An Augmented Analysis of the Perturbed Two-sided Lanczos Tridiagonalization Process☆

Christopher C. Paige[a,1,*], Ivo Panayotov[b], Jens-Peter M. Zemke[c]

[a]*School of Computer Science, McGill University, Montréal, Québec, Canada, H3A 0E9*
[b]*Oxford University Mathematical Institute, 24-29 St Giles', Oxford, England, OX1 3LB*
[c]*Institut für Mathematik, Technische Universität Hamburg-Harburg, D-21073 Hamburg, Germany.*

Abstract

We generalize an augmented rounding error result that was proven for the symmetric Lanczos process in [SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2347–2359], to the two-sided (usually unsymmetric) Lanczos process for tridiagonalizing a square matrix. We extend the analysis to more general perturbations than rounding errors in order to provide tools for the analysis of inexact and related methods. The aim is to develop a deeper understanding of the behavior of all these methods. Our results take the same form as those for the symmetric Lanczos process, except for the bounds on the backward perturbation terms (the generalizations of backward rounding errors for the augmented system). In general we cannot derive tight a priori bounds for these terms as was done for the symmetric process, but a posteriori bounds are feasible, while bounds related to certain properties of matrices would be theoretically desirable.

Keywords: Lanczos process, Finite precision, Perturbation analysis, Non-Hermitian matrix, Loss of bi-orthogonality, Augmented error analysis
2000 MSC: 65F15

1. Introduction

The Lanczos tridiagonalization process, which we refer to as the Lanczos process, was introduced in [19] for solving eigenvalue problems and was adapted in [20] for solving linear systems. The process comes in two flavors: a symmetric version that is applied to real symmetric or complex Hermitian matrices, and a two-sided version that is applicable to general square matrices. We use the term "two-sided", since it can also be applied to Hermitian matrices.

The symmetric Lanczos process in all its variants is currently the method of choice for solving many problems involving large and sparse real symmetric, and generally complex

☆In fond memory of Miki Neumann, a warm and lovely person.

*Corresponding author
Email addresses: paige@cs.mcgill.ca (Christopher C. Paige), Ivo.Panayotov@maths.ox.ac.uk (Ivo Panayotov), zemke@tu-harburg.de (Jens-Peter M. Zemke)

[1]The research of this author was supported by NSERC of Canada grant OGP0009236.

Hermitian, matrices. Its theoretical and numerical behavior has been extensively studied, the latter, for example, in [23, 24, 25, 26, 13, 14, 30]. A deep treatment of the symmetric Lanczos process and an extensive list of references is given in [21], with an elegant summary in [22]. A compilation of early references related to the (symmetric) Lanczos process and the mathematically equivalent method of conjugate gradients (CG) [16] for the period 1948–1976 can be found in [11].

In contrast to reliable direct methods, a traditional backward error analysis [40, 41, 17], or even a mixed forward-backward error analysis [17], does not apply. Greenbaum [13] proved that quantities related to a finite precision Lanczos process can be interpreted as the outcome of an exact Lanczos process applied to a certain larger matrix. This larger matrix is not known in advance nor a posteriori, yet this result helps to increase the understanding of the process, see [14].

Recently Paige [28] performed a rounding error analysis of the symmetric Lanczos process based on "augmentation", that is, a kind of analysis where the process is considered from the point of view of a larger dimensional space than the one of the original problem. He showed that from this point of view the process satisfies a relation which we will refer to as "recursive augmented stability". This shows that the computed tridiagonal matrix comes from an error free Lanczos process applied to a slightly symmetrically perturbed block diagonal matrix which is known at the start of each step: at step $k+1$ it is $\mathsf{diag}(\mathbf{T}_k, \mathbf{A})$ where $\mathbf{A}$ is the system matrix and $\mathbf{T}_k$ is the computed tridiagonal matrix from the previous step. Paige's analysis is interesting in two main aspects. The first aspect is that it can be used to readily explain some of the observed rounding error aspects of the Lanczos process applied to the eigenvalue problem, in particular a loss of memory of converged eigenvectors, leading to the formation of tight clusters of eigenvalues corresponding to simple eigenvalues of the matrix $\mathbf{A}$. The second aspect is more far-reaching, as the established relationships may be useful in further analysis of the process for the eigenvalue problem, for systems of linear equations, and all other methods based on the Lanczos process.

Following an idea of Charles Sheffield (for the history we refer to the remarks in [6]), augmented rounding error analysis has provided useful results, see [6, 7, 5, 29]. The common theme in augmented rounding error analysis is to study the behavior of algorithms which in theory construct orthonormal bases of vectors, but which in practice lose orthogonality because of the effects of rounding errors. Notably, the proof of backward stability of the MGS-GMRES method was derived in [29] via an augmented rounding error analysis. Inspired by these uses of augmented error analysis, Paige [27] defined a unitary $(n+k)\times(n+k)$ matrix based on a sequence of vectors $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k \in \mathbb{C}^n$, of unit length, for any $k, n \in \mathbb{N}$ (including the case $k > n$), that incorporated the vectors together with information about the loss of orthogonality between them. This matrix, which he called a unitary augmentation, incorporates the common ideas among the augmented rounding error analyses cited above. It was demonstrated in [27] that this unitary augmentation can be used very elegantly to simplify existing rounding error analyses related to loss of orthogonality.

The two-sided Lanczos process is currently one of the popular methods for solving eigenvalue problems of large and sparse square matrices. For such problems there are other popular methods, such as the one by Arnoldi [2]. Nevertheless, the two-sided Lanczos process is an interesting choice because of the low-storage requirements. For some history, implementation and examples of industrial applications of the two-sided Lanczos

process, see, e.g., the book by Komzsik [18]. While the symmetric Lanczos process is somewhat well understood numerically, much less is known for the two-sided process. Perhaps the first significant rounding error analysis was done by Bai [3] who extended the analysis of the symmetric process of Paige [25], and provided a componentwise rather than normwise analysis. Up to now there are not very many results on the rounding-error analysis of the two-sided Lanczos process and its generalizations, we mention explicitly the pioneering works [9, 8, 3, 35].

There exist various generalizations of the Lanczos process, like block variants, e.g., [36, 12, 4], and variants with different numbers of left-hand and right-hand side starting vectors, e.g., [1]. Recently, Lanczos processes with $s \in \mathbb{N}$ left-hand side starting vectors and one right-hand side starting vector are enjoying a renaissance: in disguise, namely, as the workhorses of the IDR [39] and IDR(s) [34, 10] based methods, see, e.g., [15, 33, 32]. To the knowledge of the authors, no error analysis of these generalizations comparable with that in [28] is available.

The results here are valid for *general* perturbation terms of *any* magnitude, not just rounding errors. Thus these results extend to the error analysis of inexact methods and some of the above generalizations, such as classical IDR [39], IDR(1) [34, 10] and BiCGStab [38, 37]. To handle all cases we will use the more general term "perturbation" in place of the more restricted term "rounding error", and "backward perturbation" in place of "backward rounding error of the augmented system".

1.1. Motivation

Our aim in the current paper is to extend the augmented rounding error analysis performed in [28] to basic (i.e., unmodified) perturbed two-sided Lanczos processes. We show that the results in [28] extend in form. However unlike that earlier case, we are in general unable to derive tight a priori bounds on the terms induced by the initial perturbations. It is hoped that future work will lead to bounds related to certain properties of matrices, such as distance from symmetry, and that practical bounds will be obtainable a posteriori. We believe that the present intermediate results will help to improve the theoretical understanding of the finite precision two-sided Lanczos process and are worth adding to the limited existing literature.

Briefly and roughly, for the finite precision case our results extend those of Bai [3] in showing that the computed tridiagonal matrix is exact for a slightly (when the algorithm is far from breakdown) perturbed augmented system with exactly bi-orthogonal matrices.

1.2. Outline

In section 2 we present the basic perturbed two-sided Lanczos process. In section 3.1 we discuss the loss of bi-orthonormality between the Lanczos vectors (to be defined below), while section 3.2 is devoted to describing the augmented matrices, defined in [27], which we call *Sheffield augmentation* in this paper, and their relation to this loss of bi-orthonormality. In section 4 we carry out an augmented analysis for the two-sided Lanczos process and interpret this, our main result. We discuss bounds on the rounding error terms defined throughout the paper in section 5.

1.3. Notation

We use standard notation. The identity matrix of size $n \times n$ is denoted by $\mathbf{I}_n \in \mathbb{C}^{n\times n}$, its columns by $\mathbf{e}_j \in \mathbb{C}^n$, $j = 1, \ldots, n$, and its elements by the Kronecker delta $\delta_{ij} \in \mathbb{C}$, $i, j = 1, \ldots, n$. The sum of all columns $\mathbf{e} \in \mathbb{C}^n$ is the vector of all ones. We define an expanded variant of the identity matrix $\underline{\mathbf{I}}_k \in \mathbb{C}^{(k+1)\times k}$ that has an additional zero row vector appended at the bottom. The zero matrix is denoted by $\mathbf{O}$, the zero column vector is denoted by $\mathbf{o}$. We use MATLAB-like operators triu and tril, see (5), and the overloaded operator diag which has to be understood in context. The operator diag extends to block matrices, e.g.,

$$\mathsf{diag}(\mathbf{T}_k, \mathbf{A}) := \begin{pmatrix} \mathbf{T}_k & \mathbf{O}_{k,n} \\ \mathbf{O}_{n,k} & \mathbf{A} \end{pmatrix} \in \mathbb{C}^{(k+n)\times(k+n)}, \quad \mathbf{T}_k \in \mathbb{C}^{k\times k}, \quad \mathbf{A} \in \mathbb{C}^{n\times n}.$$

The letter ϵ is used in two contexts: it either denotes a small, variable quantity, or, in finite precision arithmetic, the unit roundoff. Appended symbols $^\mathsf{T}$ and $^\mathsf{H}$ denote the transpose and complex conjugate transpose, respectively. We use boldface symbols for matrices and vectors, and regular font symbols for scalars. Most of the mathematical relationships in this paper come in pairs, see e.g., (1) below; we use non-hatted symbols to represent the first relationship and hatted symbols to represent the second relationship in each pair. The Lanczos process is intimately related to a sequence of unreduced tridiagonal matrices denoted by the letter $\mathbf{T}_k \in \mathbb{C}^{k\times k}$, $k \in \mathbb{N}$. Extended counterparts with an additional row at the bottom are denoted by $\underline{\mathbf{T}}_k = \mathbf{T}_{k+1}\underline{\mathbf{I}}_k \in \mathbb{C}^{(k+1)\times k}$, $k \in \mathbb{N}$. We use some terminology of [15], namely (perturbed) Hessenberg decompositions [15, p. 4, Eqn. (1.12), Footnote 2] and a unified and simplified scheme to denote names of algorithms, e.g., GMRES in place of GMRES.

2. The two-sided Lanczos process

A short but good introduction to the basic two-sided Lanczos process in exact arithmetic, as well as a basic rounding error analysis, is given in [3]. Such analyses depend on the particular implementation, and so details will not be given here.

For a general square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ we consider the basic two-sided Lanczos process with perturbations $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$ introduced by some inexact process. The processes we consider satisfy the following two perturbed Hessenberg decompositions, for the finite precision case see, e.g., [3, Theorem 3.1]:

$$\mathbf{A}\mathbf{Q}_k + \mathbf{E}_k = \mathbf{Q}_{k+1}\underline{\mathbf{T}}_k = \mathbf{Q}_k\mathbf{T}_k + \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^\mathsf{T}, \tag{1a}$$

$$\mathbf{A}^\mathsf{H}\widehat{\mathbf{Q}}_k + \widehat{\mathbf{E}}_k = \widehat{\mathbf{Q}}_{k+1}\underline{\widehat{\mathbf{T}}}_k = \widehat{\mathbf{Q}}_k\widehat{\mathbf{T}}_k + \widehat{\mathbf{q}}_{k+1}\widehat{\beta}_{k+1}\mathbf{e}_k^\mathsf{T}. \tag{1b}$$

Here,

$$\underline{\mathbf{T}}_k := \begin{pmatrix} \alpha_1 & \overline{\widehat{\beta}_2} & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \overline{\widehat{\beta}_k} \\ & & \beta_k & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix}, \quad \underline{\widehat{\mathbf{T}}}_k := \begin{pmatrix} \overline{\alpha_1} & \overline{\beta_2} & & \\ \widehat{\beta}_2 & \overline{\alpha_2} & \ddots & \\ & \ddots & \ddots & \overline{\beta_k} \\ & & \widehat{\beta}_k & \overline{\alpha_k} \\ & & & \widehat{\beta}_{k+1} \end{pmatrix}, \tag{2}$$

where as usual the over-bar signifies "complex conjugate". We denote the leading square tridiagonal matrices by $\mathbf{T}_k := \underline{\mathbf{I}}_k^{\mathsf{T}}\underline{\mathbf{T}}_k \in \mathbb{C}^{k\times k}$ and $\widehat{\mathbf{T}}_k := \underline{\mathbf{I}}_k^{\mathsf{T}}\underline{\widehat{\mathbf{T}}}_k \in \mathbb{C}^{k\times k}$. The coefficients are chosen so that in the error-free case, i.e., in the absence of perturbations $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$, $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$ are bi-orthonormal, i.e., $\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{Q}_k = \mathbf{I}_k$, and $\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{q}_{k+1} = \mathbf{Q}_k^{\mathsf{H}}\widehat{\mathbf{q}}_{k+1} = \mathbf{o}$. This excludes more refined two-sided Lanczos process variants, e.g., those based on mere bi-*orthogonality* [8]. Note that this bi-orthonormality shows $\mathbf{T}_k^{\mathsf{H}} = \widehat{\mathbf{T}}_k$ in the error free version of (1), and the computational implementation in [3] also ensures this, so we assume $\mathbf{T}_k^{\mathsf{H}} = \widehat{\mathbf{T}}_k$ here. We term the recurrences (1) the *Lanczos recurrences*. The columns of $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$ are called *right* and *left Lanczos vectors*, respectively.

In the presence of perturbations, $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$ are non-zero and bi-orthogonality between the Lanczos vectors is quickly lost. To construct $\mathbf{q}_{j+1}$ and $\widehat{\mathbf{q}}_{j+1}$, the process first generates non-normalized versions of these vectors. If either is zero this can be considered a successful completion. Even in the exact case it may happen that at a particular step $j < n$ the inner product of the non-zero non-normalized vectors is zero, in which case $\mathbf{q}_{j+1}$ and $\widehat{\mathbf{q}}_{j+1}$ can no longer be constructed. This occurrence is called a breakdown of the process. This only causes trouble in the two-sided Lanczos process, since in the symmetric version $\mathbf{q}_{j+1} = \widehat{\mathbf{q}}_{j+1}$. An exact breakdown is extremely rare in practice; a more common occurrence is a near breakdown, when the non-normalized Lanczos vectors at step j are almost orthogonal. In that case, scaling $\widehat{\mathbf{q}}_{j+1}$ and $\mathbf{q}_{j+1}$ enforces the condition $\widehat{\mathbf{q}}_{j+1}^{\mathsf{H}}\mathbf{q}_{j+1} = 1$, but causes at least one of $\widehat{\mathbf{q}}_{j+1}$ and $\mathbf{q}_{j+1}$ to have a very large norm. As we see later, large norms of the Lanczos vectors yield large bounds on the perturbation terms. Some strategies have been implemented to try to avoid the problems associated with near breakdown, e.g., in the look-ahead implementation of the process [31]. For simplicity we assume that the process is executed without look-ahead, and breakdown or near breakdown does not occur.

Since normalization of $\mathbf{q}_{k+1}$ and $\widehat{\mathbf{q}}_{k+1}$ is the last part of each step, in finite precision arithmetic it is straightforward to ensure that the diagonal of $\widehat{\mathbf{Q}}_{k+1}^{\mathsf{H}}\mathbf{Q}_{k+1}$ is almost the unit matrix except near breakdown, so that $\mathsf{diag}(\widehat{\mathbf{Q}}_{k+1}^{\mathsf{H}}\mathbf{Q}_{k+1}) = \mathbf{e} + \mathbf{c}$ for some small $\mathbf{c} \in \mathbb{C}^{k+1}$. Then with two small diagonal matrices $\widehat{\mathbf{S}}_{k+1}, \mathbf{S}_{k+1} \in \mathbb{C}^{(k+1)\times(k+1)}$ such that $\mathbf{I}_{k+1} = (\mathbf{I}_{k+1} + \widehat{\mathbf{S}}_{k+1})^{\mathsf{H}}(\mathbf{I}_{k+1} + \mathsf{diag}(\mathbf{c}))(\mathbf{I}_{k+1} + \mathbf{S}_{k+1})$, we can reformulate the Lanczos recurrences (1) as

$$\mathbf{A}\mathbf{Q}_k(\mathbf{I}_k + \mathbf{S}_k) + \mathbf{E}_k^{\text{new}} = \mathbf{Q}_{k+1}(\mathbf{I}_{k+1} + \mathbf{S}_{k+1})\underline{\mathbf{T}}_k, \tag{3a}$$

$$\mathbf{A}^{\mathsf{H}}\widehat{\mathbf{Q}}_k(\mathbf{I}_k + \widehat{\mathbf{S}}_k) + \widehat{\mathbf{E}}_k^{\text{new}} = \widehat{\mathbf{Q}}_{k+1}(\mathbf{I}_{k+1} + \widehat{\mathbf{S}}_{k+1})\underline{\widehat{\mathbf{T}}}_k, \tag{3b}$$

where the new perturbation matrices are given by

$$\mathbf{E}_k^{\text{new}} := \mathbf{E}_k - \mathbf{A}\mathbf{Q}_k\mathbf{S}_k - \mathbf{Q}_{k+1}\mathbf{S}_{k+1}\underline{\mathbf{T}}_k, \tag{4a}$$

$$\widehat{\mathbf{E}}_k^{\text{new}} := \widehat{\mathbf{E}}_k - \mathbf{A}^{\mathsf{H}}\widehat{\mathbf{Q}}_k\widehat{\mathbf{S}}_k - \widehat{\mathbf{Q}}_{k+1}\widehat{\mathbf{S}}_{k+1}\underline{\widehat{\mathbf{T}}}_k. \tag{4b}$$

We now assume that we work with the replacements: $\mathbf{Q}_{k+1}(\mathbf{I}_{k+1} + \mathbf{S}_{k+1}) \to \mathbf{Q}_{k+1}$, $\mathbf{E}_k^{\text{new}} \to \mathbf{E}_k$, similarly for hatted variants, so that $\mathsf{diag}(\widehat{\mathbf{Q}}_{k+1}^{\mathsf{H}}\mathbf{Q}_{k+1}) = \mathbf{e}$.

To describe the loss of bi-orthogonality we introduce the additive splitting

$$\widehat{\mathbf{U}}_k^{\mathsf{H}} + \mathbf{I}_k + \mathbf{U}_k = \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{Q}_k, \quad \widehat{\mathbf{U}}_k^{\mathsf{H}} := \mathsf{tril}(\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{Q}_k, -1), \quad \mathbf{U}_k := \mathsf{triu}(\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{Q}_k, 1) \tag{5}$$

which defines the strictly upper triagular $\widehat{\mathbf{U}}_k$ and $\mathbf{U}_k$, and define the vectors

$$\widehat{\mathbf{u}}_{k+1}^{\mathsf{H}} := \widehat{\mathbf{q}}_{k+1}^{\mathsf{H}}\mathbf{Q}_k \quad \text{and} \quad \mathbf{u}_{k+1} := \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{q}_{k+1}. \tag{6}$$

The strictly upper triangular matrices $\mathbf{U}_k, \widehat{\mathbf{U}}_k \in \mathbb{C}^{k\times k}$ and the vectors $\mathbf{u}_{k+1}, \widehat{\mathbf{u}}_{k+1} \in \mathbb{C}^k$ capture the loss of bi-orthogonality of the Lanczos vectors when the process is subject to perturbations. We now describe how this loss of bi-orthogonality develops as a function of the tridiagonal matrices $\mathbf{T}_k$ and $\widehat{\mathbf{T}}_k$ and the perturbations $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$. The results of the following lemma have been derived for the symmetric process in [25]. The derivation for the two-sided Lanczos process is similar; the results are contained in a proof in [3]. We provide a short proof valid for general perturbation terms in Eqns. (1).

Lemma 2.1 (Paige [25, Eqn. (22)], Bai [3, Eqns. (4.11,4.12)], [42])**.** *Consider a perturbed Lanczos process with no breakdown and Lanczos recurrences* (1)*, and let the loss of bi-orthogonality be captured by* $\mathbf{U}_k$*,* $\widehat{\mathbf{U}}_k$*,* $\mathbf{u}_{k+1}$*,* $\widehat{\mathbf{u}}_{k+1}$ *defined in* (5) *and* (6)*. Then these satisfy the perturbed Hessenberg decompositions*

$$\mathbf{T}_k\mathbf{U}_k - \begin{pmatrix}\mathbf{U}_k & \mathbf{u}_{k+1}\end{pmatrix}\underline{\mathbf{T}}_k = \mathbf{F}_k := \mathsf{triu}(\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k - \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{E}_k) + \mathbf{D}_k, \tag{7a}$$

$$\widehat{\mathbf{T}}_k\widehat{\mathbf{U}}_k - \begin{pmatrix}\widehat{\mathbf{U}}_k & \widehat{\mathbf{u}}_{k+1}\end{pmatrix}\underline{\widehat{\mathbf{T}}}_k = \widehat{\mathbf{F}}_k := \mathsf{triu}(\mathbf{E}_k^{\mathsf{H}}\widehat{\mathbf{Q}}_k - \mathbf{Q}_k^{\mathsf{H}}\widehat{\mathbf{E}}_k) + \widehat{\mathbf{D}}_k, \tag{7b}$$

where the diagonal matrices $\mathbf{D}_k$ *and* $\widehat{\mathbf{D}}_k$ *are defined by*

$$\begin{aligned}\mathbf{D}_k &:= \mathsf{triu}(\widehat{\mathbf{U}}_k^{\mathsf{H}}\mathbf{T}_k - \mathbf{T}_k\widehat{\mathbf{U}}_k^{\mathsf{H}} - \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{u}}_{k+1}^{\mathsf{H}}) \\ &= \mathsf{diag}(-\overline{\widehat{u}_{1,2}\widehat{\beta}_2}, \overline{\widehat{u}_{1,2}\widehat{\beta}_2} - \overline{\widehat{u}_{2,3}\widehat{\beta}_3}, \ldots, \overline{\widehat{u}_{k-1,k}\widehat{\beta}_k} - \overline{\widehat{u}_{k,k+1}\widehat{\beta}_{k+1}}),\end{aligned} \tag{8a}$$

and

$$\begin{aligned}\widehat{\mathbf{D}}_k &:= \mathsf{triu}(\mathbf{U}_k^{\mathsf{H}}\widehat{\mathbf{T}}_k - \widehat{\mathbf{T}}_k\mathbf{U}_k^{\mathsf{H}} - \mathbf{e}_k\overline{\beta_{k+1}}\mathbf{u}_{k+1}^{\mathsf{H}}) \\ &= \mathsf{diag}(-\overline{u_{1,2}\beta_2}, \overline{u_{1,2}\beta_2} - \overline{u_{2,3}\beta_3}, \ldots, \overline{u_{k-1,k}\beta_k} - \overline{u_{k,k+1}\beta_{k+1}}).\end{aligned} \tag{8b}$$

Proof. Since $\widehat{\mathbf{T}}_k = \mathbf{T}_k^{\mathsf{H}}$, $(1b)^{\mathsf{H}}\mathbf{Q}_k - \widehat{\mathbf{Q}}_k^{\mathsf{H}}(1a)$ equates to

$$\begin{aligned}\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k - \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{E}_k &= \underline{\widehat{\mathbf{T}}}_k^{\mathsf{H}}\widehat{\mathbf{Q}}_{k+1}^{\mathsf{H}}\mathbf{Q}_k - \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{Q}_{k+1}\underline{\mathbf{T}}_k \\ &= \mathbf{T}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{Q}_k - \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{Q}_k\mathbf{T}_k + \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}}\mathbf{Q}_k - \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} \\ &= \mathbf{T}_k(\widehat{\mathbf{U}}_k^{\mathsf{H}} + \mathbf{U}_k) - (\widehat{\mathbf{U}}_k^{\mathsf{H}} + \mathbf{U}_k)\mathbf{T}_k + \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{u}}_{k+1}^{\mathsf{H}} - \mathbf{u}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}},\end{aligned} \tag{9}$$

which can be reorganized as

$$\mathbf{T}_k\mathbf{U}_k - \begin{pmatrix}\mathbf{U}_k & \mathbf{u}_{k+1}\end{pmatrix}\underline{\mathbf{T}}_k = \widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k - \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{E}_k + \widehat{\mathbf{U}}_k^{\mathsf{H}}\mathbf{T}_k - \mathbf{T}_k\widehat{\mathbf{U}}_k^{\mathsf{H}} - \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{u}}_{k+1}^{\mathsf{H}}. \tag{10}$$

Note that the left hand side of this equation is upper triangular and the last three terms on the right hand side are lower triangular. Comparing the upper triangular parts on both sides leads to (7a). The second equality in (8a) displays the componentwise values of $\mathbf{D}_k$. Equation (7b), with (8b), is proven by considering the lower triangular part of (10), which amounts to a similar analysis to the above applied to $(1a)^{\mathsf{H}}\widehat{\mathbf{Q}}_k - \mathbf{Q}_k^{\mathsf{H}}(1b)$. □

We remark that the upper triangular perturbation matrices $\mathbf{F}_k$ and $\widehat{\mathbf{F}}_k$ are small whenever the Lanczos recurrences (1) are approximately satisfied, i.e., when $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$ are small, when $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$ are not too large, and when local bi-orthogonality is approximately obtained, i.e., the elements $u_{i,i+1}\beta_{i+1}$ and $\widehat{u}_{i,i+1}\widehat{\beta}_{i+1}$ corresponding to the first super-diagonals of $\mathbf{U}_k$ and $\widehat{\mathbf{U}}_k$ are small for $i = 1, \ldots, k$, see (8). In the finite precision case these have been proven to be small in good implementations of the symmetric process, but this will not hold in the standard two-sided algorithms if $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$ are large. In particular the assumption of local orthogonality made in [30, (13.13)] for the symmetric case cannot be made for local bi-orthogonality in the two-sided case, and the terms in (8) have to be bounded.

There is a structural symmetry here if we define throughout $\widehat{\mathbf{A}} := \mathbf{A}^{\mathsf{H}}$, $\widehat{\mathbf{I}}_k := \mathbf{I}_k$, and $\widehat{\mathbf{O}}_{k,k} := \mathbf{O}_{k,k}$. Since the original Lanczos recurrences (1) remain valid by interchanging hatted versions of symbols with non-hatted ones, and since our relationships come in pairs, every relationship that we derive will hold by swapping hatted symbols with non-hatted ones. For example, in (8), $\mathbf{D}_k$ turns to $\widehat{\mathbf{D}}_k$ and vice versa upon interchanging hatted letters with non-hatted counterparts.

3. Obtaining true bi-orthogonality via an augmented process

In this section we present two augmented matrices as defined in [27] which incorporate $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$ together with information about their loss of bi-orthogonality.

3.1. Indicators of loss of bi-orthogonality

First we define an alternative way to measure the loss of bi-orthogonality between the Lanczos vectors. Following [27, Eqn. (7.1), Theorem 7.1], we define

$$\begin{aligned} \mathbf{R}_k &:= (\mathbf{I}_k + \mathbf{U}_k)^{-1}\mathbf{U}_k = -\sum_{j=1}^{k-1}(-\mathbf{U}_k)^j, \\ \widehat{\mathbf{R}}_k &:= (\mathbf{I}_k + \widehat{\mathbf{U}}_k)^{-1}\widehat{\mathbf{U}}_k = -\sum_{j=1}^{k-1}(-\widehat{\mathbf{U}}_k)^j, \end{aligned} \tag{11}$$

and

$$\mathbf{r}_{k+1} := (\mathbf{I}_k + \mathbf{U}_k)^{-1}\mathbf{u}_{k+1}, \quad \widehat{\mathbf{r}}_{k+1} := (\mathbf{I}_k + \widehat{\mathbf{U}}_k)^{-1}\widehat{\mathbf{u}}_{k+1}. \tag{12}$$

The second equalities in (11) are based on finite Neumann series expansions possible for nilpotent matrices. Equation (11) shows that $\mathbf{R}_k$ and $\widehat{\mathbf{R}}_k$ are polynomials in $\mathbf{U}_k$ and $\widehat{\mathbf{U}}_k$, respectively, and that they are also strictly upper triangular. By [27, Eqns. (7.2)+(7.3)] the two sets $\{\mathbf{U}_k, \widehat{\mathbf{U}}_k\}$ and $\{\mathbf{R}_k, \widehat{\mathbf{R}}_k\}$ are related by the following rules,

$$\mathbf{U}_k = (\mathbf{I}_k - \mathbf{R}_k)^{-1}\mathbf{R}_k = \mathbf{R}_k(\mathbf{I}_k - \mathbf{R}_k)^{-1}, \quad (\mathbf{I}_k - \mathbf{R}_k)^{-1} = \mathbf{I}_k + \mathbf{U}_k, \tag{13a}$$

$$\widehat{\mathbf{U}}_k = (\mathbf{I}_k - \widehat{\mathbf{R}}_k)^{-1}\widehat{\mathbf{R}}_k = \widehat{\mathbf{R}}_k(\mathbf{I}_k - \widehat{\mathbf{R}}_k)^{-1}, \quad (\mathbf{I}_k - \widehat{\mathbf{R}}_k)^{-1} = \mathbf{I}_k + \widehat{\mathbf{U}}_k. \tag{13b}$$

We can write reverse polynomial relationships based on finite Neumann series as

$$\mathbf{U}_k = \sum_{j=1}^{k-1}\mathbf{R}_k^j, \qquad \widehat{\mathbf{U}}_k = \sum_{j=1}^{k-1}\widehat{\mathbf{R}}_k^j. \tag{14}$$

We see that there is a one-to-one correspondence between these two sets measuring the loss of bi-orthogonality of the Lanczos vectors. From (5) the set $\{\mathbf{U}_k, \widehat{\mathbf{U}}_k\}$ appears to provide the more natural measure, but the set $\{\mathbf{R}_k, \widehat{\mathbf{R}}_k\}$ has been useful, see [27, p. 568], and is the one needed here.

Note that by the strictly upper triangular structure, $\mathbf{r}_{k+1}$ and $\widehat{\mathbf{r}}_{k+1}$ are analogously related to $\mathbf{R}_k$, $\mathbf{R}_{k+1}$ and $\widehat{\mathbf{R}}_k$, $\widehat{\mathbf{R}}_{k+1}$ as $\mathbf{u}_{k+1}$ and $\widehat{\mathbf{u}}_{k+1}$ are related to $\mathbf{U}_k$, $\mathbf{U}_{k+1}$ and $\widehat{\mathbf{U}}_k$, $\widehat{\mathbf{U}}_{k+1}$, i.e.,

$$\mathbf{R}_{k+1} = \begin{pmatrix} \mathbf{R}_k & \mathbf{r}_{k+1} \\ \mathbf{o}^{\mathsf{T}} & 0 \end{pmatrix}, \quad \widehat{\mathbf{R}}_{k+1} = \begin{pmatrix} \widehat{\mathbf{R}}_k & \widehat{\mathbf{r}}_{k+1} \\ \mathbf{o}^{\mathsf{T}} & 0 \end{pmatrix}. \tag{15}$$

3.2. Sheffield augmentation

Let $\mathbf{q}_j \in \mathbb{C}^n$ and $\widehat{\mathbf{q}}_j \in \mathbb{C}^n$, $1 \leqslant j \leqslant k$, denote the columns of the matrices $\mathbf{Q}_k, \widehat{\mathbf{Q}}_k \in \mathbb{C}^{n\times k}$. In [27, §7., Eqn. (7.1), p. 580] the strictly upper triangular matrices $\mathbf{R}_k$ and $\widehat{\mathbf{R}}_k$ are used to define two bi-orthonormal matrices based on $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$:

Definition 3.1 ([27, Eqn. (7.1)])**.** *Let $\mathbf{Q}_k \in \mathbb{C}^{n\times k}$ and $\widehat{\mathbf{Q}}_k \in \mathbb{C}^{n\times k}$ be matrices with $\widehat{\mathbf{q}}_j^{\mathsf{H}}\mathbf{q}_j = 1$ for $1 \leqslant j \leqslant k$. The* Sheffield augmentation *of these two matrices is given by the two matrices $\mathbf{P}^{(k)}, \widehat{\mathbf{P}}^{(k)} \in \mathbb{C}^{(n+k)\times(n+k)}$, defined as*

$$\mathbf{P}^{(k)} := \mathbf{P} := \begin{pmatrix} \mathbf{P}_1 & \mathbf{P}_2 \end{pmatrix} := \begin{pmatrix} \mathbf{R}_k & (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} \\ \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k) & \mathbf{I}_n - \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} \end{pmatrix}, \tag{16a}$$

$$\widehat{\mathbf{P}}^{(k)} := \widehat{\mathbf{P}} := \begin{pmatrix} \widehat{\mathbf{P}}_1 & \widehat{\mathbf{P}}_2 \end{pmatrix} := \begin{pmatrix} \widehat{\mathbf{R}}_k & (\mathbf{I}_k - \widehat{\mathbf{R}}_k)\mathbf{Q}_k^{\mathsf{H}} \\ \widehat{\mathbf{Q}}_k(\mathbf{I}_k - \widehat{\mathbf{R}}_k) & \mathbf{I}_n - \widehat{\mathbf{Q}}_k(\mathbf{I}_k - \widehat{\mathbf{R}}_k)\mathbf{Q}_k^{\mathsf{H}} \end{pmatrix}, \tag{16b}$$

where $\mathbf{R}_k$, $\widehat{\mathbf{R}}_k$ are the polynomials in $\mathbf{U}_k$, $\widehat{\mathbf{U}}_k$ defined by (11) and $\mathbf{U}_k$, $\widehat{\mathbf{U}}_k$ are the strictly upper triangular matrices defined by (5).

We remark again that the case $k > n$ is not excluded.

In [27, Theorem 7.1] it is proven that $\widehat{\mathbf{P}}^{\mathsf{H}}\mathbf{P} = \mathbf{I}_{n+k}$, i.e., that the columns form bi-orthonormal bases of $\mathbb{C}^{n+k}$, $\widehat{\mathbf{P}}^{\mathsf{H}} = \mathbf{P}^{-1}$. For a single set of unit length column vectors, given by a single matrix $\mathbf{Q}_k \in \mathbb{C}^{n\times k}$, in [27] one unitary augmented matrix was defined, which was called a *unitary augmentation* of $\mathbf{Q}_k$. No name was given to the bi-orthonormal counterparts for the case of two matrices $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$, and we suggest $\mathbf{P}^{(k)}$ and $\widehat{\mathbf{P}}^{(k)}$ be called the *Sheffield augmentation* of $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$.

4. Two-sided augmented analysis

Before our main results, we obtain the equivalent of Eqns. (7) for $\mathbf{R}_k$ and $\widehat{\mathbf{R}}_k$, our new measures of loss of bi-orthogonality. This lemma extends [28, Eqn. (3.10)].

Lemma 4.1. *The strictly upper triangular matrices $\mathbf{R}_k$ and $\widehat{\mathbf{R}}_k$ satisfy the two perturbed Hessenberg decompositions*

$$\mathbf{T}_k\mathbf{R}_k - \begin{pmatrix} \mathbf{R}_k & \mathbf{r}_{k+1} \end{pmatrix}\underline{\mathbf{T}}_k = (\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k), \tag{17a}$$

$$\widehat{\mathbf{T}}_k\widehat{\mathbf{R}}_k - \begin{pmatrix} \widehat{\mathbf{R}}_k & \widehat{\mathbf{r}}_{k+1} \end{pmatrix}\underline{\widehat{\mathbf{T}}}_k = (\mathbf{I}_k - \widehat{\mathbf{R}}_k)\widehat{\mathbf{F}}_k(\mathbf{I}_k - \widehat{\mathbf{R}}_k). \tag{17b}$$

Proof. Since $\mathbf{R}_k$ is *strictly* upper triangular, its last row is zero, which implies that we can modify the last column of a matrix to the left without changing the result of the multiplication, e.g.,

$$\mathbf{e}_k^\mathsf{T}\mathbf{R}_k = \mathbf{o}_k^\mathsf{T} \quad \Rightarrow \quad \mathbf{R}_k\mathbf{T}_k\mathbf{R}_k = \begin{pmatrix}\mathbf{R}_k & \mathbf{r}_{k+1}\end{pmatrix}\underline{\mathbf{T}}_k\mathbf{R}_k. \tag{18}$$

We restate the Hessenberg decomposition (7a) for $\mathbf{U}_k$,

$$\mathbf{T}_k\mathbf{U}_k - \begin{pmatrix}\mathbf{U}_k & \mathbf{u}_{k+1}\end{pmatrix}\underline{\mathbf{T}}_k = \mathbf{F}_k.$$

Replacing $\mathbf{U}_k$ and $\mathbf{u}_{k+1}$ using (13a) and (12) gives

$$\mathbf{T}_k\mathbf{R}_k(\mathbf{I}_k - \mathbf{R}_k)^{-1} - \begin{pmatrix}(\mathbf{I}_k - \mathbf{R}_k)^{-1}\mathbf{R}_k & (\mathbf{I}_k - \mathbf{R}_k)^{-1}\mathbf{r}_{k+1}\end{pmatrix}\underline{\mathbf{T}}_k = \mathbf{F}_k.$$

Multiplication by the unit upper triangular matrix $\mathbf{I}_k - \mathbf{R}_k$ from the left and the right results in

$$(\mathbf{I}_k - \mathbf{R}_k)\mathbf{T}_k\mathbf{R}_k - \begin{pmatrix}\mathbf{R}_k & \mathbf{r}_{k+1}\end{pmatrix}\underline{\mathbf{T}}_k(\mathbf{I}_k - \mathbf{R}_k) = (\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k),$$

which, when added to (18), yields the perturbed Hessenberg decomposition (17a). The corresponding "hatted" variant (17b) follows completely analogously. □

From (17a) it immediately follows that

$$(\mathbf{I}_k - \mathbf{R}_k)\mathbf{T}_k = \mathbf{T}_k(\mathbf{I}_k - \mathbf{R}_k) + \mathbf{r}_{k+1}\beta_{k+1}\mathbf{e}_k^\mathsf{T} + (\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k). \tag{19}$$

We refer to both equation (17a) and equation (19) in the proof of Theorem 4.1. The following observations will also be useful:

$$\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{q}_k = \widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{Q}_k\mathbf{e}_k = (\widehat{\mathbf{U}}_k^\mathsf{H} + \mathbf{I}_k + \mathbf{U}_k)\mathbf{e}_k = (\mathbf{I}_k + \mathbf{U}_k)\mathbf{e}_k,$$

so that from the relation (13a) between $\mathbf{R}_k$ and $\mathbf{U}_k$

$$(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{q}_k = \mathbf{e}_k. \qquad (\text{Similarly } (\mathbf{I}_k - \widehat{\mathbf{R}}_k)\mathbf{Q}_k^\mathsf{H}\widehat{\mathbf{q}}_k = \mathbf{e}_k.) \tag{20}$$

From (12), (13a) and (6) we have

$$\mathbf{r}_{k+1} = (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{q}_{k+1}. \qquad (\text{Similarly } \widehat{\mathbf{r}}_{k+1} = (\mathbf{I}_k - \widehat{\mathbf{R}}_k)\mathbf{Q}_k^\mathsf{H}\widehat{\mathbf{q}}_{k+1}.) \tag{21}$$

It will also help to define the matrix update $\mathbf{A}_k$ of $\mathbf{A}$ of rank at most two by

$$\mathbf{A}_k := \mathbf{A} - \mathbf{q}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^\mathsf{H} - \mathbf{q}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^\mathsf{H}. \tag{22}$$

It then follows from (21) and (20) that

$$\begin{aligned}(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{A}_k &\\ = (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{A} - (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{q}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^\mathsf{H} - (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{q}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^\mathsf{H}&\\ = (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^\mathsf{H}\mathbf{A} - \mathbf{r}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^\mathsf{H} - \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^\mathsf{H}.&\end{aligned} \tag{23}$$

We now prove our main theorem.

Theorem 4.1. *For general* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *consider a perturbed two-sided Lanczos process captured by the Lanczos recurrences* (1) *with the tridiagonal matrices defined by equations* (2)*. The matrices* $\mathbf{U}_k$ *and* $\widehat{\mathbf{U}}_k$ *are defined in equations* (5)*,* $\mathbf{u}_{k+1}$ *and* $\widehat{\mathbf{u}}_{k+1}$ *in equations* (6)*,* $\mathbf{F}_k$ *and* $\widehat{\mathbf{F}}_k$ *in equations* (7)*,* $\mathbf{R}_k$ *and* $\widehat{\mathbf{R}}_k$ *in equations* (11)*, and the Sheffield augmentation* $\mathbf{P}_k$ *and* $\widehat{\mathbf{P}}_k$ *in equations* (16)*. Then with* $\mathbf{A}_k$ *in* (22)*,*

$$\widehat{\mathbf{P}}^{\mathsf{H}} \left(\mathsf{diag}(\mathbf{T}_k, \mathbf{A}) + \mathbf{H}\right) \mathbf{P} = \begin{pmatrix} \mathbf{T}_k & \mathbf{e}_k \overline{\widehat{\beta}_{k+1}} \widehat{\mathbf{q}}_{k+1}^{\mathsf{H}} \\ \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} & \mathbf{A}_k \end{pmatrix}, \tag{24}$$

where

$$\mathbf{H} := \mathbf{N}_k(\mathbf{F}_k - \widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k)\widehat{\mathbf{N}}_k^{\mathsf{H}} - \begin{pmatrix} \mathbf{O}_{k,k} \\ \mathbf{E}_k \end{pmatrix} \widehat{\mathbf{N}}_k^{\mathsf{H}} - \mathbf{N}_k \begin{pmatrix} \mathbf{O}_{k,k} & \widehat{\mathbf{E}}_k^{\mathsf{H}} \end{pmatrix} \tag{25}$$

with

$$\mathbf{N}_k := \begin{pmatrix} \mathbf{I}_k \\ -\mathbf{Q}_k \end{pmatrix} (\mathbf{I}_k - \mathbf{R}_k) = \begin{pmatrix} \mathbf{I}_k \\ \mathbf{O}_{n,k} \end{pmatrix} - \mathbf{P}_1, \tag{26a}$$

$$\widehat{\mathbf{N}}_k := \begin{pmatrix} \mathbf{I}_k \\ -\widehat{\mathbf{Q}}_k \end{pmatrix} (\mathbf{I}_k - \widehat{\mathbf{R}}_k) = \begin{pmatrix} \mathbf{I}_k \\ \mathbf{O}_{n,k} \end{pmatrix} - \widehat{\mathbf{P}}_1. \tag{26b}$$

Proof. Define

$$\mathbf{G} := \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix} := \begin{pmatrix} \mathbf{T}_k & \mathbf{O}_{k,n} \\ \mathbf{O}_{n,k} & \mathbf{A} \end{pmatrix} \mathbf{P} - \mathbf{P} \begin{pmatrix} \mathbf{T}_k & \mathbf{e}_k \overline{\widehat{\beta}_{k+1}} \widehat{\mathbf{q}}_{k+1}^{\mathsf{H}} \\ \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} & \mathbf{A}_k \end{pmatrix}. \tag{27}$$

We first find expressions for the components of $\mathbf{G}$ and then show that relation (24) is satisfied with

$$\mathbf{H} = -\mathbf{G}\widehat{\mathbf{P}}^{\mathsf{H}}, \qquad \mathbf{H}\mathbf{P} = -\mathbf{G}. \tag{28}$$

Just as in [28], the proof is based on consideration of the four blocks of $\mathbf{G}$ separately.

The first block $\mathbf{G}_{11}$ can be rewritten utilizing the definition of $\mathbf{P}$ in equation (16a), using (21), and utilizing the new Hessenberg decomposition (17a), respectively,

$$\begin{aligned} \mathbf{G}_{11} &= \mathbf{T}_k\mathbf{R}_k - \mathbf{R}_k\mathbf{T}_k - (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} \\ &= \mathbf{T}_k\mathbf{R}_k - \mathbf{R}_k\mathbf{T}_k - \mathbf{r}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} \\ &= \mathbf{T}_k\mathbf{R}_k - \begin{pmatrix} \mathbf{R}_k & \mathbf{r}_{k+1} \end{pmatrix} \underline{\mathbf{T}}_k = (\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k). \end{aligned} \tag{29}$$

The second block $\mathbf{G}_{21}$ is rewritten similarly. Using the definition of $\mathbf{P}$ (16a), the additional Hessenberg decomposition (19), substituting (21), observing that $\mathbf{e}_k^{\mathsf{T}}\mathbf{R}_k = \mathbf{o}^{\mathsf{T}}$, see (18), and finally using the right Lanczos recurrence (1a) gives:

$$\begin{aligned} \mathbf{G}_{21} &= \mathbf{A}\mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k) - \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\mathbf{T}_k \\ &\quad - \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} + \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} \\ &= \mathbf{A}\mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k) \\ &\quad - \mathbf{Q}_k\mathbf{T}_k(\mathbf{I}_k - \mathbf{R}_k) - \mathbf{Q}_k\mathbf{r}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} - \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k) \\ &\quad - \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} \qquad + \mathbf{Q}_k\mathbf{r}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} \\ &= (\mathbf{A}\mathbf{Q}_k - \mathbf{Q}_k\mathbf{T}_k - \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}})(\mathbf{I}_k - \mathbf{R}_k) - \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k) \\ &= -\Big(\mathbf{E}_k + \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k\Big)(\mathbf{I}_k - \mathbf{R}_k). \end{aligned} \tag{30}$$

We consider the third block $\mathbf{G}_{12}$. By the definition of $\mathbf{P}$ in Eqn. (16a) the first line of Eqn. (31) follows. Using the Hessenberg decomposition (19), the expansion (23), and finally the complex conjugate transpose of (1b), gives:

$$\begin{aligned}
\mathbf{G}_{12} &= \mathbf{T}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \mathbf{R}_k\mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}} - (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{A}_k \\
&= (\mathbf{I}_k - \mathbf{R}_k)\mathbf{T}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \mathbf{r}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}}\widehat{\mathbf{Q}}_k^{\mathsf{H}} - (\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} \\
&\quad - \mathbf{R}_k\mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}} - (\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{A} + \mathbf{r}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^{\mathsf{H}} + \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}} \\
&= (\mathbf{I}_k - \mathbf{R}_k)\Big(\mathbf{T}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{A} + \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}}\Big) \\
&\quad - (\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} \\
&= (\mathbf{I}_k - \mathbf{R}_k)\Big(\widehat{\mathbf{E}}_k^{\mathsf{H}} - \mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\Big).
\end{aligned} \tag{31}$$

We develop $\mathbf{G}_{22}$ in steps. By the definition of $\mathbf{P}$ in Eqn. (16a), $\mathbf{G}_{22}$ is given by

$$\mathbf{A} - \mathbf{A}\mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}} - \mathbf{A}_k + \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{A}_k.$$

Using (22) to give $\mathbf{A} - \mathbf{A}_k$, then (23), gives

$$\begin{aligned}
\mathbf{G}_{22} &= \mathbf{q}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^{\mathsf{H}} + \mathbf{q}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}} - \mathbf{A}\mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} \\
&\quad + \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)(\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{A} - \mathbf{e}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}}) - \mathbf{Q}_k\mathbf{r}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^{\mathsf{H}} - \mathbf{q}_k\overline{\widehat{\beta}_{k+1}}\widehat{\mathbf{q}}_{k+1}^{\mathsf{H}}.
\end{aligned}$$

Cancelling, rearranging, using (1a) and (1b)$^{\mathsf{H}}$ with $\mathbf{e}_k^{\mathsf{T}}\mathbf{R}_k = \mathbf{o}$, then (19), gives

$$\begin{aligned}
\mathbf{G}_{22} &= \mathbf{q}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^{\mathsf{H}} - (\mathbf{Q}_k\mathbf{T}_k + \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}} - \mathbf{E}_k)(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} \\
&\quad + \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)(\mathbf{T}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \widehat{\mathbf{E}}_k^{\mathsf{H}}) - \mathbf{Q}_k\mathbf{r}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^{\mathsf{H}} \\
&= -\mathbf{Q}_k\mathbf{T}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} + \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\mathbf{T}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \mathbf{Q}_k\mathbf{r}_{k+1}\beta_{k+1}\widehat{\mathbf{q}}_k^{\mathsf{H}} \\
&\quad + \mathbf{E}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{E}}_k^{\mathsf{H}} \\
&= \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} + \mathbf{E}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}} - \mathbf{Q}_k(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{E}}_k^{\mathsf{H}}.
\end{aligned}$$

With the definition of $\mathbf{N}_k$ in (26a), by examining the individual submatrices on the right below, it can be seen that $\mathbf{G}$ can be expressed as a sum of three terms as follows:

$$\begin{aligned}
\mathbf{G} = \mathbf{N}_k\mathbf{F}_k(\mathbf{I}_k - \mathbf{R}_k)\begin{pmatrix}\mathbf{I}_k & -\widehat{\mathbf{Q}}_k^{\mathsf{H}}\end{pmatrix} \\
- \begin{pmatrix}\mathbf{O}_{k,k} \\ \mathbf{E}_k\end{pmatrix}(\mathbf{I}_k - \mathbf{R}_k)\begin{pmatrix}\mathbf{I}_k & -\widehat{\mathbf{Q}}_k^{\mathsf{H}}\end{pmatrix} + \mathbf{N}_k\begin{pmatrix}\mathbf{O}_{k,k} & \widehat{\mathbf{E}}_k^{\mathsf{H}}\end{pmatrix}.
\end{aligned} \tag{32}$$

From (16a),

$$(\mathbf{I}_k - \mathbf{R}_k)\begin{pmatrix}\mathbf{I}_k & -\widehat{\mathbf{Q}}_k^{\mathsf{H}}\end{pmatrix} = \begin{pmatrix}\mathbf{I}_k - \mathbf{R}_k & -(\mathbf{I}_k - \mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\end{pmatrix} = \begin{pmatrix}\mathbf{I}_k & \mathbf{O}_{k,n}\end{pmatrix}(\mathbf{I}_{k+n} - \mathbf{P}),$$

so that with (16b)

$$(\mathbf{I}_k - \mathbf{R}_k)\begin{pmatrix}\mathbf{I}_k & -\widehat{\mathbf{Q}}_k^{\mathsf{H}}\end{pmatrix}\widehat{\mathbf{P}}^{\mathsf{H}} = \begin{pmatrix}\mathbf{I}_k & \mathbf{O}_{k,n}\end{pmatrix}(\widehat{\mathbf{P}}^{\mathsf{H}} - \mathbf{I}_{k+n}) = -\widehat{\mathbf{N}}_k^{\mathsf{H}}. \tag{33}$$

Also,

$$\begin{aligned}\begin{pmatrix}\mathbf{O}_{k,k} & \widehat{\mathbf{E}}_k^{\mathsf{H}}\end{pmatrix}\widehat{\mathbf{P}}^{\mathsf{H}} &= \begin{pmatrix}\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k(\mathbf{I}_k-\widehat{\mathbf{R}}_k^{\mathsf{H}}) & \widehat{\mathbf{E}}_k^{\mathsf{H}}-\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k(\mathbf{I}_k-\widehat{\mathbf{R}}_k^{\mathsf{H}})\widehat{\mathbf{Q}}_k^{\mathsf{H}}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{O}_{k,k} & \widehat{\mathbf{E}}_k^{\mathsf{H}}\end{pmatrix}+\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k\widehat{\mathbf{N}}_k^{\mathsf{H}}. \end{aligned} \tag{34}$$

Multiplying (32) on the right by $\widehat{\mathbf{P}}^{\mathsf{H}}$ and using (33) and (34) gives (25). □

If we define $\widehat{\mathbf{H}}$ by replacing every quantity in the definition of $\mathbf{H}$ in Eqn. (25) by its hatted variant, where $\widehat{\mathbf{O}}_{k,k}=\mathbf{O}_{k,k}$ and $\widehat{\mathbf{I}}_k=\mathbf{I}_k$, then we will show that $\mathbf{H}^{\mathsf{H}}=\widehat{\mathbf{H}}$, which we call "hat"-symmetry. The sum of the last two terms in the sum (25) is obviously "hat"-symmetric, so we need only show that $(\mathbf{F}_k-\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k)^{\mathsf{H}}=\widehat{\mathbf{F}}_k-\mathbf{E}_k^{\mathsf{H}}\widehat{\mathbf{Q}}_k$. The off-diagonal part of this equation is satisfied, since with Eqns. (7)

$$\begin{aligned}\mathbf{F}_k-\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k &= \mathbf{D}_k-\mathsf{tril}(\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k,-1)-\mathsf{triu}(\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{E}_k,1)-\mathsf{diag}(\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{E}_k),\\ \widehat{\mathbf{F}}_k-\mathbf{E}_k^{\mathsf{H}}\widehat{\mathbf{Q}}_k &= \widehat{\mathbf{D}}_k-\mathsf{tril}(\mathbf{E}_k^{\mathsf{H}}\widehat{\mathbf{Q}}_k,-1)-\mathsf{triu}(\mathbf{Q}_k^{\mathsf{H}}\widehat{\mathbf{E}}_k,1)-\mathsf{diag}(\mathbf{Q}_k^{\mathsf{H}}\widehat{\mathbf{E}}_k).\end{aligned}$$

It remains to prove that

$$\mathbf{D}_k-\mathsf{diag}(\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{E}_k)=(\widehat{\mathbf{D}}_k-\mathsf{diag}(\mathbf{Q}_k^{\mathsf{H}}\widehat{\mathbf{E}}_k))^{\mathsf{H}},$$

i.e., with $u_{0,1}:=0$ and (8), we have to prove that for $1\leqslant j\leqslant k$,

$$\overline{\widehat{u}_{j-1,j}\widehat{\beta}_j}-\overline{\widehat{u}_{j,j+1}\widehat{\beta}_{j+1}}-\widehat{\mathbf{q}}_j^{\mathsf{H}}\mathbf{E}_k\mathbf{e}_j=u_{j-1,j}\beta_j-u_{j,j+1}\beta_{j+1}-\mathbf{e}_j^{\mathsf{T}}\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{q}_j.$$

But this follows from the diagonal of (9) by using $\widehat{\mathbf{T}}_k^{\mathsf{H}}=\mathbf{T}_k$. Therefore $\mathbf{H}^{\mathsf{H}}=\widehat{\mathbf{H}}$.

Corollary 4.1. *With the assumptions and expressions for* $\mathbf{H}$, $\mathbf{N}_k$ *and* $\widehat{\mathbf{N}}_k$ *in Theorem 4.1 and with* $\mathbf{r}_{k+1}$, $\widehat{\mathbf{r}}_{k+1}$ *in* (12), *we have*

$$\big(\mathsf{diag}(\mathbf{T}_k,\mathbf{A})+\mathbf{H}\big)\mathbf{P}_1=\mathbf{P}_1\mathbf{T}_k+\mathbf{p}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}}, \tag{35a}$$

$$\big(\mathsf{diag}(\mathbf{T}_k,\mathbf{A})+\mathbf{H}\big)^{\mathsf{H}}\widehat{\mathbf{P}}_1=:\big(\mathsf{diag}(\widehat{\mathbf{T}}_k,\widehat{\mathbf{A}})+\widehat{\mathbf{H}}\big)\widehat{\mathbf{P}}_1=\widehat{\mathbf{P}}_1\widehat{\mathbf{T}}_k+\widehat{\mathbf{p}}_{k+1}\widehat{\beta}_{k+1}\mathbf{e}_k^{\mathsf{T}}, \tag{35b}$$

where

$$\mathbf{p}_{k+1}=\begin{pmatrix}\mathbf{r}_{k+1}\\ \mathbf{q}_{k+1}-\mathbf{Q}_k\mathbf{r}_{k+1}\end{pmatrix},\qquad \widehat{\mathbf{p}}_{k+1}=\begin{pmatrix}\widehat{\mathbf{r}}_{k+1}\\ \widehat{\mathbf{q}}_{k+1}-\widehat{\mathbf{Q}}_k\widehat{\mathbf{r}}_{k+1}\end{pmatrix}. \tag{36}$$

The vectors $\mathbf{p}_{k+1}$ *and* $\widehat{\mathbf{p}}_{k+1}$ *are also the* $(k+1)$*st columns of* $\mathbf{P}^{(k+1)}$ *and* $\widehat{\mathbf{P}}^{(k+1)}$ *respectively with the* $(k+1)$*st zero element removed.*

Proof. From (21) $\mathbf{r}_{k+1}=(\mathbf{I}_k-\mathbf{R}_k)\widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{q}_{k+1}$ and $\widehat{\mathbf{r}}_{k+1}=(\mathbf{I}_k-\widehat{\mathbf{R}}_k)\mathbf{Q}_k^{\mathsf{H}}\widehat{\mathbf{q}}_{k+1}$. Equation (35a), with $\mathbf{p}_{k+1}$ in (36), follows by multiplying (24) on the left by $\mathbf{P}$. Analogously, equation (35b), with $\widehat{\mathbf{p}}_{k+1}$ in (36), follows by applying the $^{\mathsf{H}}$ operator on (24) and then multiplying on the left by $\widehat{\mathbf{P}}$.

Finally from (16a) and (15), the $(k+1)$st column of $\mathbf{P}^{(k+1)}$ is given by

$$\mathbf{P}_1^{(k+1)}\mathbf{e}_{k+1}=\begin{pmatrix}\mathbf{R}_{k+1}\\ \mathbf{Q}_{k+1}(\mathbf{I}_{k+1}-\mathbf{R}_{k+1})\end{pmatrix}\mathbf{e}_{k+1}=\begin{pmatrix}\mathbf{r}_{k+1}\\ 0\\ \mathbf{q}_{k+1}-\mathbf{Q}_k\mathbf{r}_{k+1}\end{pmatrix}.$$

This shows that $\mathbf{p}_{k+1}$ is the $(k+1)$st column of $\mathbf{P}^{(k+1)}$ with the $(k+1)$st zero element removed; the proof is similar for $\widehat{\mathbf{p}}_{k+1}$. □

Corollary 4.1 indicates that running k steps of the Lanczos process on $\mathbf{A} \in \mathbb{C}^{n\times n}$ in the presence of perturbations is equivalent to running k steps of an exact Lanczos process on a perturbation of the augmented matrix $\mathsf{diag}\,(\mathbf{T}_k, \mathbf{A})$. More precisely, if $\mathbf{P}_{1,-k}$ denotes $\mathbf{P}_1$ in (16) without its zero kth row, $\mathbf{H}_{-k}$ denotes $\mathbf{H}$ in (25) without its kth column, and similarly for $\widehat{\mathbf{P}}_1$ and $\widehat{\mathbf{H}} = \mathbf{H}^{\mathsf{H}}$, then (35a) and (35b) can be rewritten

$$\left(\begin{pmatrix} \underline{\mathbf{T}}_{k-1} & \mathbf{O}_{k,n} \\ \mathbf{O}_{n,k-1} & \mathbf{A} \end{pmatrix} + \mathbf{H}_{-k}\right) \mathbf{P}_{1,-k} = (\mathbf{P}_1, \mathbf{p}_{k+1})\underline{\mathbf{T}}_k, \tag{37a}$$

$$\left(\begin{pmatrix} \underline{\widehat{\mathbf{T}}}_{k-1} & \mathbf{O}_{k,n} \\ \mathbf{O}_{n,k-1} & \mathbf{A}^{\mathsf{H}} \end{pmatrix} + \widehat{\mathbf{H}}_{-k}\right) \widehat{\mathbf{P}}_{1,-k} = (\widehat{\mathbf{P}}_1, \widehat{\mathbf{p}}_{k+1})\underline{\widehat{\mathbf{T}}}_k. \tag{37b}$$

This shows, for example, how both $\mathbf{A}$ and $\underline{\mathbf{T}}_{k-1}$ contribute to forming $\mathbf{p}_{k+1}$ and the new column of $\underline{\mathbf{T}}_k$. This contribution by $\underline{\mathbf{T}}_{k-1}$ is an indication that the Lanczos process loses memory and is likely to recompute eigenvalues already found at previous iterations. In an exact (i.e., unperturbed) process we have $\mathbf{H} = \mathbf{O}$ while (35a) and (35b) reduce to (1a) and (1b), except for the extra k by k zero blocks, with $\mathbf{E}_k = \mathbf{O}$ and no loss of bi-orthogonality. Also, if we only assume that bi-orthogonality is perfectly preserved, then $\mathbf{U}_k$, $\mathbf{u}_{k+1}$, $\mathbf{R}_k$, $\mathbf{r}_{k+1}$, $\mathbf{D}_k$ and the corresponding hatted variants are all zero, $\mathbf{P}_1$, $\mathbf{p}_{k+1}$, and the corresponding hatted variants in Eqns. (16) and Eqns. (36) collapse to

$$\mathbf{P}_1 = \begin{pmatrix} \mathbf{O}_k \\ \mathbf{Q}_k \end{pmatrix}, \quad \widehat{\mathbf{P}}_1 = \begin{pmatrix} \mathbf{O}_k \\ \widehat{\mathbf{Q}}_k \end{pmatrix}, \quad \mathbf{p}_{k+1} = \begin{pmatrix} \mathbf{o}_k \\ \mathbf{q}_{k+1} \end{pmatrix}, \quad \widehat{\mathbf{p}}_{k+1} = \begin{pmatrix} \mathbf{o}_k \\ \widehat{\mathbf{q}}_{k+1} \end{pmatrix}. \tag{38}$$

Thus, since $\mathbf{HP} = -\mathbf{G}$ from (28), and by definition of $\widehat{\mathbf{H}}$, the leading blocks of Eqns. (35) reduce with (29) to

$$\mathbf{H}_{1,2}\mathbf{Q}_k = -\mathbf{G}_{11} = -\mathbf{F}_k = \mathbf{O}_k, \quad \mathbf{H}_{2,1}^{\mathsf{H}}\widehat{\mathbf{Q}}_k = \widehat{\mathbf{H}}_{1,2}\widehat{\mathbf{Q}}_k = -\widehat{\mathbf{F}}_k = \mathbf{O}_k, \tag{39}$$

with $\mathbf{H}_{1,2}$ the upper right n by k block of $\mathbf{H}$ and $\mathbf{H}_{2,1}$ the lower left k by n block of $\mathbf{H}$, similar for the hatted variants. This is trivially true, as both $\mathbf{F}_k$ and $\widehat{\mathbf{F}}_k$ are zero by Eqns. (7). The trailing blocks of Eqns. (35) imply a standard backward error result

$$\begin{aligned}
(\mathbf{A} + \mathbf{H}_{2,2})\ \mathbf{Q}_k &= (\mathbf{A} \ + \mathbf{E}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}} + \mathbf{Q}_k\widehat{\mathbf{E}}_k^{\mathsf{H}} - \mathbf{Q}_k\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}})\mathbf{Q}_k = \mathbf{Q}_k\mathbf{T}_k + \mathbf{q}_{k+1}\beta_{k+1}\mathbf{e}_k^{\mathsf{T}}, \\
(\mathbf{A} + \mathbf{H}_{2,2})^{\mathsf{H}}\widehat{\mathbf{Q}}_k &= (\mathbf{A}^{\mathsf{H}} + \widehat{\mathbf{E}}_k\mathbf{Q}_k^{\mathsf{H}} + \widehat{\mathbf{Q}}_k\mathbf{E}_k^{\mathsf{H}} - \widehat{\mathbf{Q}}_k\mathbf{E}_k^{\mathsf{H}}\widehat{\mathbf{Q}}_k\mathbf{Q}_k^{\mathsf{H}})\widehat{\mathbf{Q}}_k = \widehat{\mathbf{Q}}_k\widehat{\mathbf{T}}_k + \widehat{\mathbf{q}}_{k+1}\widehat{\beta}_{k+1}\mathbf{e}_k^{\mathsf{T}},
\end{aligned}$$

with $\mathbf{H}_{2,2} = \mathbf{E}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}} + \mathbf{Q}_k\widehat{\mathbf{E}}_k^{\mathsf{H}} - \mathbf{Q}_k\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k\widehat{\mathbf{Q}}_k^{\mathsf{H}} = (\widehat{\mathbf{E}}_k\mathbf{Q}_k^{\mathsf{H}} + \widehat{\mathbf{Q}}_k\mathbf{E}_k^{\mathsf{H}} - \widehat{\mathbf{Q}}_k\mathbf{E}_k^{\mathsf{H}}\widehat{\mathbf{Q}}_k\mathbf{Q}_k^{\mathsf{H}})^{\mathsf{H}} = \widehat{\mathbf{H}}_{2,2}^{\mathsf{H}}$ the lowest right n by n block of $\mathbf{H}$. The "hat"-symmetry follows again from (9), which implies that in case of bi-orthogonality, $\widehat{\mathbf{E}}_k^{\mathsf{H}}\mathbf{Q}_k = \widehat{\mathbf{Q}}_k^{\mathsf{H}}\mathbf{E}_k$. The latter is a strong assumption on the structure of the perturbations: the left and right perturbations have to be highly structured to result in preservation of bi-orthogonality.

Theorem 4.1 and Corollary 4.1 extend Theorem 3.1 and Corollary 3.2 of [28] to the two-sided Lanczos process. If one removes the hats from (35b), then (35a) and (35b) become the same relation, identical to [28, Equation 3.24]. The two sets of results are analogous except for the bounds on $\mathbf{H}$ that we now discuss for the case of perturbations caused by execution in finite precision arithmetic.

5. Round-off error bounds

For the Hermitian (symmetric) application of the Lanczos process, it was shown in [25] that $\|\mathbf{E}_k\|_{2,F}, \|\mathbf{F}_k\|_{2,F} \leqslant \|\mathbf{A}\|_{2,F} O(\epsilon)$, and in [27] and [28] that $\|\mathbf{R}_k\|_2 \leqslant 1$, $\|\mathbf{N}_k\|_2 \leqslant 2$. Moreover, the columns of $\mathbf{Q}_k$ are essentially unit length (up to rounding error). These bounds were used in [28] to show that for the Hermitian Lanczos process

$$\|\mathbf{H}\|_{2,F} \leqslant 4\left(\|\mathbf{E}_k\|_{2,F} + \|\mathbf{F}_k\|_{2,F}\right) \leqslant \|\mathbf{A}\|_{2,F} O(\epsilon),$$

giving a kind of augmented stability result.

For the two sided Lanczos process, we are unable to produce clear a priori bounds on the size of $\mathbf{H}$. In the absence of breakdown of the process, in [3, Theorem 3.1] it was shown for the implementation described there that the rounding error matrices $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$ satisfy the componentwise relations

$$\begin{aligned}
\left|\mathbf{E}_k\right| &\leqslant (3+m)\epsilon|\mathbf{A}||\mathbf{Q}_k| + 4\epsilon|\mathbf{Q}_k||\mathbf{T}_k| + O(\epsilon^2), \\
\left|\widehat{\mathbf{E}}_k\right| &\leqslant (3+m)\epsilon|\mathbf{A}^{\mathsf{H}}|\widehat{\mathbf{Q}}_k| + 4\epsilon|\widehat{\mathbf{Q}}_k||\widehat{\mathbf{T}}_k| + O(\epsilon^2),
\end{aligned}$$

where m refers to the maximum number of non-zero elements per row of $\mathbf{A}$. Bounds of a similar size to these can be found for the elements of $\mathbf{D}_k$ and $\widehat{\mathbf{D}}_k$ in (8). These bounds depend on the sizes of $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$. Similary, the sizes of $\mathbf{F}_k$, $\widehat{\mathbf{F}}_k$, $\mathbf{N}_k$, $\widehat{\mathbf{N}}_k$, $\mathbf{R}_k$, $\widehat{\mathbf{R}}_k$ and hence the size of $\mathbf{H}$, and values of $\mathbf{p}_{k+1}$ and $\widehat{\mathbf{p}}_{k+1}$ depend on $\mathbf{Q}_k$ and $\widehat{\mathbf{Q}}_k$, whose columns may grow unboundedly when the algorithm reaches a state of near breakdown. For that reason we believe that it is impossible, at least in general, to produce an a priori bound on $\mathbf{H}$. Hence the two-sided Lanczos process is not numerically "stable" in the augmented sense that the Hermitian Lanczos process was shown to be.

Nevertheless, for any specific computation, one may keep track of the sizes of $\mathbf{Q}_k$, $\widehat{\mathbf{Q}}_k$, for example by accumulating their Frobenius norms as suggested by [3, p. 221]. Moreover, it is possible to compute $\mathbf{R}_k$, $\widehat{\mathbf{R}}_k$, $\mathbf{r}_{k+1}$, $\widehat{\mathbf{r}}_{k+1}$ essentially exactly when $k \ll n$. All this can be done without too much computational overhead. Therefore, for a specific problem instance $\mathbf{H}$ can be bounded a posteriori, and the vectors $\mathbf{p}_{k+1}$ and $\widehat{\mathbf{p}}_{k+1}$ can be calculated essentially exactly (up to small rounding errors).

Such a computational procedure is only feasible for reasonably small k, and will not often be practical. But since $\mathbf{H}$ is nicely bounded in the Hermitian case, it might be possible to obtain bounds in the general case in terms of the distance of $\mathbf{A}$ from being Hermitian, or perhaps in terms of other matrix properties. This would be useful for the theoretical understanding of the process, or of practical use for certain classes of matrices. For general applications we would like to obtain practical computable bounds on $\mathbf{H}$ for all k, but this will take time, since unlike the Hermitian case, we at present have very little understanding of the matrices $\mathbf{P}^{(k)}$ and $\widehat{\mathbf{P}}^{(k)}$ in Eqns. (16).

For inexact methods we need similar developments to the above in order to understand how $\mathbf{H}$ develops in terms of general $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$ in Eqns. (1).

6. Conclusion

In this paper we extended an augmented backward rounding error result, proven in [28] for the symmetric Lanczos process, to the two-sided Lanczos process subject to

perturbations such as those caused by finite precision arithmetic or inexact methods. Our analysis shows that the form of the results (24) and (35) for the two sided Lanczos process are completely analogous to those for the symmetric Lanczos process. We are unable in general to derive tight a priori bounds for the backward perturbation term $\mathbf{H}$ in (24) and (25), and believe that this might grow unboundedly in cases of near breakdown. However, since in the case of finite precision computations with real symmetric $\mathbf{A}$ this term is always bounded proportionally to the introduced rounding error $\mathbf{E}_k$ in (1a), it seems possible that in the general case a bound could be found in terms of the introduced errors $\mathbf{E}_k$ and $\widehat{\mathbf{E}}_k$ in Eqns. (1) and the distance from symmetry of $\mathbf{A}$, and that practical a posteriori measures could be computed.

References

[1] J.I. Aliaga, D.L. Boley, R.W. Freund, V. Hernández, A Lanczos-type method for multiple starting vectors, Math. Comp. (2000) 1577–1601.

[2] W.E. Arnoldi, The principle of minimized iteration in the solution of the matrix eigenvalue problem, Quart. Appl. Math. 9 (1951) 17–29.

[3] Z. Bai, Error analysis of the Lanczos algorithm for the nonsymmetric eigenvalue problem, Math. Comput. 62 (1994) 209–226.

[4] Z. Bai, D. Day, Q. Ye, ABLE: an adaptive block Lanczos method for non-Hermitian eigenvalue problems, SIAM J. Matrix Anal. Appl. 20 (1999) 1060–1082. Sparse and structured matrices and their applications (Coeur d'Alene, ID, 1996).

[5] J.L. Barlow, N. Bosner, Z. Drmač, A new stable bidiagonal reduction algorithm, Linear Algebra Appl. 397 (2005) 35–84.

[6] Å. Björck, C.C. Paige, Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm, SIAM J. Matrix Anal. Appl. 13 (1992) 176–190.

[7] Å. Björck, C.C. Paige, Solution of augmented linear systems using orthogonal factorizations, BIT 34 (1994) 1–24.

[8] D. Day, An efficient implementation of the nonsymmetric Lanczos algorithm, SIAM J. Matrix Anal. Appl. 18 (1997) 566–589.

[9] D.M. Day, III, Semi-duality in the two-sided Lanczos algorithm, Ph.D. thesis, University of California, Berkeley, Berkeley, 1993.

[10] M.B. van Gijzen, P. Sonneveld, Algorithm 913: An elegant IDR(s) variant that efficiently exploits biorthogonality properties, ACM TOMS 38 (2011).

[11] G.H. Golub, D.P. O'Leary, Some history of the conjugate gradient and Lanczos algorithms: 1948–1976, SIAM Rev. 31 (1989) 50–102.

[12] G.H. Golub, R. Underwood, The block Lanczos method for computing eigenvalues, in: Mathematical software, III (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1977), Academic Press, New York, 1977, pp. 361–377. Publ. Math. Res. Center, No. 39.

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

[14] A. Greenbaum, Z. Strakoš, Predicting the behavior of finite precision Lanczos and conjugate gradient computations, SIAM J. Matrix Anal. Appl. 13 (1992) 121–137.

[15] M.H. Gutknecht, J.P.M. Zemke, Eigenvalue computations based on IDR, Bericht 145, TUHH, Institute of Numerical Simulation, 2010. Online available at http://doku.b.tu-harburg.de/volltexte/2010/875/.

[16] M.R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, J. Research Nat. Bur. Standards 49 (1952) 409–436 (1953).

[17] N.J. Higham, Accuracy and stability of numerical algorithms, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[18] L. Komzsik, The Lanczos method, volume 15 of *Software, Environments, and Tools*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003. Evolution and application.

[19] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Research Nat. Bur. Standards 45 (1950) 255–282.

[20] C. Lanczos, Solution of systems of linear equations by minimized iterations, J. Research Nat. Bur. Standards 49 (1952) 33–53.

[21] G. Meurant, The Lanczos and conjugate gradient algorithms, volume 19 of *Software, Environments, and Tools*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. From theory to finite precision computations.
[22] G. Meurant, Z. Strakoš, The Lanczos and conjugate gradient algorithms in finite precision arithmetic, Acta Numer. 15 (2006) 471–542.
[23] C.C. Paige, The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices, Ph.D. thesis, London University Institute of Computer Science, 1971.
[24] C.C. Paige, Computational variants of the Lanczos method for the eigenproblem, J. Inst. Math. Appl. 18 (1976) 373–381.
[25] C.C. Paige, Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix, J. Inst. Math. Appl. 18 (1976) 341–349.
[26] C.C. Paige, Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem, Linear Algebra Appl. 34 (1980) 235–258.
[27] C.C. Paige, A useful form of unitary matrix obtained from any sequence of unit 2-norm n-vectors, SIAM J. Matrix Anal. Appl. 31 (2009) 565–583.
[28] C.C. Paige, An augmented stability result for the Lanczos Hermitian matrix tridiagonalization process, SIAM J. Matrix Anal. Appl. 31 (2010) 2347–2359.
[29] C.C. Paige, M. Rozložník, Z. Strakoš, Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES, SIAM J. Matrix Anal. Appl. 28 (2006) 264–284 (electronic).
[30] B.N. Parlett, The symmetric eigenvalue problem, volume 20 of *Classics in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.
[31] B.N. Parlett, D.R. Taylor, Z.A. Liu, A look-ahead Lanczos algorithm for unsymmetric matrices, Math. Comp. 44 (1985) 105–124.
[32] G.L. Sleijpen, M.B. van Gijzen, Exploiting BiCGstab(ℓ) strategies to induce dimension reduction, SIAM J. Sci. Comput. 32 (2010) 2687–2709.
[33] G.L. Sleijpen, P. Sonneveld, M.B. van Gijzen, Bi-CGSTAB as an induced dimension reduction method, Appl. Numer. Math. 60 (2010) 1100–1114.
[34] P. Sonneveld, M.B. van Gijzen, IDR(s): a family of simple and fast algorithms for solving large nonsymmetric systems of linear equations, SIAM J. Sci. Comput. 31 (2008/09) 1035–1062.
[35] C.H. Tong, Q. Ye, Analysis of the finite precision bi-conjugate gradient algorithm for nonsymmetric linear systems, Math. Comp. 69 (2000) 1559–1575.
[36] R. Underwood, An iterative block Lanczos method for the solution of large sparse symmetric eigenproblems, Ph.D. thesis, Computer Science Department, School of Humanities and Sciences, Stanford University, 1975.
[37] H.A. van der Vorst, Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 13 (1992) 631–644.
[38] H.A. van der Vorst, P. Sonneveld, CGSTAB, a more smoothly converging variant of CG-S, Report 90-50, Department of Mathematics and Informatics, Delft University of Technology, 1990.
[39] P. Wesseling, P. Sonneveld, Numerical experiments with a multiple grid and a preconditioned Lanczos type method, in: Approximation methods for Navier-Stokes problems (Proc. Sympos., Univ. Paderborn, Paderborn, 1979), volume 771 of *Lecture Notes in Math.*, Springer, Berlin, 1980, pp. 543–562.
[40] J.H. Wilkinson, Rounding errors in algebraic processes, Prentice-Hall Inc., Englewood Cliffs, N.J., 1963.
[41] J.H. Wilkinson, The algebraic eigenvalue problem, Clarendon Press, Oxford, 1965.
[42] J.P.M. Zemke, Krylov Subspace Methods in Finite Precision: A Unified Approach, Dissertation, Arbeitsbereich Technische Informatik III 4-04, Technische Universität Hamburg-Harburg, 2003.