assumptions. We have already established that κ_k can reach the order $\mathcal{O}(\varepsilon^{-1})$. In order to grow even further, the perturbation $f_{k+1}^{(0)}$ would have to lie almost completely in $\mathrm{img}(\tilde{V}_k)$. This is very unlikely as long as the perturbations can be interpreted as being random.

So far we have only considered the distance to orthogonality of just one vector to the previous basis $\tilde{V}_k$. Now we shift attention to the whole cross product matrix aiming for bounds on the distance of D_k or its Cholesky factor C_k from the identity. Abusing notation, we formulate the bounds for both cases (with and without reorthogonalization) as follows.

Corollary 1 *Let D_k, C_k, and $\tilde{H}_k$ result from k iterations of Algorithm 1 applied to $A \in \mathbb{C}^{n\times n}$ and $\tilde{v}_1 \in \mathbb{C}^n$ with $\|\tilde{v}_1\|_2 = 1$. Let $\ell \in \{0,1\}$ be the number of reorthogonalizations used in ComGS. Let $\kappa_k := \|\tilde{h}_{1:k,k}\|_2/\tilde{h}_{k+1,k}$ and $\kappa_{\max,k} := \max_{i=1,\dots,k}\kappa_i$. For $\ell = 0$ assume that the perturbations in steps 2 and 4 of Algorithm 2 are bounded by $\|f_k^{(0)}\|_2 \le k\varepsilon\|\tilde{w}_{k+1}\|_2$ and $\|f_k^{(S)}\|_2 \le \varepsilon\|\tilde{l}_{k+1}\|_2$ for some $\varepsilon < (\sqrt{(k+1)^5}(2+\kappa_{\max,k}) + 3k + 5)^{-1}$. Similarly, for $\ell = 1$ assume that the perturbations in steps 2, 4, and 6 of Algorithm 3 are bounded by $\|f_{k+1}^{(0)}\|_2 \le k\varepsilon\|\tilde{w}_{k+1}\|_2$, $\|f_{k+1}^{(1)}\|_2 \le$*

$k\varepsilon\|\tilde{l}_{k+1}^{(0)}\|_2$, *and* $\|f_k^{(S)}\|_2 \le \varepsilon\|\tilde{l}_{k+1}^{(1)}\|_2$ *for some* $\varepsilon < (1/2)\nu\left(1-\left((1/4)\sqrt{(k+1)^7}\kappa_{\max,k}\right)\nu^2\right)$ *with* $\nu = 1/\left(\sqrt{(k+1)^5}+1+4(k+1)\right)$.

Define the two non-negative sequences $\{\delta_k\}$ *and* $\{\zeta_k\}$ *by* $\delta_1^2 := 4\varepsilon^2/(1-\varepsilon)$ *and*

$$\delta_{k+1}^2 := \delta_k^2 + 2\left(\frac{\varepsilon\min\{\sqrt{k},1+\zeta_k\}\left(k+1+\min\{\sqrt{k},1+\zeta_k\}k(k\varepsilon)^\ell\kappa_k\right)}{1-(k(\ell+1)+2)\varepsilon}\right)^2 + \frac{4\varepsilon^2}{1-\varepsilon} \quad \textit{and}$$

$$\zeta_k := \frac{\delta_k}{\sqrt{2}(1-\delta_k)} \quad \textit{for } k = 1,2,\ldots. \tag{10}$$

Then D_k *is positive definite and for* $* \in \{2,F\}$, D_k *and its Cholesky factor* C_k *satisfy*

$$\|D_k - I_k\|_* \le \delta_k \le \sqrt{k^5}\left(2+(k\varepsilon)^\ell\kappa_{\max,k-1}\right)\cdot\frac{\varepsilon}{1-(k(\ell+3)+2)\varepsilon}, \ \textit{with } \kappa_{\max,0}=0 \tag{11}$$

$$\|C_k - I_k\|_* \le \zeta_k \le \sqrt{k^5/2}\left(2+(k\varepsilon)^\ell\kappa_{\max,k-1}\right)\cdot\frac{\varepsilon}{1-(\sqrt{k^5}(2+(k\varepsilon)^\ell\kappa_{\max,k-1})+k(\ell+1)+2)\varepsilon}. \tag{12}$$

For the proof we need the following theorem.

Theorem 3 *[41, Theorem 1.4] Let* A *be an* $n\times n$ *positive-definite matrix with the Cholesky factorization* $A = C^HC$. *If* ΔA *is a* $n\times n$ *Hermitian matrix satisfying*

$$\|A^{-1}\|_2\|\Delta A\|_F < 1,$$

then there is a unique Cholesky factorization

$$A+\Delta A = (C+\Delta C)(C+\Delta C)^H,$$

and

$$\frac{\|\Delta C\|_F}{\|C\|_2} \le \frac{\kappa(A)\epsilon}{\sqrt{2}\,(1-\kappa(A)\epsilon)},$$

where $\epsilon = \frac{\|\Delta A\|_F}{\|A\|_2}$ *and* $\kappa(A) = \|A\|_2\|A^{-1}\|_2$.

Proof (of Corollary 1)

The right inequalities in (11) and (12) on δ_k and ζ_k are obtained by using $\min\{\sqrt{k},1+\zeta_k\} \le \sqrt{k}$ in the definition of $\{\delta_k\}$. We have

$$\begin{aligned}
\delta_k^2 &= \delta_1^2 + \sum_{j=1}^{k-1}(\delta_{j+1}^2-\delta_j^2)\\
&\le \frac{4\varepsilon^2}{1-\varepsilon} + \sum_{j=1}^{k-1}\frac{4\varepsilon^2}{1-\varepsilon} + 2\left(\frac{\varepsilon}{1-(k(\ell+1)+2)\varepsilon}\sqrt{j}(k+1+\sqrt{j}k(k\varepsilon)^\ell\kappa_j)\right)^2\\
&\le \frac{4k\varepsilon^2}{1-\varepsilon} + 2\left(\frac{\varepsilon}{1-(k(\ell+1)+2)\varepsilon}\right)^2\left(k+1+\sqrt{k^3}(k\varepsilon)^\ell\kappa_{\max,k-1}\right)^2\sum_{j=1}^{k-1}j\\
&\le \left(\frac{2\sqrt{k}\varepsilon}{1-(k(\ell+1)+2)\varepsilon}\right)^2 + \left(\frac{\varepsilon}{1-(k(\ell+1)+2)\varepsilon}\right)^2\left(k+1+\sqrt{k^3}(k\varepsilon)^\ell\kappa_{\max,k-1}\right)^2(k^2-k)\\
&\le \left(\frac{\varepsilon}{1-(k(\ell+1)+2)\varepsilon}\left(2+(k\varepsilon)^\ell\kappa_{\max,k-1}\right)\sqrt{k^5}\right)^2.
\end{aligned}$$

In the last step we have used that omitting the first term is more than compensated by also replacing $k^2 - k$ by k^2. This proves the right inequality of (11). Note that by the assumed bounds on ε, we have $\delta_i < 1$ for all $i = 1, \dots, k+1$. Hence, from (10) we obtain

$$\zeta_k = \frac{\delta_k}{\sqrt{2}(1-\delta_k)} \leq \frac{\sqrt{k^5}\left(2 + (k\varepsilon)^\ell \kappa_{\max,k-1}\right)}{\sqrt{2}} \cdot \frac{\varepsilon}{1 - (\sqrt{k^5}(2k(k\varepsilon)^\ell \kappa_{\max,k-1}) + k(\ell+1) + 2)\varepsilon}$$

proving the right inequality of (12).

The proof of the left inequalities in (11) and(12) is by induction over k. For $k = 1$ we have $D_1 - I_1 = \tilde{v}_1^H \tilde{v}_1 - 1 \leq (1+\varepsilon)^2 - 1 \leq 2\varepsilon/\sqrt{1-\varepsilon} = \delta_1$. Moreover, $C_1 - I_1 = \sqrt{\|\tilde{v}_1\|_2^2} - 1 \leq \varepsilon \leq \zeta_1$, i.e., the assertion holds for $k = 1$. Now, suppose that $\|D_k - I_k\|_F \leq \delta_k$ and $\|C_k - I_k\|_F \leq \zeta_k$ holds for some positive integer k. We will show the assertion for $k+1$. From Step 5 of Algorithm 2, we have

$$\|D_{k+1} - I_{k+1}\|_F^2 = \|D_k - I_k\|_F^2 + 2\|\tilde{V}_k^H \tilde{v}_{k+1}\|_2^2 + (\|\tilde{v}_{k+1}\|_2^2 - 1)^2.$$

Hence, it follows from the induction hypothesis and Theorem 1, respectively Theorem 2, that

$$\|D_{k+1} - I_{k+1}\|_F^2 \leq \delta_k^2 + 2\left(\frac{\varepsilon}{1 - k(\ell+1)\varepsilon}\|\tilde{V}_k\|_2 \left(k + 1 + k(k\varepsilon)^\ell \|\tilde{V}_k\|_2 \kappa_k\right)\right)^2 + (\|\tilde{v}_{k+1}\|_2^2 - 1)^2. \quad (13)$$

Using the assumption on $\|f_{k+1}^{(S)}\|_2$ we obtain for the rightmost term

$$(\|\tilde{v}_{k+1}\|_2^2 - 1)^2 \leq ((1+\varepsilon)^2 - 1)^2 \leq (2\varepsilon(1 + \tfrac{1}{2}\varepsilon))^2 \leq \frac{4\varepsilon^2}{1-\varepsilon}.$$

For $\|\tilde{V}_k\|_2$ we find

$$\|\tilde{V}_k\|_2 = \|C_k\|_2 \leq \|I_k\|_2 + \|C_k - I_k\|_2 \leq 1 + \zeta_k,$$

because the matrix $\tilde{V}_k C_k^{-1} = \hat{V}_k$ has orthonormal columns. Together with Lemma 1 we have $\|\tilde{V}_k\|_2 \leq \min\{\sqrt{k}(1+\varepsilon), 1+\zeta_k\} \leq (1+\varepsilon)\min\{\sqrt{k}, 1+\zeta_k\}$. Substitution into (13) yields

$$\begin{aligned}
\|D_{k+1} - I_{k+1}\|_F^2 &\leq \delta_k^2 + 2\left(\frac{\varepsilon}{1 - k(\ell+1)\varepsilon}(1+\varepsilon)\min\{\sqrt{k}, 1+\zeta_k\} \cdot \right. \\
&\qquad \left. \left(k + 1 + (1+\varepsilon)\min\{\sqrt{k}, 1+\zeta_k\} k(k\varepsilon)^\ell \kappa_k\right)\right)^2 + \frac{4\varepsilon^2}{1-\varepsilon} \\
&\leq \delta_k^2 + 2\left(\frac{\varepsilon}{1 - (k(\ell+1)+2)\varepsilon}\min\{\sqrt{k}, 1+\zeta_k\} \cdot \right. \\
&\qquad \left. \left(k + 1 + \min\{\sqrt{k}, 1+\zeta_k\} k(k\varepsilon)^\ell \kappa_k\right)\right)^2 + \frac{4\varepsilon^2}{1-\varepsilon} = \delta_{k+1}^2.
\end{aligned}$$

In order to treat the Cholesky factor, we consider Theorem 3 with $A = I_{k+1}$, $\Delta A = D_{k+1} - I_{k+1}$, $C = I_{k+1}$, and $\Delta C = C_{k+1} - I_{k+1}$. Since

$$\|I_{k+1}^{-1}\|_2 \|D_{k+1} - I_{k+1}\|_F < 1 \cdot \delta_{k+1} < 1$$

by the assumed bound on ε, Theorem 3 is applicable and yields

$$\|C_{k+1} - I_{k+1}\|_F \leq \frac{\|D_{k+1} - I_{k+1}\|_F}{\sqrt{2}(1 - \|D_{k+1} - I_{k+1}\|_F)} \leq \frac{\delta_{k+1}}{\sqrt{2}(1-\delta_{k+1})} = \zeta_{k+1}.$$

With that the induction proof is completed. Moreover, the assumed bound on ε implies $\delta_k < 1$ and hence ensures non-negativity of $\{\zeta_k\}$.

Finally, using the norm property $\|A\|_2 \leq \|A\|_F$ (cf. [14]) concludes the proof of (11), (12). □

In (11), (12) we have presented two bounds each for $\|D_k - I\|_2$ and $\|C_k - I\|_2$. As before (in theorems 1 and 2) the first ones are sharper, whereas the second ones are more concise. The sharper bounds, being recursively defined, are useful in practice, but do not convey much theoretical insight. Both concise bounds are explicit and are of the form $\alpha_1\varepsilon/(1-\alpha_2\varepsilon)$. So, they are useful if $\max\{\alpha_1, \alpha_2\} \ll \varepsilon^{-1}$. This is the case whenever $\varepsilon^{\ell-2}\kappa_{\max,k}$ is not large. Then, D_k and C_k differ from the identity by order ε. This means that our algorithms generate an almost orthonormal basis $\tilde{V}_k$.

Concluding this section, we show a relation between algorithms 2 and 3, more precisely, that Algorithm 3 is a special case of Algorithm 2.

Theorem 4 *The results of Algorithm 1 using Algorithm 2 and Algorithm 1 using Algorithm 3 coincide if the perturbations are related by*

$$f_{k+1}^{(M,no)} = f_{k+1}^{(M,re)} + f_{k+1}^{(0,re)}, \quad f_{k+1}^{(0,no)} = f_{k+1}^{(1,re)}, \quad f_{k+1}^{(S,no)} = f_{k+1}^{(S,re)} \quad \text{for all} \quad k = 1, 2, \ldots. \tag{14}$$

Here, $f_{k+1}^{(M,no)}$, $f_{k+1}^{(0,no)}$, and $f_{k+1}^{(S,no)}$ are the perturbations[1] *occuring in Algorithm 1 using Algorithm 2 whereas $f_{k+1}^{(M,re)}$, $f_{k+1}^{(0,re)}$, $f_{k+1}^{(1,re)}$, and $f_{k+1}^{(S,re)}$ are the perturbations occurring in Algorithm 1 using Algorithm 3.*

Proof The k-th iteration of Algorithm 1 using Algorithm 2 generates

$$\begin{aligned}
\tilde{h}_{1:k,k} &= D_k^{-1}\tilde{V}_k^H(A\tilde{v}_k - f_{k+1}^{(M,no)}),\\
\tilde{h}_{k+1,k} &= \|A\tilde{v}_k - f_{k+1}^{(M,no)} - \tilde{V}_k\tilde{h}_{1:k,k} - f_{k+1}^{(0,no)}\|_2,\\
\tilde{v}_{k+1}\tilde{h}_{k+1,k} &= A\tilde{v}_k - f_{k+1}^{(M,no)} - \tilde{V}_k\tilde{h}_{1:k,k} - f_{k+1}^{(0,no)} - f_{k+1}^{(S,no)},
\end{aligned}$$

whereas the k-th iteration of Algorithm 1 using Algorithm 3 yields

$$\begin{aligned}
\tilde{h}_{1:k,k} &= s^{(0)} + D_k^{-1}\tilde{V}_k^H(A\tilde{v}_k - f_{k+1}^{(M,re)} - \tilde{V}_k^H s^{(0)} - f_{k+1}^{(0,re)})\\
\tilde{h}_{k+1,k} &= \|A\tilde{v}_k - f_{k+1}^{(M,re)} - \tilde{V}_k\tilde{h}_{1:k,k} - f_{k+1}^{(0,re)} - f_{k+1}^{(1,re)}\|_2,\\
\tilde{v}_{k+1}\tilde{h}_{k+1,k} &= A\tilde{v}_k - f_{k+1}^{(M,re)} - \tilde{V}_k\tilde{h}_{1:k,k} - f_{k+1}^{(0,re)} - f_{k+1}^{(1,re)} - f_{k+1}^{(S,re)}.
\end{aligned}$$

Both sets of results coincide if (14) holds. □

Thus, in the following analysis we only consider Algorithm 1 using Algorithm 2. Due to the fact that reorthogonalization requires extra scalar products and is only necessary for large $\kappa_{\max,k-1}$, we propose to only carry out the reorthogonalization step if $\kappa_{\max,k-1}$ exceeds a certain threshold. Also this hybrid technique fits in the framework of Algorithm 2.

Finally, we recall that next to $\tilde{V}_k$ we have also implicit access to an orthonormal basis $\hat{V}_k$ of the same subspace, see (3).

4 Krylov-like relations

The classical Arnoldi method produces an orthonormal basis and a Hessenberg matrix satisfying an Arnoldi relation (1). Due to the perturbations, this is no longer true for the results of Algorithm 1. Here we want to derive a relation between A, $\tilde{V}_k$, and $\tilde{H}_k$ that is close to an Arnoldi relation in order to admit a backward error analysis of our method. From Step 3 of Algorithm 1 and steps 2 and 4 of Algorithm 2 for $j = 1, \ldots, k$ we have

$$\begin{aligned}
A\tilde{v}_j &= \tilde{w}_{j+1} + f_{j+1}^{(M)}\\
&= \tilde{V}_j\tilde{h}_{1:j,j} + \tilde{l}_{j+1} + f_{j+1}^{(0)} + f_{j+1}^{(M)}\\
&= \tilde{V}_j\tilde{h}_{1:j,j} + \|\tilde{l}_{j+1}\|\,\tilde{v}_{j+1} + f_{j+1}^{(S)} + f_{j+1}^{(0)} + f_{j+1}^{(M)}\\
&= \tilde{V}_{k+1}\begin{bmatrix}\tilde{h}_{1:j,j}\\ \|\tilde{l}_{j+1}\|\end{bmatrix} + f_{j+1}^{(S)} + f_{j+1}^{(0)} + f_{j+1}^{(M)}.
\end{aligned}$$

[1] where the superscripts (re) and (no) indicate that reorthogonalization was used (re)/was not used (no)

Hence, after k steps we obtain the relation

$$A\tilde{V}_k = \tilde{V}_k\tilde{H}_k + \tilde{v}_{k+1}\tilde{h}_{k+1,k}e_k^T + F_k. \tag{15}$$

Here $F_k = [f_2^{(M)} + f_2^{(0)} + f_2^{(S)}, \ldots, f_{k+1}^{(M)} + f_{k+1}^{(0)} + f_{k+1}^{(S)}]$ is the matrix consisting of the individual perturbations occurring throughout the Algorithms 1 and 2. In order to bound F_k, we assume that the perturbations are bounded by $\|f_{j+1}^{(0)}\|_2 \leq \varepsilon\|\tilde{w}_{j+1}\|_2$, $\|f_{j+1}^{(S)}\|_2 \leq \varepsilon\tilde{h}_{j+1,j}$ (as in Theorem 1), and $\|f_{j+1}^{(M)}\|_2 \leq \varepsilon\|A\|_2$. The latter assumption is reasonable since $f_{j+1}^{(M)}$ can be interpreted as error in the matrix-vector-product which is proportional to the norms of the factors A and $\tilde{v}_i$. Then each column of F can be bounded by

$$\|Fe_j\|_2 \leq \|f_{j+1}^{(M)}\|_2 + \|f_{j+1}^{(0)}\|_2 + \|f_{j+1}^{(S)}\|_2 \leq \varepsilon(\|A\|_2 + \|\tilde{w}_{j+1}\|_2 + \tilde{h}_{j+1,j}) \leq 3\varepsilon\|A\|_2.$$

Thus,

$$\|F_k\|_2 \leq \|F_k\|_F \leq 3\sqrt{k}\|A\|_2\varepsilon. \tag{16}$$

Since $\tilde{H}_k$ is a Hessenberg matrix and the norm of F_k is small, (15) is close to an Arnoldi relation. Consequently, $\tilde{\mathcal{K}}_k = \mathrm{img}(\tilde{V}_k)$ is close to a Krylov subspace of A. However, (15) fails to be an exact Arnoldi relation for two reasons: (i) the presence of F_k and (ii) the non-orthogonality of $\tilde{V}_k$.

We address the latter issue by switching from $\tilde{V}_k$ to $\hat{V}_k = \tilde{V}_k C_k^{-1}$, the known orthonormal basis of $\tilde{\mathcal{K}}_k$. Post-multiplication of Equation (15) by C_k^{-1} yields

$$\begin{aligned} A\tilde{V}_k C_k^{-1} &= (\tilde{V}_k\tilde{H}_k + \tilde{v}_{k+1}\tilde{h}_{k+1,k}e_k^T)C_k^{-1} + F_kC_k^{-1} \\ &= [\tilde{V}_k, \tilde{v}_{k+1}]C_{k+1}^{-1}C_{k+1}\begin{bmatrix}\tilde{H}_k \\ \tilde{h}_{k+1,k}e_k^T\end{bmatrix}C_k^{-1} + F_kC_k^{-1}. \end{aligned}$$

Exploiting the upper triangular structure of C_k we define a Hessenberg matrix $\hat{H}_k$ and a scalar $\hat{h}_{k+1,k}$ by

$$\begin{bmatrix}\hat{H}_k \\ \hat{h}_{k+1,k}e_k^T\end{bmatrix} := C_{k+1}\begin{bmatrix}\tilde{H}_k \\ \tilde{h}_{k+1,k}e_k^T\end{bmatrix}C_k^{-1}.$$

Introducing further $\hat{F}_k := F_kC_k^{-1}$ and using $[\hat{V}_k, \hat{v}_{k+1}] = \hat{V}_{k+1} = \tilde{V}_{k+1}C_{k+1}^{-1} = [\tilde{V}_k, \tilde{v}_{k+1}]C_{k+1}^{-1}$ results in

$$A\hat{V}_k = \hat{V}_k\hat{H}_k + \hat{v}_{k+1}\hat{h}_{k+1,k}e_k^T + \hat{F}_k. \tag{17}$$

This would be an Arnoldi relation if $\hat{F}_k$ was not present. Therefore we now remove $\hat{F}$ by interpreting it as backward error of A. Indeed, relation (17) is equivalent to

$$(A + E_k)\hat{V}_k = \hat{V}_k\hat{H}_k + \hat{v}_{k+1}\hat{h}_{k+1,k}e_k^T, \tag{18}$$

whenever $E_k \in \mathbb{C}^{n\times n}$ fulfills

$$E_k\hat{V}_k = -\hat{F}_k. \tag{19}$$

E.g., E_k may be chosen as $E_k = -\hat{F}_k\hat{V}_k^H$ as in [36, 39, 40]. In this way we have managed to arrive at a correct Arnoldi relation (18). In the remainder of this section we consider Hermitian matrices A. In this case choosing $E_k = -\hat{F}_k\hat{V}_k^H$ is not satisfying, because this choice of E_k is in general not Hermitian which does not seem appropriate when A is. Unfortunately, the following lemma rules out the existence of a Hermitian matrix E_k satisfying (19).

Lemma 2 *Let $V \in \mathbb{C}^{n\times k}$ have orthonormal columns and $F \in \mathbb{C}^{n\times k}$. Then there is a Hermitian E with $EV = F$ if and only if V^HF is Hermitian.*

Proof The proof is simple and, for $k = 1$, is contained in [27]. We give it for completeness. Let E be any matrix such that $EV = F$. Now, if E is Hermitian, then so is $V^HEV = V^HF$. Hence, if V^HF is not Hermitian, then there is no Hermitian E with $EV = F$. On the other hand, let V^HF be Hermitian, i.e., $V^HF = F^HV$. Then $E = FV^H + V^HF - VF^HVV^H$ is Hermitian and $EV = F$. □

It turns out that a solution to this dilemma is to replace the (non-Hermitian) Hessenberg matrix $\hat{H}_k$ with a Hermitian matrix B_k. In fact any Hermitian matrix will do. To this end we relax the concept of the Arnoldi relation by allowing non-Hessenberg H_k. We say that $A \in \mathbb{C}^{n\times n}, V_{k+1} = [V_k, v_{k+1}] \in \mathbb{C}^{n\times k+1}, H_k \in \mathbb{C}^{k\times k}$ satisfy a *Krylov relation* (as introduced by Stewart in [39]), if (1) holds and V_{k+1} has orthonormal columns. We have the following result.

Theorem 5 *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian, let $\hat{V}_{k+1} = [\hat{V}_k, \hat{v}_{k+1}] \in \mathbb{C}^{n\times k+1}$ have orthonormal columns, and suppose that $\hat{H}_k \in \mathbb{C}^{k\times k}$, $\hat{h}_{k+1,k} \in \mathbb{C}$ and $\hat{F}_k \in \mathbb{C}^{n\times k}$ are such that* (17) *holds.*

Then for every Hermitian matrix $B_k \in \mathbb{C}^{k\times k}$ there exists a Hermitian matrix $E_k \in \mathbb{C}^{n\times n}$ such that

$$(A + E_k)\hat{V}_k = \hat{V}_k B_k + \hat{v}_{k+1}\hat{h}_{k+1,k}e_{k+1}^T \tag{20}$$

is a Krylov relation and for $* \in \{2, F\}$ E_k is bounded by

$$\alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \le \|E_k\|_* \le \|B_k - \hat{S}_k\|_* + \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \le \|B_k - \hat{S}_k\|_* + \alpha_*\|\hat{F}_k\|_* \tag{21}$$

with $\hat{S}_k := \hat{V}_k^H A\hat{V}_k$, $P_{\tilde{\mathcal{K}}_k}$ as in (3)*, and $\alpha_2 = 1, \alpha_F = \sqrt{2}$.*

The lower bound holds for any Hermitian $E_k \in \mathbb{C}^{n\times n}$ for which a $B_k \in \mathbb{C}^{k\times k}$ exists such that (20) *holds.*

Proof We prove the lower bound first. Assume that E_k and B_k are such that (20) holds and let $\hat{V}_\perp \in \mathbb{C}^{n\times n-k}$ be any matrix such that $[\hat{V}_k, \hat{V}_\perp]$ is unitary. Pre-multiplicating (20) by $\hat{V}_\perp$ gives $\hat{V}_\perp^H(A + E_k)\hat{V}_k = 0$, i.e., $\hat{V}_\perp^H A\hat{V}_k = -\hat{V}_\perp^H E_k\hat{V}_k$. Pre-multiplicating (17) by $\hat{V}_\perp$ gives $\hat{V}_\perp^H A\hat{V}_k = \hat{V}_\perp^H\hat{F}_k$. Together, $\hat{V}_\perp^H E_k\hat{V}_k = -\hat{V}_\perp^H\hat{F}_k$. Thus, since E_k is Hermitian, it must be of the form

$$E_k = [\hat{V}_k, \hat{V}_\perp]\begin{bmatrix} E_{11} & -\hat{F}_k^H\hat{V}_\perp \\ -\hat{V}_\perp^H\hat{F}_k & E_{22}\end{bmatrix}[\hat{V}_k, \hat{V}_\perp]^H, \tag{22}$$

where E_{11}, E_{22} are still undetermined. Let E_* be the matrix that is obtained by setting E_{11} and E_{22} in (22) to zero. Then $\|E_k\|_* \ge \|E_*\|_* = \alpha_*\|\hat{V}_\perp^H\hat{F}_k\|_* = \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_*$, proving the lower bound on $\|E_k\|_*$ in (21).

For the upper bound, let $B_k \in \mathbb{C}^{k\times k}$ be Hermitian. Pre-multiplying equation (17) with $\hat{V}_k^H$ shows that $\hat{S}_k$, $\hat{H}_k$, and $\hat{F}_k$ are related via

$$\hat{S}_k = \hat{H}_k + \hat{V}_k^H\hat{F}_k. \tag{23}$$

The proof is constructive: choose $E_k = E_* + \hat{V}_k(B_k - \hat{S}_k)\hat{V}_k^H$ and note that with A and B_k also $\hat{S}_k$ and E_k are Hermitian. Then

$$\begin{aligned}(A + E_k)\hat{V}_k &= A\hat{V}_k + E_*\hat{V}_k + \hat{V}_k(B_k - \hat{S}_k)\hat{V}_k^H\hat{V}_k \\ &= (\hat{V}_k\hat{H}_k + \hat{v}_{k+1,k}e_k^T + \hat{F}_k) - \hat{V}_\perp\hat{V}_\perp^H\hat{F}_k + \hat{V}_k(B_k - (\hat{H}_k + \hat{V}_k^H\hat{F}_k)) \\ &= \hat{V}_k B_k + \hat{v}_{k+1,k}e_k^T,\end{aligned}$$

where we have used $\hat{V}_k\hat{V}_k^H + \hat{V}_\perp\hat{V}_\perp^H = I$. This proves (20). For the norm of E_k we have

$$\|E_k\|_* = \|E_* + \hat{V}_k(B_k - \hat{S}_k)\hat{V}_k^H\|_* \le \|E_*\|_* + \|B_k - \hat{S}_k\|_* \le \alpha_*\|\hat{F}_k\|_* + \|B_k - \hat{S}_k\|_*$$

which concludes the proof. □

It follows from equation (20) that $\tilde{\mathcal{K}}_k$ is a Krylov subspace for the Hermitian matrix $A + E_k$. Note that the perturbation E_k depends on k. Hence, although $\tilde{\mathcal{K}}_j$ is a Krylov subspace of $A + E_j$, in general it is not a Krylov subspace of $A + E_k$ for $j < k$.

Theorem 5 opens up some freedom in the choice of B_k. Which matrix should be used in practice? One criterion could be that for positive definite A also B_k should be definite. By left multiplication of (20) by $\hat{V}_k^H$ we see that $B_k = \hat{V}_k^H(A + E_k)\hat{V}_k$, implying that B_k is guaranteed to be definite, whenever $A + E_k$ is. Hence, $\|E_k\|$ should be small. Corollary 2 below will provide bounds on the norm of E_k corresponding to the following choices of B_k. In light of

B_k	α_{2,B_k}	β_{2,B_k}	α_{F,B_k}	β_{F,B_k}
$\hat{S}_k$	1	0	$\sqrt{2}$	0
$\frac{1}{2}(\hat{H}_k + \hat{H}_k^H)$	2	0	$1+\sqrt{2}$	0
$\frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H)$	2	1	$1+\sqrt{2}$	1
$T_{\hat{S}_k}$	$1+\sqrt{2k}$	0	$2\sqrt{2}$	0
$T_{\hat{H}_k}$	$2+\sqrt{k}$	0	$2+\sqrt{2}$	0
$T_{\tilde{H}_k}$	$2+\sqrt{k}$	$1+\sqrt{k}$	$2+\sqrt{2}$	2.

Table 1 Coefficients of the error bound (25) for different choices of B_k

our bound (21), a perfect choice is $B_k = \hat{S}_k = \hat{V}_k^H A \hat{V}_k$, as it minimizes this bound on $\|E_k\|_*$. Unfortunately, constructing $\hat{S}_k$ requires matrix-vector-products with A, which are not possible without perturbations. In the special case when $f_k^{(M)} = 0$ for all k, $\hat{S}_k$ can be computed as $\hat{S}_k = C_k^{-H} \tilde{S}_k C_k^{-1}$, where $\tilde{S}_k := \tilde{V}_k^H A \tilde{V}_k$ can be updated along V and $\tilde{H}_k$ in Algorithm 1 by

$$\tilde{S}_k = \begin{bmatrix} \tilde{S}_{k-1} & \tilde{V}_{k-1}^H \tilde{w}_{k+1} \\ \tilde{w}_{k+1}^H \tilde{V}_{k-1} & \tilde{v}_k^H \tilde{w}_{k+1} \end{bmatrix}.$$

When the matrix-vector-multiplication is inexact, other choices for B_k have to be found. Considering the two terms in the bound (21), a B_k is acceptable, whenever $\|B_k - \hat{S}_k\|_*$ is not much larger than $\|\hat{F}_k\|_*$. By (23) $\hat{H}_k$ is close to $\hat{S}_k$. Since $\hat{H}_k$ itself is non-Hermitian, we propose to use its Hermitian part, $\frac{1}{2}(\hat{H}_k + \hat{H}_k^H)$.

In some situations B_k may be required to be tridiagonal. Possible reasons for this restriction may be theoretical (it implies that $\tilde{\mathcal{K}}_j$ is a Krylov subspace for $A+E_k$ for $j \le k$) or just practical (computation of eigenvalues for tridiagonal matrices is faster and more accurate than for general Hermitian matrices [9]). For these situations we will analyze the choices $B_k = T_{\hat{S}_k}$, $B_k = T_{\hat{H}_k}$, where T_H denotes the tridiagonal part of the Hermitian part of H, i.e.,

$$(T_H)_{i,j} := \begin{cases} \frac{1}{2}(h_{i,j} + \overline{h_{j,i}})/2, & \text{for} \quad |i-j| \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

Finally, we will also look at $B_k = \frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H)$ and $B_k = T_{\tilde{H}_k}$. The reasoning behind these choices is that without perturbations, Algorithm 1 would reduce to the Lanczos method and $\tilde{H}_k$ would be Hermitian and tridiagonal. With perturbations, $\tilde{H}_k$ is neither, but should still be close. Thus, its Hermitian or Hermitian tridiagonal parts should be good approximations of the original tridiagonal matrix. In particular, the choice $B_k = T_{\tilde{H}_k}$ is used in the ARPACK [25] for Hermitian A. We obtain the following bounds on E_k.

Corollary 2 *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian and let $\hat{V}_{k+1} = [\hat{V}_k, \hat{v}_{k+1}] \in \mathbb{C}^{n\times k+1}$ have orthonormal columns. Suppose that $\hat{H}_k \in \mathbb{C}^{k\times k}$ is Hessenberg, $\hat{h}_{k+1,k} \in \mathbb{C}$ and $\hat{F}_k \in \mathbb{C}^{n\times k}$ is such that (17) holds. Let $\hat{S}_k := \hat{V}_k^H A \hat{V}_k$ and $\underline{\hat{H}}_k := [\hat{H}_k^T, \hat{h}_{k+1,k} e_k]^T$. Let $C_k \in \mathbb{C}^{k\times k}$ and $C_{k+1} \in \mathbb{C}^{k+1\times k+1}$ be invertible upper triangular matrices such that $\|C_k - I_k\|_2 \le \zeta_k$ and $\|C_{k+1} - I_{k+1}\|_2 \le \zeta_{k+1} < 1$. Moreover, define $\tilde{H}_k := [I_k, 0] C_{k+1}^{-1} \underline{\hat{H}}_k C_k$ and $T_{\hat{S}_k}, T_{\hat{H}_k}, T_{\tilde{H}_k}$ as in (24).*

Then, for $B_k \in \{S_k, T_{\hat{S}_k}, \frac{1}{2}(\hat{H}_k + \hat{H}_k^H), T_{\hat{H}_k}, \frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H), T_{\tilde{H}_k}\}$ there exists a Hermitian matrix $E_k \in \mathbb{C}^{n\times n}$ such that (20) is a Krylov relation and for $* \in \{2, F\}$ $\|E_k\|_*$ is bounded by

$$\|E_k\|_* \le \alpha_{*,B_k} \|\hat{F}_k\|_* + \beta_{*,B_k} \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}} \tag{25}$$

with constants α_{*,B_k}, β_{*,B_k} given in Table 1.

For the proof we need the following perturbation lemma, see, e.g., [14, Lemma 2.3.3].

Lemma 3 *(Perturbation Lemma) Let $* \in \{2, F\}$ and $C \in \mathbb{C}^{n\times n}$ with $\|C - I\|_* \le \zeta < 1$. Then i) C is invertible, ii) $C^{-1} = \sum_{i=0}^{\infty} (I - C)^i$, iii) $\|C^{-1} - I\|_* \le \frac{\zeta}{1-\zeta}$, and iv) $\|C^{-1}\|_* \le \frac{1}{1-\zeta}$.*

Proof (of Corollary 2) The error bounds of the different choices of B_k are proved separately.
• *Case* $B_k = \hat{S}_k$: This case follows directly from (21).
• *Case* $B_k = \frac{1}{2}(\hat{H}_k + \hat{H}_k^H)$: Using relation (23) and that $\hat{S}_k$ is Hermitian, we have

$$\tfrac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k = \tfrac{1}{2}(\hat{S}_k - \hat{V}_k^H\hat{F}_k + \hat{S}_k^H - (\hat{V}_k^H\hat{F}_k)^H) - \hat{S}_k = -\tfrac{1}{2}(\hat{V}_k^H\hat{F}_k + \hat{F}_k^H\hat{V}_k)$$

implying that

$$\|\tfrac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k\|_* = \tfrac{1}{2}\|\hat{V}_k^H\hat{F}_k + \hat{F}_k^H\hat{V}_k\|_* \le \|\hat{V}_k^H\hat{F}\|_* \le \|P_{\tilde{\mathcal{K}}_k}\hat{F}\|_*. \tag{26}$$

Hence by (21), the backward error corresponding to $B_k = \frac{1}{2}(\hat{H}_k + \hat{H}_k^H)$ is bounded by

$$\begin{aligned}\|E_k\|_* \le \|\tfrac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k\|_* + \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \le \|P_{\tilde{\mathcal{K}}_k}\hat{F}_k\|_* + \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \\ \le \alpha_{*,\frac{1}{2}(\hat{H}_k+\hat{H}_k^H)}\|\hat{F}_k\|_2\end{aligned}$$

with α_* as in Theorem 5.
• *Case* $B_k = T_{\hat{H}_k}$: For $B_k = T_{\hat{H}_k}$ we consider a splitting $\hat{H}_k = \hat{T} + \hat{U}$, where $\hat{T}$ is the tridiagonal part of $\hat{H}_k$ and $\hat{U} = \hat{H}_k - \hat{T}$ is strictly upper triangular. Then

$$T_{\hat{H}_k} = \tfrac{1}{2}(\hat{T} + \hat{T}^H) = \tfrac{1}{2}(\hat{H} + \hat{H}^H) - \tfrac{1}{2}(\hat{U} + \hat{U}^H). \tag{27}$$

From (23) we have

$$\hat{H}_k - \hat{H}_k^H = \hat{S}_k - \hat{V}_k^H\hat{F}_k - (\hat{S}_k^H - \hat{F}_k^H\hat{V}_k) = \hat{F}_k^H\hat{V}_k - \hat{V}_k^H\hat{F}_k. \tag{28}$$

Hence, we have

$$\|\hat{U}+\hat{U}^H\|_* \le \|\hat{U}+\hat{U}^H\|_F = \|\hat{U}-\hat{U}^H\|_F \le \|\hat{H}_k-\hat{H}_k^H\|_F = \|\hat{F}_k^H\hat{V}_k-\hat{V}_k^H\hat{F}_k\|_F \le 2\|\hat{V}_k^H\hat{F}_k\|_F.$$

Thus, together with (27) and (26) we have

$$\|T_{\hat{H}_k} - \hat{S}_k\|_* \le \|\tfrac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k\|_* + \tfrac{1}{2}\|\hat{U} + \hat{U}^H\|_* \le \|P_{\tilde{\mathcal{K}}_k}\hat{F}_k\|_* + \|P_{\tilde{\mathcal{K}}_k}\hat{F}_k\|_F.$$

Hence, with (21), the backward error E_k corresponding to $B_k = T_{\hat{H}_k}$ is bounded by

$$\begin{aligned}\|E_k\|_* \le \|T_{\hat{H}_k} - \hat{S}_k\|_* + \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \le \|P_{\tilde{\mathcal{K}}_k}\hat{F}_k\|_* + \|P_{\tilde{\mathcal{K}}_k}\hat{F}_k\|_F + \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \\ \le \alpha_{*,T_{\hat{H}_k}}\|\hat{F}_k\|_*.\end{aligned}$$

• *Case* $B_k = T_{\hat{S}_k}$: For $B_k = T_{\hat{S}_k}$ we consider a splitting $\hat{S}_k = T_{\hat{S}_k} + U + U^H$, where U denotes the upper triangular part of $\hat{S}_k - T_{\hat{S}_k}$ (note that the main and the first super diagonals of U are zero). Since $\hat{H}_k$ is a Hessenberg matrix it follows from (23) that $\hat{S}_k$ and $\hat{V}_k^H\hat{F}_k$ coincide below the first subdiagonal. Hence, $\|U^H\|_F \le \|\hat{V}_k^H\hat{F}_k\|_F$ and it follows

$$\|\hat{S}_k - T_{\hat{S}_k}\|_* \le \|\hat{S}_k - T_{\hat{S}_k}\|_F = \|U + U^H\|_F = \sqrt{2}\|U^H\|_F \le \sqrt{2}\|\hat{V}_k^H\hat{F}_k\|_F = \sqrt{2}\|P_{\tilde{\mathcal{K}}_k}\hat{F}_k\|_F$$

Thus, using (21), the backward error E_k corresponding to $B_k = T_{\hat{S}_k}$ is bounded by

$$\|E_k\|_* \le \|\hat{S}_k - T_{\hat{S}_k}\|_* + \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \le \sqrt{2}\|P_{\tilde{\mathcal{K}}_k}\hat{F}_k\|_F + \alpha_*\|(I - P_{\tilde{\mathcal{K}}_k})\hat{F}_k\|_* \le \alpha_{*,T_{\hat{S}_k}}\|\hat{F}_k\|_*.$$

• *Case* $B_k = \frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H)$: Define $\underline{\tilde{H}}_k := C_{k+1}^{-1}\underline{\hat{H}}_k C_k$ and note that $\tilde{H}_k$ consists of the top k rows of $\underline{\tilde{H}}_k$. Then we have $\underline{\tilde{H}}_k = (C_{k+1}^{-1} - I_{k+1} + I_{k+1})\underline{\hat{H}}_k(C_k - I_k + I_k)$ implying that

$$\underline{\tilde{H}}_k = \underline{\hat{H}}_k + (C_{k+1}^{-1} - I_{k+1})\underline{\hat{H}}_k + \underline{\hat{H}}_k(C_k - I_k) + (C_{k+1}^{-1} - I_{k+1})\underline{\hat{H}}_k(C_k - I_k).$$

With $\|C_k - I_k\|_2 = \zeta_k$ and (using Lemma 3 iii)) $\|C_{k+1}^{-1} - I_{k+1}\|_2 \leq \frac{\zeta_{k+1}}{1-\zeta_{k+1}}$, we have

$$\begin{aligned}\|\tilde{H}_k - \hat{H}_k\|_* &\leq \|\underline{\tilde{H}}_k - \underline{\hat{H}}_k\|_* \\ &\leq \|C_{k+1}^{-1} - I_{k+1}\|_2 \|\underline{\hat{H}}_k\|_* + \|\underline{\hat{H}}_k\|_* \|C_k - I_k\|_2 + \|C_{k+1}^{-1} - I_{k+1}\|_2 \|\underline{\hat{H}}_k\|_* \|C_k - I_k\|_2 \\ &\leq \|\underline{\hat{H}}_k\|_* \left(\frac{\zeta_{k+1}}{1-\zeta_{k+1}} + \zeta_k + \frac{\zeta_k \zeta_{k+1}}{1-\zeta_{k+1}} \right) = \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}}. \quad (29)\end{aligned}$$

Consequently, with (26) and $\|\frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H) - \frac{1}{2}(\hat{H}_k + \hat{H}_k^H)\|_* \leq \|\tilde{H}_k - \hat{H}_k\|_*$, we have

$$\begin{aligned}\|\tfrac{1}{2}(\tilde{H}_k + \tilde{H}_k^H) - \hat{S}_k\|_* &\leq \|\tfrac{1}{2}(\tilde{H}_k + \tilde{H}_k^H) - \tfrac{1}{2}(\hat{H}_k + \hat{H}_k^H)\|_* + \|\tfrac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k\|_* \\ &\leq \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}} + \|P_{\hat{\mathcal{K}}_k} \hat{F}_k\|_*. \quad (30)\end{aligned}$$

Thus, using (21), the backward error E_k corresponding to $B_k = \frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H)$ is bounded by

$$\begin{aligned}&\|E_k\|_* \leq \|\hat{S}_k - \tfrac{1}{2}(\tilde{H}_k + \tilde{H}_k^H)\|_* + \alpha_* \|(I - P_{\hat{\mathcal{K}}_k})\hat{F}_k\|_* \\ &\leq \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}} + \|P_{\hat{\mathcal{K}}_k} \hat{F}_k\|_* + \alpha_* \|(I - P_{\hat{\mathcal{K}}_k})\hat{F}_k\|_* \leq \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}} + \alpha_{*, \frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H)} \|\hat{F}_k\|_*.\end{aligned}$$

•*Case* $B_k = T_{\tilde{H}_k}$: Using a splitting $\tilde{H}_k = \tilde{T} + \tilde{U}$, where $\tilde{T}$ is the tridiagonal part of $\tilde{H}_k$ and $\tilde{U} = \tilde{H}_k - \tilde{T}$ is strictly upper triangular, leads to

$$T_{\tilde{H}_k} = \tfrac{1}{2}(\tilde{T} + \tilde{T}^H) = \tfrac{1}{2}(\tilde{H}_k + \tilde{H}_k^H) - \tfrac{1}{2}(\tilde{U} + \tilde{U}^H). \quad (31)$$

Using (28) and (29) we have

$$\|\tilde{H}_k - \tilde{H}_k^H\|_* \leq \|\hat{H}_k - \hat{H}_k^H\|_* + 2\|\tilde{H}_k - \hat{H}_k\|_* \leq \|\hat{F}_k^H \hat{V}_k - \hat{V}_k^H \hat{F}_k\|_* + 2\|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}}.$$

Hence, with $\gamma_2 := \sqrt{k}, \gamma_F := 1$ it follows that

$$\begin{aligned}&\|\tilde{U} + \tilde{U}^H\|_* \leq \|\tilde{U} + \tilde{U}^H\|_F = \|\tilde{U} - \tilde{U}^H\|_F \leq \|\tilde{H}_k - \tilde{H}_k^H\|_F \leq \gamma_* \|\tilde{H}_k - \tilde{H}_k^H\|_* \\ &\leq \gamma_* \left(\|\hat{F}_k^H \hat{V}_k - \hat{V}_k^H \hat{F}_k\|_* + 2\|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}} \right) \leq 2\gamma_* \|P_{\hat{\mathcal{K}}_k} \hat{F}_k\|_* + 2\gamma_* \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}}. \quad (32)\end{aligned}$$

Consequently, with (31), (30), and (32)

$$\begin{aligned}\|T_{\tilde{H}_k} - \hat{S}_k\|_* &\leq \|\tfrac{1}{2}(\tilde{H}_k + \tilde{H}_k^H) - \hat{S}_k\|_* + \tfrac{1}{2}\|\tilde{U} + \tilde{U}^H\|_* \\ &\leq \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}} + \|P_{\hat{\mathcal{K}}_k} \hat{F}_k\|_* + \gamma_* \|P_{\hat{\mathcal{K}}_k} \hat{F}_k\|_* + \gamma_* \|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}} \\ &= (1+\gamma_*)\|P_{\hat{\mathcal{K}}_k} \hat{F}_k\|_* + (1+\gamma_*)\|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}}.\end{aligned}$$

Hence, by (21), the backward error corresponding to $B_k = T_{\tilde{H}_k}$ is bounded by

$$\begin{aligned}\|E_k\|_* &\leq \|T_{\tilde{H}_k} - \hat{S}_k\|_* + \alpha_* \|(I - P_{\hat{\mathcal{K}}_k})\hat{F}_k\|_* \\ &\leq (1+\gamma_*)\|P_{\hat{\mathcal{K}}_k} \hat{F}_k\|_* + (1+\gamma_*)\|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}} + \alpha_* \|(I - P_{\hat{\mathcal{K}}_k})\hat{F}_k\|_* \\ &\leq (1+\gamma_*+\alpha_*)\|\hat{F}_k\|_* + (1+\gamma_*)\|\underline{\hat{H}}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1-\zeta_{k+1}}\end{aligned}$$

concluding the proof. □

Not surprisingly, the bound is best for $\hat{S}_k$, but it is not much worse for $\frac{1}{2}(\hat{H}_k + \hat{H}_k^H)$ making it a good candidate when $\hat{S}_k$ is not available. All the bounds for tridiagonal B_k involve a $\sqrt{k}$ factor. This stems from the use of the Frobenius norm and is likely to be an overestimation in our case. Moreover, it is noteworthy that for those B_k involving $\tilde{H}_k$ the bounds contain a term $\|\hat{H}_k\|_2\|C_k - I\|_2$.

Remark 1 Combining the bounds of Corollary 2 with those of Corollary 1 and (16), we can show that E_k can be bounded in terms of $\|A\|_2\varepsilon$. For example, for $B_k = T_{\hat{H}_k}$ we obtain

$$\begin{aligned}\|E_k\|_F \le (2+\sqrt{2})\|\hat{F}\|_F \le (2+\sqrt{2})\|F\|_F(1+\mathcal{O}(\varepsilon)) &\le 3(2+\sqrt{2})\sqrt{k}\|A\|_2\varepsilon + \mathcal{O}(\varepsilon^2)\\ &\le 11\sqrt{k}\|A\|_2\varepsilon + \mathcal{O}(\varepsilon^2)\end{aligned}$$

where we used $\|C_k^{-1}\|_2 = 1 + \mathcal{O}(\varepsilon)$ (by Corollary 1).

For $B_k = T_{\tilde{H}_k}$ we obtain (using $\|\hat{H}_k\|_2 \le \|A + E_k\|_2 = \|A\|_2(1 + \mathcal{O}(\varepsilon))$ and $\|C_k^{-1} - I_k\|_2 = \sqrt{2k^3}(1 + \varepsilon\kappa_{\max,k})\varepsilon + \mathcal{O}(\varepsilon^2)$ which follows from the coarse bound in Corollary 1)

$$\begin{aligned}\|E_k\|_F &\le (2+\sqrt{2})\|\hat{F}\|_F + 2\|\hat{H}_k\|_F\tfrac{\zeta_k+\zeta_{k+1}}{1-\zeta_{k+1}}\\ &\le (2+\sqrt{2})\|F\|_F(1+\mathcal{O}(\varepsilon)) + 2\sqrt{k}\|\hat{H}_k\|_2(2\sqrt{(k+1)^5/2}(2+(k+1)\varepsilon\kappa_{\max,k+1})\varepsilon + \mathcal{O}(\varepsilon^2))\\ &\le (2+\sqrt{2})3\sqrt{k}\|A\|_2\varepsilon + 2\sqrt{k}\|A\|_2 2\sqrt{(k+1)^5/2}(2+(k+1)\varepsilon\kappa_{\max,k+1})\varepsilon + \mathcal{O}(\varepsilon^2)\\ &\le \left(7(k+1)^3 + 3(k+1)^4\varepsilon\kappa_{\max,k+1}\right)\|A\|_2\varepsilon + \mathcal{O}(\varepsilon^2).\end{aligned}$$

Comparing these results suggests that $T_{\hat{H}_k}$ should be preferred over $T_{\tilde{H}_k}$. The numerical Example 3 below shows that this is indeed the case when κ is large. Otherwise these two choices of B_k actually perform rather similar.

5 Numerical results

In this section we present some numerical experiments that verify the previous theoretical results.

Example 1 Our first numerical example assesses the robustness of the orthogonalization variant compensated Gram Schmidt (ComGS) in comparison to the well known classical (CGS) and modified (MGS) Gram Schmidt schemes and their variants with reorthogonalization (CGSre, MGSre). In this test these schemes are used in an (inexact) QR factorization of a general $n \times m$ matrix A.

Algorithm 4 *QR factorization (inexact)*
Input: $A = [a_1, \ldots, a_m] \in \mathbb{C}^{n\times m}$, $\varepsilon \ge 0$
Output: $\tilde{V} \in \mathbb{C}^{n\times m}, \tilde{R} \in \mathbb{C}^{m\times m}$ upper triangular such that $\tilde{V}^H\tilde{V} \approx I, A \approx \tilde{V}\tilde{R}$
1: **for** $k = 0, \ldots m-1$ **do**
2: $\quad[\tilde{v}_{k+1}, \tilde{r}_{k+1,k+1}, \tilde{r}_{1:k,k+1}] = inexact\_orthonormalize(a_{k+1}, \tilde{V}_k, D_k, \varepsilon)$
3: **end for**

This algorithm consists exclusively of orthonormalizations. So any deficiency in the numerical results can be directly traced back to a weakness in the orthonormalization scheme.

For completeness we state here what we mean by inexact CGS and MGS with and without reorthogonalization. All variants obtain as

Input: $\tilde{w}_{k+1} \in \mathbb{C}^n$, $\tilde{V}_k \in \mathbb{C}^{n\times k}$, $\varepsilon \ge 0$ and return as
Output: $\tilde{v}_{k+1} \in \mathbb{C}^n$, $\tilde{h}_{1:k+1,k} \in \mathbb{C}^{k+1}$ such that $\tilde{V}_k^H\tilde{v}_{k+1} \approx 0$ and $\tilde{w}_{k+1} \approx [\tilde{V}_k, \tilde{v}_{k+1}]\tilde{h}_{1:k+1,k}$.

As before all vector operations yielding a vector are inexact, while the scalar products are exact. The inexactness is expressed in the vectors $f_{k+1}^{(*)}$.

Algorithm 5 *CGS (inexact)*

$\tilde{h}_{1:k,k} = \tilde{V}_k^H \tilde{w}_{k+1}$
$\tilde{l}_{k+1} = \tilde{w}_{k+1} - \tilde{V}_k \tilde{h}_{1:k,k} - f_{k+1}^{(0)}$
$\tilde{h}_{k+1,k} = \|\tilde{l}_{k+1}\|_2$
$\tilde{v}_{k+1} = (\tilde{l}_{k+1} - f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$

Algorithm 6 *CGSre (inexact)*

$s^{(0)} = \tilde{V}_k^H \tilde{w}_{k+1}$
$\tilde{l}_{k+1}^{(0)} = \tilde{w}_{k+1} - \tilde{V}_k s^{(0)} - f_{k+1}^{(0)}$
$s^{(1)} = \tilde{V}_k^H \tilde{l}_{k+1}^{(0)}$
$\tilde{l}_{k+1}^{(1)} = \tilde{l}_{k+1}^{(0)} - \tilde{V}_k s^{(1)} - f_{k+1}^{(1)}$
$\tilde{h}_{k+1,k} = \|\tilde{l}_{k+1}^{(1)}\|_2$, $\tilde{h}_{1:k,k} = s^{(0)} + s^{(1)}$
$\tilde{v}_{k+1} = (\tilde{l}_{k+1}^{(1)} - f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$

Algorithm 7 *MGS (inexact)*

$\tilde{l}_{k+1}^{(0)} = \tilde{w}_{k+1}$
for $i = 1, \ldots, k$ **do**
    $\tilde{h}_{i,k} = \tilde{v}_i^H \tilde{l}_{k+1}^{(i-1)}$
    $\tilde{l}_{k+1}^{(i)} = \tilde{l}_{k+1}^{(i-1)} - \tilde{v}_i \tilde{h}_{i,k} - f_{k+1}^{(i)}$
end for
$\tilde{h}_{k+1,k} = \|\tilde{l}_{k+1}^{(k)}\|_2$
$\tilde{v}_{k+1} = (\tilde{l}_{k+1}^{(k)} - f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$

Algorithm 8 *MGSre (inexact)*

$\tilde{l}_{k+1}^{(0)} = \tilde{w}_{k+1}$
for $i = 1, \ldots, k$ **do**
    $s_i^{(0)} = \tilde{v}_i^H \tilde{l}_{k+1}^{(i-1)}$
    $\tilde{l}_{k+1}^{(i)} = \tilde{l}_{k+1}^{(i-1)} - \tilde{v}_i s_i^{(0)} - f_{k+1}^{(i)}$
end for
for $i = 1, \ldots, k$ **do**
    $s_i^{(1)} = \tilde{v}_i^H \tilde{l}_{k+1}^{(k+i-1)}$
    $\tilde{l}_{k+1}^{(k+i)} = \tilde{l}_{k+1}^{(k+i-1)} - \tilde{v}_i s_i^{(1)} - f_{k+1}^{(k+i)}$
end for
$\tilde{h}_{k+1,k} = \|\tilde{l}_{k+1}^{(2k)}\|_2$, $\tilde{h}_{1:k,k} = s^{(0)} + s^{(1)}$
$\tilde{v}_{k+1} = (\tilde{l}_{k+1}^{(2k)} - f_{k+1}^{(S)})/\tilde{h}_{k+1,k}$

The norm of the error vectors $f_{k+1}^{(*)}$ depend on ε as follows: i) In all variants $\|f_{k+1}^{(S)}\|_2 \leq \varepsilon \|l\|_2$ where l is the vector to be scaled. ii) In CGS and CGSre $\|f_{k+1}^{(0)}\|_2$ and $\|f_{k+1}^{(1)}\|_2$ are bounded by $k\|\tilde{w}_{k+1}\|_2$ and $k\|\tilde{l}_{k+1}^{(0)}\|_2$, respectively. iii) In MGS and MGSre $\|f_{k+1}^{(i)}\|_2$ is bounded by $\|\tilde{l}_{k+1}^{(i-1)}\|_2$, $i = 1, \ldots, 2k$. In the experiments we have used $\varepsilon = 10^{-10}$.

In our test we chose A as a (notoriously ill-conditioned) Vandermonde matrix, more precisely, $a_{ij} = (j/m)^{i-1}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ (as in [17, section 20.7]) with $n = 300$, $m = 180$. The condition number of A_k (the matrix consisting of the leading k columns of A) grows rapidly with k, e.g., already for $k = 9$ the condition number reaches $\kappa(A_9) \approx 10^{16}$.

Algorithm 4 computes the QR factorization column by column, i.e., $\tilde{V}_k \tilde{R}_k = [\tilde{v}_1, \ldots, \tilde{v}_k][\tilde{r}_{ij}]_{i,j=1}^k$ is a QR factorization of A_k. Thus, we are computing a sequence of nested inexact QR factorizations and it makes sense to monitor its quality as it evolves with the number of processed columns k.

Figure 1 illustrates the loss of orthogonality of $\tilde{V}_k$ computed by the different variants measured by $\|\tilde{V}_k^H \tilde{V}_k - I\|_F$.

We observe that, unsurprisingly, CGS performs worst, completely losing orthogonality already after just three processed columns. Perhaps somewhat more surprising is that all variants without reorthogonalzation lose orthogonality after five columns. On the other hand, every variant with reorthogonalization achieves orthogonality to ε-level – at least until 90 columns. At this point CGSre loses orthogonality, as well. However, ComGSre and MGSre keep the loss of orthogonality to order ε for the whole process. We note that ComGS and MGS behave similarly with and without reorthogonalization.

The main advantage of ComGS(re) compared to the other methods lies in the availability of the cross product matrix $D_k = \tilde{V}_k^H \tilde{V}_k$, which provides implicitly an orthonormal basis $\hat{V}_k$ of the search space. It can be seen from Figure 1 that $\hat{V}_k$ is orthogonal to machine precision, i.e., $\hat{V}_k$ is even closer to orthogonality than the basis obtained from MGS with reorthogonalization. Note, that this holds for plain ComGS (without reorthogonalization) as well, although $\tilde{V}_k$ loses orthogonality completely. In other words, even without reorthogonalization ComGS gives as

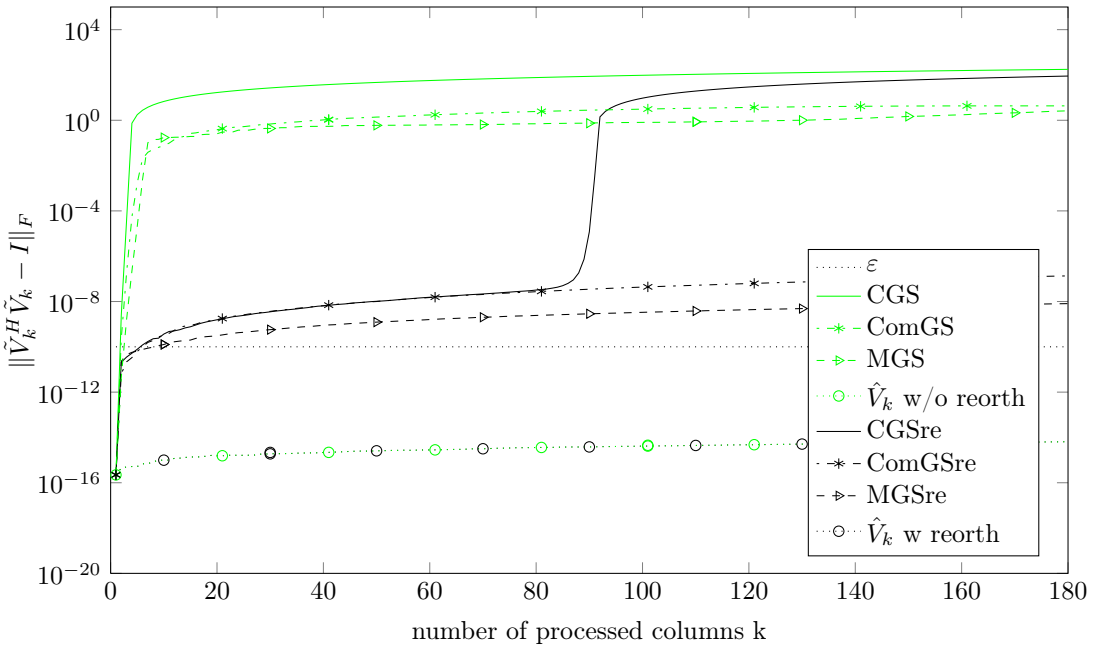

Fig. 1 Loss of orthogonality in the inexact QR factorization using the Gram Schmidt variants: Classical Gram Schmidt (CGS), modified Gram Schmidt (MGS), compensated Gram Schmidt (ComGS, $\tilde{V}_k$ and $\hat{V}_k$), each in individual black lines, and additionally each method with reorthogonalization (gray lines, appended "re" to the variant name).

good results as in the case with reorthogonalization. In other words, reorthogonalization is not necessary if $\hat{V}$ from ComGS is used even for this pathological example.

Thus for all following experiments we have used ComGS(re).

Additionally, we have checked $\|A_k - V_k R_k\|_F$, which is of order ε for every variant.

Example 2: In the second test we verify the bounds of Corollary 1. We applied the inexact Arnoldi method, Algorithm 1 to a matrix A built by the following MATLAB command

$$\text{A=}\mathbf{diag}\text{([10, 9, 8, 7, 0.1+0.9*}\mathbf{rand}\text{(1,n-4)])}$$

i.e., a diagonal matrix with four large eigenvalues at $7, 8, 9, 10$ and the remaining eigenvalues between 0.1 and 1. So, $\|A\|_2 = 10$. Since Arnoldi's method is invariant under unitary similarity transformations of A [3, 308 f.], diagonal matrices represent the general case.

We used $n = 10^5$, $\varepsilon = 10^{-10}$, did 10 Arnoldi steps and set $\tilde{v}_1$ to

$$\text{v1=[}\mathbf{randn}\text{(4,1);}\mathbf{zeros}\text{(n-4,1)]}$$

normalized to unit norm, i.e., $\tilde{v}_1$ is nonzero in the first four components only. Thus it is in the invariant subspace of A corresponding to the four large eigenvalues and the exact Arnoldi method would experience a lucky breakdown after four iterations. The inexact Arnoldi iteration will experience a near breakdown at that point.

As orthonormalization scheme we used ComGS with and without reorthogonalization. Figure 2 plots the distance to orthogonality of $\tilde{V}_k$ (evolving with k) and its two bounds given in Corollary 1. Additionally, the distance to orthogonality of $\hat{V}_k$ is depicted.

We observe that for the first 4 steps the loss of orthogonality in $\tilde{V}_k$ is of order ε. Then, for plain ComGS orthogonality is lost in the 5th iteration (as predicted by a huge $\kappa_4 \approx 10^6$). On the other hand using ComGSre, $\tilde{V}_k$ stays orthogonal to order ε throughout all iterations. Moreover, as in Example 1, $\hat{V}_k$ is orthogonal to machine precision in both cases.

Turning towards the bounds we observe that, first of all, they hold. The recursive bound δ_k overestimates (by an almost constant margin of roughly 10^2), but closely follows the qualitative

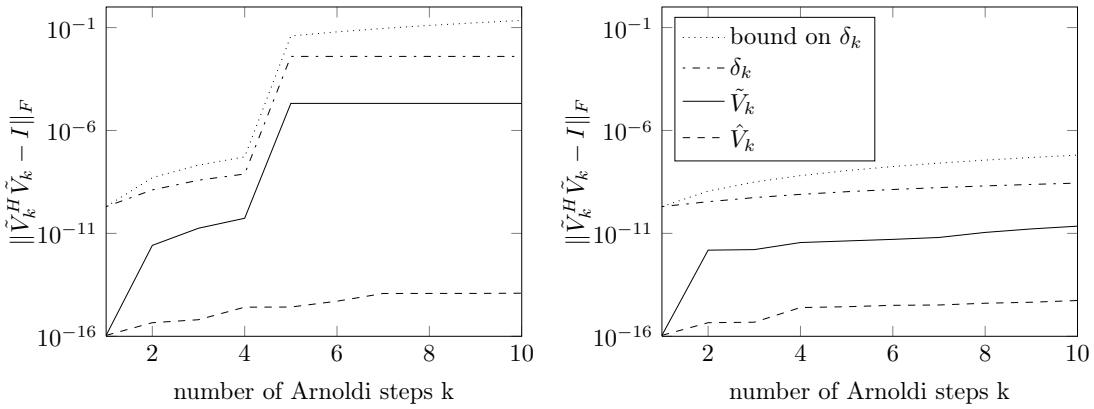

Fig. 2 Loss of orthogonality in the inexact Arnoldi algorithm without (left) and with (right) reorthogonalization.

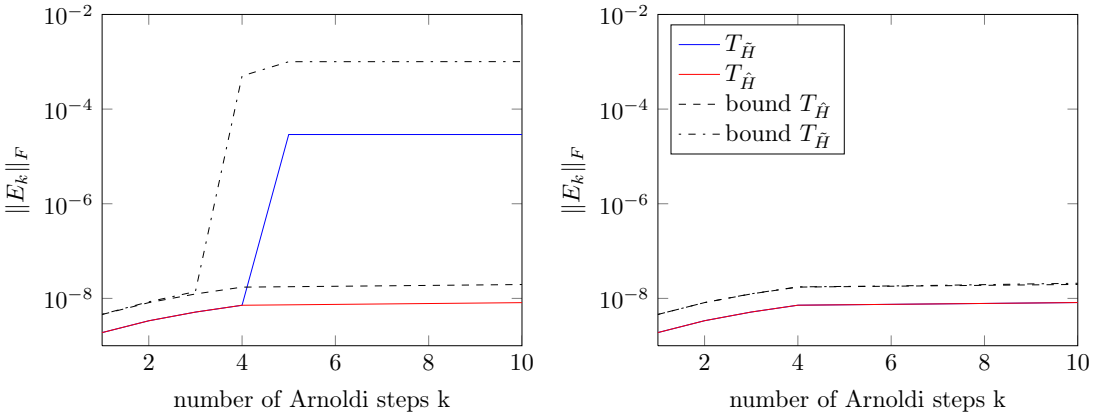

Fig. 3 Backward error of the inexact Arnoldi algorithm and its bounds without (left) and with (right) reorthogonalization.

trends. Especially the jump from step 4 to step 5 (in the left plot) is correctly captured. The closed form bound is (unsurprisingly) less sharp (as can be explained by the crude estimation of $\min\{\sqrt{k}(1+\varepsilon), 1+\zeta_k\}$ by $\sqrt{k}(1+\varepsilon)$ in the derivation).

Example 3: In our third test we investigate the backward error matrices E_k that are necessary to turn (20) into an exact Krylov relation for given $\hat{V}_{k+1}$ and several choices of B_k. We illustrate the Frobenius norm of E_k and verify its bound of Corollary 2.

The matrix A as well as all other parameters are reused form Example 2.

Figure 3 shows the results for the choices $B_k = T_{\tilde{H}}$ and $B_k = T_{\hat{H}}$ and their respective bounds from (25) as they evolve over k. Since the matrix E_k is never actually formed, we state how we obtain its Frobenius norm. From (22) we have $\|E_k\|_F = (2\|E_k\hat{V}_k\|_F^2 + \|B_k - S_k\|_F^2)^{\frac{1}{2}}$, where the term $\|E_k\hat{V}_k\|_F$ can in turn be evaluated (using (20)) using

$$E_k\hat{V}_k = A\hat{V}_k - \hat{V}_k B_k - \hat{v}_{k+1}e_k^T\hat{h}_{k+1,k}.$$

Again, the results depend on whether reorthogonalization is used. Without reorthogonalization $\|E_k\|$ is of order ε for $B_k = T_{\hat{H}}$. The bound correctly captures the trends of the curve and overestimates the actual value by an almost constant factor of just 2.

Using $\tilde{H}_k$ instead of $\hat{H}$ yields similar results – for the first 4 steps. Starting from step 5 (when the near breakdown happened) this choice requires a much larger perturbation E_k. The bound also captures the trends, and overestimates by a factor of ≈ 10. However, the bound jumps one

step earlier than the actual value (this can be explained by the occurrence of the term ζ_{k+1} in (25) that seems to be overly conservative).

With reorthogonalization, there is no relevant difference between $T_{\check{H}_k}$ and $T_{\hat{H}_k}$, neither in $\|E_k\|_F$, nor in its bounds. All these quantities are of order ε.

6 Conclusions

We have investigated the behavior of Arnoldi's method in settings where matrix-vector multiplication, vector addition and -scaling are inexact, but scalar products can be evaluated without error. We have devised a variant of Gram Schmidt orthogonalization, called compensated Gram Schmidt (ComGS), taylored to this scenario. We have shown that ComGS (possibly enhanced by reorthogonalization) produces a basis that is orthogonal to the same level of accuracy ε as the vector operations themselves. Moreover, ComGS implicitly provides a second basis that is orthogonal to machine precision – even without reorthogonalization. Numerical tests confirm the proven bounds on the loss of orthogonality for ComGS.

We then went on to show that for Hermitian A the inexact Arnoldi method yields an exact Krylov relation of a nearby matrix $A + E$. The key idea was to replace the non-Hermitian Hessenberg matrix. Several choices are possible (some of them even tridiagonal) and we proved bounds for the corresponding E. In numerical tests the bounds were confirmed to correctly predict the trends of the curve, and to be accurate within one order of magnitude.

We conclude that if the improved basis and corresponding Hessenberg matrix are used, then the norm of the backward error E is kept to ε-level. This holds even under unfavorable conditions like the lack of reorthogonalization and occurrence of near breakdowns of the method.

A convergence analysis of the Arnoldi method under backward perturbations is work in progress.

Acknowledgements We thank Volker Mehrmann, André Gaul and Agnieszka Miedlar (TU Berlin) as well as Mario Arioli (Rutherford Appleton Lab, UK), Rich Lehoucq (Sandia National Lab, US) and Daniel Kressner (EPF Lausanne, Switzerland) for insightful discussions on the topic.

References

1. Arnoldi, W.E.: The principle of minimized iteration in the solution of the matrix eigenvalue problem. Quart. Appl. Math. **9**, 17–29 (1951)
2. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H. (eds.): Templates for the solution of algebraic eigenvalue problems, *Software, Environments, and Tools*, vol. 11. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2000). A practical guide
3. Björck, Å.: Numerics of Gram-Schmidt orthogonalization. Linear Algebra Appl. **197/198**, 297–316 (1994)
4. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math. **59**(8), 1207–1223 (2006)
5. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k-term approximation. J. Amer. Math. Soc. **22**(1), 211–231 (2009)
6. Daniel, J.W., Gragg, W.B., Kaufman, L., Stewart, G.W.: Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization. Math. Comp. **30**(136), 772–795 (1976)
7. Davenport, M.A., Duarte, M.F., Eldar, Y.C., Kutyniok, G.: Introduction to compressed sensing. In: Compressed sensing, pp. 1–64. Cambridge Univ. Press, Cambridge (2012)
8. Demmel, J.W.: Applied Numerical Linear Algebra. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1997)
9. Dhillon, I.S., Parlett, B.N., Vömel, C.: The design and implementation of the MRRR algorithm. ACM Trans. Math. Software **32**(4), 533–560 (2006)
10. Elman, H.C., Meerbergen, K., Spence, A., Wu, M.: Lyapunov inverse iteration for identifying Hopf bifurcations in models of incompressible flow. SIAM J. Sci. Comput. **34**(3), A1584–A1606 (2012)
11. van den Eshof, J., Sleijpen, G.L.G.: Inexact Krylov subspace methods for linear systems. SIAM J. Matrix Anal. Appl. **26**(1), 125–153 (2004)
12. Giraud, L., Langou, J., Rozloznik, M.: The loss of orthogonality in the Gram-Schmidt orthogonalization process. Comput. Math. Appl. **50**(7), 1069–1075 (2005)
13. Giraud, L., Langou, J., Rozložník, M., van den Eshof, J.: Rounding error analysis of the classical Gram-Schmidt orthogonalization process. Numer. Math. **101**(1), 87–100 (2005)

14. Golub, G.H., Van Loan, C.F.: Matrix Computations, third edn. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD (1996)
15. Grasedyck, L.: Hierarchical singular value decomposition of tensors. SIAM J. Matrix Anal. Appl. **31**(4), 2029–2054 (2009/10)
16. Hackbusch, W., Khoromskij, B.N., Tyrtyshnikov, E.E.: Hierarchical Kronecker tensor-product approximations. J. Numer. Math. **13**(2), 119–156 (2005)
17. Higham, N.J.: Accuracy and stability of numerical algorithms, second edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2002)
18. Hoffmann, W.: Iterative algorithms for Gram-Schmidt orthogonalization. Computing **41**(4), 335–348 (1989)
19. Holtz, S., Rohwedder, T., Schneider, R.: On manifolds of tensors of fixed TT-rank. Numer. Math. **120**(4), 701–731 (2012)
20. Huckle, T., Waldherr, K., Schulte-Herbrüggen, T.: Computations in quantum tensor networks. Linear Algebra Appl. **438**(2), 750–781 (2013)
21. Jaimoukha, I.M., Kasenally, E.M.: Krylov subspace methods for solving large Lyapunov equations. SIAM J. Numer. Anal. **31**(1), 227–251 (1994)
22. Jaimoukha, I.M., Kasenally, E.M.: Oblique projection methods for large scale model reduction. SIAM J. Matrix Anal. Appl. **16**(2), 602–627 (1995)
23. Kressner, D.: Numerical methods for general and structured eigenvalue problems, *Lecture Notes in Computational Science and Engineering*, vol. 46. Springer-Verlag, Berlin (2005)
24. Kressner, D., Tobler, C.: htucker - a matlab toolbox for tensors in hierarchical tucker format. Tech. rep., ETH Z"urich (2012)
25. Lehoucq, R.B., Sorensen, D.C., Yang, C.: ARPACK users' guide - Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods, *Software, Environments, and Tools*, vol. 6. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1998)
26. Liesen, J., Strakos, Z.: Krylov Subspace Methods: Principles and Analysis. Oxford University Press (2012)
27. Mackey, D.S., Mackey, N., Tisseur, F.: Structured mapping problems for matrices associated with scalar products. I. Lie and Jordan algebras. SIAM J. Matrix Anal. Appl. **29**(4), 1389–1410 (2007)
28. Oseledets, I., Tyrtyshnikov, E.: TT-cross approximation for multidimensional arrays. Linear Algebra Appl. **432**(1), 70–88 (2010)
29. Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**(5), 2295–2317 (2011)
30. Paige, C.C.: Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. J. Inst. Math. Appl. **18**(3), 341–349 (1976)
31. Parlett, B.N.: The Symmetric Eigenvalue Problem. Prentice-Hall Inc., Englewood Cliffs, N.J. (1980). Prentice-Hall Series in Computational Mathematics
32. Saad, Y.: Iterative methods for sparse linear systems, second edn. Society for Industrial and Applied Mathematics, Philadelphia, PA (2003)
33. Saad, Y.: Numerical methods for large eigenvalue problems, rev. ed. edn. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (2011)
34. Schaller, G.: Adiabatic preparation without quantum phase transitions. Phys. Rev. A **78**, 032,328 (2008)
35. Schützhold, R., Schaller, G.: Adiabatic quantum algorithms as quantum phase transitions: First versus second order. Phys. Rev. A **74**, 060,304 (2006)
36. Simoncini, V.: A matrix analysis of Arnoldi and Lanczos methods. Numer. Math. **81**(1), 125–141 (1998)
37. Simoncini, V.: Variable accuracy of matrix-vector products in projection methods for eigencomputation. SIAM J. Numer. Anal. **43**(3), 1155–1174 (2005)
38. Simoncini, V., Szyld, D.B.: Theory of inexact Krylov subspace methods and applications to scientific computing. SIAM J. Sci. Comput. **25**(2), 454–477 (2003)
39. Stewart, G.W.: A Krylov-Schur algorithm for large eigenproblems. SIAM J. Matrix Anal. Appl. **23**(3), 601–614 (electronic) (2001/02)
40. Stewart, G.W.: Backward error bounds for approximate Krylov subspaces. Linear Algebra Appl. **340**, 81–86 (2002)
41. Sun, J.G.: Perturbation bounds for the Cholesky and QR factorizations. BIT **31**(2), 341–352 (1991)