

## IMPLICITLY RESTARTED ARNOLDI WITH PURIFICATION FOR THE SHIFT-INVERT TRANSFORMATION

KARL MEERBERGEN AND ALASTAIR SPENCE

ABSTRACT. The need to determine a few eigenvalues of a large sparse generalised eigenvalue problem  $Ax = \lambda Bx$  with positive semidefinite  $B$  arises in many physical situations, for example, in a stability analysis of the discretised Navier-Stokes equation. A common technique is to apply Arnoldi's method to the shift-invert transformation, but this can suffer from numerical instabilities as is illustrated by a numerical example. In this paper, a new method that avoids instabilities is presented which is based on applying the implicitly restarted Arnoldi method with the  $B$  semi-inner product and a purification step. The paper contains a rounding error analysis and ends with brief comments on some extensions.

### 1. INTRODUCTION

The problem of finding a few eigenvalues of large sparse  $N \times N$  generalised eigenvalue problems of the form

$$(1) \quad Ax = \lambda Bx,$$

with  $A$  nonsymmetric and  $B$  symmetric positive semidefinite, arises in many applications. For example, the block structured eigenvalue problem

$$(2) \quad \begin{bmatrix} K & C \\ C^T & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \lambda \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix},$$

with  $N = n + m$ ,  $C \in \mathbb{R}^{n \times m}$  of full rank,  $M \in \mathbb{R}^{n \times n}$  positive definite appears in the stability analysis of steady state solutions of Stokes ( $K$  symmetric) and Navier-Stokes ( $K$  nonsymmetric) equations for incompressible flow, where  $u \in \mathbb{C}^n$  denotes the velocity component and  $p \in \mathbb{C}^m$  the pressure, see for example Cliffe, Garratt, and Spence [2]. Here  $M$  is the mass matrix of the velocity elements and  $K$  is nonsymmetric because of the linearisation of the convection term in the Navier-Stokes equations. As is well known, see Malkus [9], Ericsson [5] and Cliffe, Garratt, and Spence [3], (2) can have infinite eigenvalues, corresponding to eigenvectors of the form  $(0^T, p^T)^T$ . These have no physical relevance and in applications one would only be concerned about the calculation of a small number of 'stability determining' finite eigenvalues.

A common approach for finding a few eigenvalues of (1) close to a given  $\alpha \in \mathbb{C}$  is Arnoldi's method applied to the shift-invert transformation  $S = (A - \alpha B)^{-1}B$ .

---

Received by the editor May 9, 1995 and, in revised form, November 5, 1995.

1991 *Mathematics Subject Classification*. Primary 65F15, 65F50.

*Key words and phrases*. Sparse generalised eigenvalue problems, shift-invert, semi-inner product, implicitly restarted Arnoldi.

If  $(\lambda, x)$  is an eigenpair of (1), then  $(\theta, x)$  with  $\theta = 1/(\lambda - \alpha)$  is an eigenpair of  $S$ , and this relation allows the solution of (1) by finding eigenpairs of  $S$ . Throughout the paper, we take  $\alpha = 0$  and use

$$(3) \quad S = A^{-1}B,$$

since there is no loss of generality at least in theory by making the shift  $A \leftarrow A - \alpha B$ . In fact, in many applications, a zero shift is very common, since the stability determining eigenvalues often lie close to the origin, though in other situations the accurate determination of an appropriate  $\alpha$  is a major challenge, which we do not discuss here (see for example Grimes, Lewis and Simon [8]).

To illustrate the influence of a singular  $B$  on  $S$ , consider first the following simple example. Let  $n = 2$ ,  $m = 1$  (so  $N = 3$ ), denote the columns of  $I_3$  by  $e_i, i = 1, 2, 3$ , and let

$$K = \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

with  $\beta \neq 0$ . Then

$$S = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \beta^{-1} & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

and clearly  $(\beta^{-1}, e_2)$  is an eigenpair of  $S$ , 0 is a double algebraic, simple geometric eigenvalue,  $e_3$  is the corresponding eigenvector and  $e_1$  is the generalised eigenvector or principal eigenvector of grade 2. This behaviour is generic in  $N \times N$  problems with the block structure of (2) as was shown by Malkus [9], who considered the Weierstrass-Kronecker canonical form of (2), and Ericsson [5], who considered the Jordan form for the shift-invert transformation of a variety of generalised eigenvalue problems. (Incidentally, both authors restrict attention to problems with symmetric  $A$  but several of their results, at least to do with Jordan structure, extend to the case when  $A$  is nonsymmetric). To summarise the important results on  $S$ , where  $A$  and  $B$  have the block structure of (2), we have the following theorem, which can be readily deduced from Theorem 2.7 in Ericsson [5]:

**Theorem 1.**  *$S$  defined by (2) and (3) has  $n - m$  nonzero eigenvalues, a zero eigenvalue of algebraic multiplicity  $2m$  and geometric multiplicity  $m$ . The order of the Jordan blocks corresponding to the defective eigenvalue 0 is two. Clearly,  $\mathcal{N} := \text{Null}(S) = \text{Null}(B)$  has dimension  $m$ , and  $\mathcal{G} := \text{Null}(S^2) \setminus \text{Null}(S)$  (the generalised nullspace) also has dimension  $m$ . If  $\mathcal{R} := \text{Range}(S^2)$  then  $\mathbb{C}^N$  can be decomposed as*

$$\mathbb{C}^N = \mathcal{R} + \mathcal{N} + \mathcal{G}.$$

Note finally that  $S\mathcal{G} = \mathcal{N}$  and  $S^2\mathcal{G} = S\mathcal{N} = \{0\}$ .

If  $B$  in (1) were nonsingular, then Arnoldi's method applied to  $A^{-1}B$  would be expected to find the required eigenvalues fairly easily. However, in the case we consider here with  $B$  singular, Arnoldi's method can find approximations to the zero eigenvalue of  $S$ , with consequent confusion after back transformation using  $\lambda = \theta^{-1}$ . These approximations are known as 'spurious' eigenvalues and are sometimes hard to distinguish from approximations to wanted eigenvalues. Several techniques have been proposed to reduce the risk of computing spurious eigenvalues for the symmetric nondefective problem [10, 5] and the defective problem [5]. In this paper we shall concentrate on the nonsymmetric defective case, as exemplified by  $S$  derived from (2), because of its importance in applications. In exact arithmetic, when Arnoldi's

method is applied with the initial vector  $v_1 \in \mathcal{R}$ , then only approximations to the nonzero eigenvalues of  $S$  can be computed. Such an initial vector can be chosen as  $v_1 = S^2v$  with  $v \in \mathbb{C}^N$  since  $S^2(\mathcal{N} + \mathcal{G}) = \{0\}$ . Unfortunately, in practice, round-off errors perturb the exact Arnoldi vectors producing components in  $\mathcal{N} + \mathcal{G}$  with a consequent corruption of the approximate eigenvalues and eigenvectors (we show an example of this in Section 3), and this paper is concerned with the efficient control of these unwanted directions.

For the symmetric problem, both Ericsson [5] and Nour-Omid, Parlett, Ericsson and Jensen [10] use the Lanczos method with  $B$ -orthogonalisation, i.e. use the  $B$  semi-inner product  $x^HBy$  instead of the classical  $x^Hy$ . The important point is that  $A^{-1}B$  is self adjoint with respect to the  $B$  semi-inner product and hence the Lanczos method can be used to produce the usual tridiagonal matrix. An important numerical feature is that this tridiagonal Lanczos matrix is *not* corrupted by the perturbations of the Lanczos vectors in  $\mathcal{N}$ , but *is* corrupted by perturbations of the Lanczos vectors in  $\mathcal{G}$ . Hence the Ritz values of  $S$  can be spurious. This is discussed by Ericsson [5, Theorem 3.8], and for completeness, we prove this statement later in Section 2. This cannot occur when  $S$  is nondefective, which is the case considered by Nour-Omid, Parlett, Ericsson and Jensen [10]. The Ritz vectors are more severely affected, being corrupted by round-off errors in both the  $\mathcal{N}$  and  $\mathcal{G}$  subspaces, but Ericsson shows that these errors may be eliminated by *two* applications of  $S$ , though in practice only one is in fact needed by use of a clever trick, called ‘purification’ in [10], but also discussed in [6] and [5]. In fact, we find it convenient to use the expression ‘purification’ in a more general way, so that it refers to the general operation of forming  $Sx$  from  $x \in \mathbb{C}^N$ . This arises in several places in the paper, though not always in relation to elimination of round-off errors.

Throughout this paper, we use the terminology ‘ $B$ -orthogonal Arnoldi’ to refer to Arnoldi’s method applied with the  $B$  (semi)-inner product (see Algorithm 1) in comparison with the ‘standard Arnoldi’ method, where the usual inner product is employed.

The main numerical aim of our paper is to introduce and analyse a new improved way of carrying out the two purification steps needed when  $S$  is formed from nonsymmetric problems with the block structure in (2). The new approach is based on the use of the implicitly restarted Arnoldi method (IRA) [13] combined with the  $B$  semi-inner product. This leads to an improvement over the approach in [5] on three counts. First, the Arnoldi upper Hessenberg matrix  $H_k$  is not seriously corrupted by round-off errors due to components in either  $\mathcal{N}$  or  $\mathcal{G}$ , whereas in [5] significant round-off errors could be present due to components in  $\mathcal{G}$ . Second, only *one* purification step is needed rather than two in [5], since the Arnoldi vectors could contain significant error components in  $\mathcal{N}$  but *not* in  $\mathcal{G}$ . Third, in cases when both methods fail due to severely corrupted  $H_k$  the new method has a simple test to indicate a spurious eigenvalue of  $H_k$ . One other feature of our analysis is that from a theoretical viewpoint, we are able to show an equivalence result between the  $B$ -orthogonal Arnoldi method applied to  $S$  and an  $M$ -orthogonal Arnoldi method applied to a reduced eigenvalue problem (see Section 2), which helps our understanding of the absence of corruption in  $H_k$  in the nondefective case and the presence of corruptions of  $H_k$  due to components in  $\mathcal{G}$  in the defective case.

Throughout the paper, we consider Arnoldi’s method for nonsymmetric problems, though if  $K$  in (2) were symmetric we would be able to employ the Lanczos method, with corresponding simplifications to the theory.

The plan of the paper is as follows. In Section 2, we make some theoretical observations, and give a theoretical explanation of the  $B$ -orthogonal Arnoldi method with purification. In Section 3, we explain Ericsson's approach for the defective problem and present and analyse our method based on the implicitly restarted Arnoldi method with  $B$  inner product, followed by purification. We compare Ericsson's approach and our method by a numerical example. In Section 4, we analyse the error propagation in  $\mathcal{N}$  and  $\mathcal{G}$ . Section 5 concludes with some comments and extensions.

## 2. THE $B$ -ORTHOGONAL ARNOLDI METHOD : THEORY

In this section we make three main theoretical observations. First, motivated by the approach in [5] we derive in Section 2.1 a reduced eigenvalue problem in  $\mathbb{C}^n$  rather than the full problem in  $\mathbb{C}^{n+m}$  for the case where  $B$  has the block structure in (2). Second, in Section 2.2 we prove an equivalence result about the  $B$ -orthogonal Arnoldi method applied to the full problem and the  $M$ -orthogonal Arnoldi algorithm applied to the reduced problem. Third, in Section 2.3, we discuss the reduced eigenvalue problem for the special case where  $A$  has the block structure in (2).

**2.1. The reduced eigenvalue problem.** Using (3), the generalised eigenproblem (1) is transformed to the standard eigenvalue problem

$$(4) \quad Sx = \theta x, \quad x \in \mathbb{C}^{n+m},$$

that can now be solved by Arnoldi's method. Following Ericsson [5], we first present a decomposition of  $S$ , when  $B$  has the blockstructure in (2). In this case,  $S$  has the structure

$$S = A^{-1} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} S_1 & 0 \\ S_2 & 0 \end{bmatrix}, \quad S_1 \in \mathbb{R}^{n \times n}, S_2 \in \mathbb{R}^{m \times n},$$

and this leads to a useful reduction to a problem in  $\mathbb{C}^n$ . Of course  $S_1$  and  $S_2$  depend on blocks in  $A^{-1}$  which are unlikely to be known or be helpful in a practical application, so this decomposition of  $S$  is only of theoretical interest. Let  $x = \begin{pmatrix} u \\ p \end{pmatrix}$ ,  $u \in \mathbb{C}^n$ ,  $p \in \mathbb{C}^m$ , be an eigenvector of  $S$ , and so

$$\begin{bmatrix} S_1 & 0 \\ S_2 & 0 \end{bmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \theta \begin{pmatrix} u \\ p \end{pmatrix}$$

is equivalent to

$$(5) \quad S_1 u = \theta u \quad \text{and} \quad S_2 u = \theta p.$$

If  $\theta \neq 0$  (which is the case if  $\lambda$  is finite), then

$$(6) \quad p = \theta^{-1} S_2 u$$

is a 'slave' to  $u$  and so the reduced problem

$$(7) \quad S_1 u = \theta u$$

determines precisely the nonzero eigenvalues and corresponding eigenvectors of  $S$ . Hence, in theory, it is sufficient to solve the reduced problem to solve (4). From (5), it follows that the zero eigenvalue of  $S$  has eigenvectors of the form  $(0^T, p^T)^T$ , but these obviously play no role in (7). Note that the theory in [5] which leads to Theorem 1 is derived from analysis of this decomposition of  $S$ .

**2.2. The  $B$ -orthogonal Arnoldi method and the equivalence theorem.** In most practical cases we use the  $B$ -orthogonal Arnoldi method in a direct attempt to solve (4), rather than try to solve the reduced problem, so we first write down the method in detail.

**Algorithm 1.** *The  $B$ -orthogonal Arnoldi method*

Given  $v_1$  with  $\|v_1\|_B^2 = v_1^H B v_1 = 1$ .

**for**  $i = 1$  **to**  $k$  **do**

    Compute  $w_{i+1} = A^{-1} B v_i$ .

    Form  $h_{ji} = v_j^H B w_{i+1}$ ,  $j = 1, \dots, i$ .

    Form  $w_{i+1} = w_{i+1} - \sum_{j=1}^i v_j h_{ji}$ .

    Compute  $h_{i+1,i} = \|w_{i+1}\|_B$ .

    Let  $v_{i+1} = w_{i+1}/h_{i+1,i}$ .

**end for**

Let  $H_k = [h_{ji}]$ .

Compute the eigenpairs  $(\theta_j, z_j)$   $j = 1, \dots, k$  of  $H_k$ .

Compute the approximate eigenpairs  $(\lambda_j, x_j)$  of  $Ax = \lambda Bx$  with  $\lambda_j = 1/\theta_j$  and  $x_j = V_k z_j$ .

The  $B$ -orthogonal Arnoldi algorithm is merely the standard Arnoldi algorithm with the usual inner product  $x^H y$  replaced by the semi-inner product  $x^H B y$ . To be precise, the method computes a  $B$ -orthonormal basis  $v_1, \dots, v_{k+1}$  of the Krylov space

$$\mathcal{K}_{k+1}(v_1, S) = \text{span}\{v_1, S v_1, S^2 v_1, \dots, S^k v_1\}, \quad S = A^{-1} B,$$

with the matrix  $H_k = V_k^H B S V_k$  being upper Hessenberg, and  $V_k^H B V_k = I$  with  $V_k = [v_1, \dots, v_k]$ . It looks strange that a singular matrix is used to compute the inner product, but as we shall see later in this section, it is a natural thing to do for generalised problems of the form (2). The approximate eigenpairs of  $(\theta, x)$  are computed from the eigenpairs of  $H_k$  in the usual way, see Algorithm 1. Similarly, one can use the  $M$ -orthogonal Arnoldi method to solve the reduced problem. As we now prove, there is a precise connection between the  $B$ -orthogonal Arnoldi method applied to  $S$  and the  $M$ -orthogonal Arnoldi method applied to  $S_1$ , which we state in Theorem 2. We split the proof into two parts presented as lemmas.

**Lemma 1.** *Consider the Krylov space  $\mathcal{K}_k(v_1, S)$  with  $v_1^T = (y_1^T, p_1^T)$ ,  $y_1 \in \mathbb{C}^n$  and  $p_1 \in \mathbb{C}^m$ . Let  $v^T = (y^T, p^T)$ ,  $v \in \mathcal{K}_k(v_1, S)$ , then  $y \in \mathcal{K}_k(y_1, S_1)$ .*

*Proof.* Following the block structure of  $S$ ,

$$S \begin{pmatrix} y_1 \\ p_1 \end{pmatrix} = \begin{pmatrix} S_1 y_1 \\ S_2 y_1 \end{pmatrix}$$

and by induction on  $j = 2, \dots, k$ , it follows that

$$S^j \begin{pmatrix} y_1 \\ p_1 \end{pmatrix} = S \left[ S^{j-1} \begin{pmatrix} y_1 \\ p_1 \end{pmatrix} \right] = \begin{pmatrix} S_1^j y_1 \\ S_2 S_1^{j-1} y_1 \end{pmatrix},$$

which shows the lemma. □

**Lemma 2.** *The  $B$ -orthogonal projection of  $Sx = \theta x$  on the space  $\text{Range}(V_k)$  with  $V_k^T = [Y_k^T \ P_k^T]$ ,  $Y_k \in \mathbb{C}^{n \times k}$ ,  $P_k \in \mathbb{C}^{m \times k}$  produces the same approximate eigenvalues as the  $M$ -orthogonal projection of  $S_1 u = \theta u$  on  $\text{Range}(Y_k)$ . Moreover, the first  $n$  components of the approximate eigenvectors of  $S$  computed by the  $B$ -orthogonal*

projection correspond to the approximate eigenvectors of  $S_1$  under  $M$ -orthogonal projection.

*Proof.* The  $B$ -orthogonal projection of  $Sx = \theta x$  on  $\text{Range}(V_k)$  produces  $(\theta, x)$  such that

$$(8) \quad Sx - \theta x \perp_B \text{Range}(V_k), \quad x \in \text{Range}(V_k),$$

where  $x \perp_B y$  means  $x^H B y = 0$ . Recall that

$$S = \begin{bmatrix} S_1 & 0 \\ S_2 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}.$$

Decompose  $x^T = (u^T, p^T)$  with  $u \in \mathbb{C}^n$  and  $p \in \mathbb{C}^m$ , then (8) is equivalent to

$$\begin{pmatrix} S_1 u - \theta u \\ S_2 u - \theta p \end{pmatrix} \perp_B \text{Range}(V_k), \quad x \in \text{Range}(V_k),$$

and because  $x^H B V_k = u^H M Y_k$ , (8) can further be reduced to

$$S_1 u - \theta u \perp_M \text{Range}(Y_k), \quad u \in \text{Range}(Y_k),$$

which is precisely the  $M$ -orthogonal projection of the reduced problem on  $\text{Range}(Y_k)$ .  $\square$

The following theorem gives the connection between  $M$ -orthogonal Arnoldi and  $B$ -orthogonal Arnoldi and is an immediate consequence of Lemmas 1 and 2.

**Theorem 2** (Equivalence Theorem). *If  $k$  steps of  $M$ -orthogonal Arnoldi applied to  $S_1$  starting with  $y_1 \in \mathbb{C}^n$  produces an eigenpair  $(\theta, u)$ , then  $k$  steps of  $B$ -orthogonal Arnoldi applied to  $S$  starting with  $v_1$ , given by*

$$(9) \quad v_1 = \begin{pmatrix} y_1 \\ q_1 \end{pmatrix}$$

with  $q_1 \in \mathbb{C}^m$  arbitrary, produces an eigenpair  $(\theta, x)$  with  $x^T = (u^T, p^T)$ .

The equivalence theorem shows that solving (4) with  $B$ -orthogonal Arnoldi and (7) with  $M$ -orthogonal Arnoldi for the initial vector chosen as in (9) compute the same eigenvalues. The  $u$ -component of  $x$  corresponds to the computed eigenvector of  $S_1$ . However, the  $p$ -component does not play a role in the  $B$ -orthogonal Arnoldi method and cannot be guaranteed to be correct. The correct  $p$  is obtained by the slave relation (6), so that if  $(\theta, u)$  is an exact eigenpair of  $S_1$ , then, following (6),  $(\theta, \begin{pmatrix} u \\ \theta^{-1} S_2 u \end{pmatrix})$  is an exact eigenpair of  $S$ . If  $x = \begin{pmatrix} u \\ p \end{pmatrix}$  is computed by  $B$ -orthogonal Arnoldi, (6) should be computed explicitly. The practical way to do this is to apply  $S$  to  $x$ , since

$$Sx = \begin{bmatrix} S_1 & 0 \\ S_2 & 0 \end{bmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} S_1 u \\ S_2 u \end{pmatrix} = \theta \begin{pmatrix} u \\ \theta^{-1} S_2 u \end{pmatrix}$$

is an eigenvector belonging to  $\theta$ . In the  $B$ -orthogonal Arnoldi method,  $Sx$  can be computed without extra vector operations with  $S$ . The Arnoldi vectors namely satisfy the relation

$$(10) \quad S V_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^H,$$

with  $e_k$  the  $k$ -th unit vector of length  $k$ , which is also written as

$$(11) \quad S V_k = V_{k+1} \bar{H}_k \quad \text{with} \quad \bar{H}_k = \begin{bmatrix} H_k \\ h_{k+1,k} e_k^T \end{bmatrix} \in \mathbb{C}^{k+1 \times k}.$$

Hence

$$(12) \quad Sx = SV_k z = V_{k+1} \bar{H}_k z$$

and since  $z$  is an eigenvector of  $H_k$ ,

$$(13) \quad SV_k z = \theta_j V_k z + h_{k+1,k} v_{k+1} e_k^T z .$$

This expression was used by Ericsson and Ruhe [6] as a means of improving accuracy in eigenvectors, but Nour-Omid, Parlett, Ericsson and Jensen [10] considered especially the case  $B$  singular and called the correction of  $V_k z$  ‘purification’, because it has the potential for removing significant round-off error components in the  $\mathcal{N}$  space. Since this is a rather important result for our method, we describe why this is so in Section 4.

Note that the  $B$  inner product is not affected by the portions of  $v_j$  in the nullspace of  $B$ , since for  $x = (u^T, p^T)^T$  and  $y = (v^T, q^T)^T$ ,  $x^H B y = u^H M v$ . Since  $B$  and  $S$  have the same nullspace,  $H_k$  is independent of components of  $V_k$  in the  $\mathcal{N}$  space.

There is another way to explain the  $B$ -orthogonal Arnoldi method. Since  $B$  is positive semidefinite there exists a Cholesky like factorisation  $B = P^T P$  with  $P \in \mathbb{R}^{N \times N}$  singular. Nour-Omid, Parlett, Ericsson and Jensen [10] point out to an equivalence result between  $B$ -orthogonal Arnoldi applied to  $A^{-1}B$  and standard Arnoldi applied to  $PA^{-1}P^T$ . It appears that with appropriate initial vectors, the eigenvalues computed by both methods are equal and that one needs a purification step to find the eigenvectors of  $A^{-1}B$  in the  $B$ -orthogonal Arnoldi method.

Finally, note that if  $S_1$  is nonsingular, the  $B$ -orthogonal Arnoldi method applied to  $S$  would not be expected to compute approximations to the zero eigenvalue of  $S$  since the method is equivalent to  $M$ -orthogonal Arnoldi applied to  $S_1$ . This is essentially the case described in [10].

**2.3. Application of  $B$ -orthogonal Arnoldi to the Navier-Stokes problem.**

Now consider the problem with block structure (2) which provides the main motivation for this paper. From  $S = A^{-1}B$  we have that

$$\begin{bmatrix} K & C \\ C^T & 0 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ S_2 & 0 \end{bmatrix} = \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} ,$$

which yields  $C^T S_1 = 0$ . Also, since  $C \in \mathbb{R}^{n \times m}$  is of full rank,  $S_1 \in \mathbb{R}^{n \times n}$  has at least  $m$  independent left null vectors and so,  $\text{rank}(S_1) \leq n - m$ . Since  $A$  is nonsingular,  $\text{rank}(S) = \text{rank}(B) = n$ . Since  $S_2 \in \mathbb{R}^{m \times n}$ ,  $\text{rank}(S_2) \leq m$ , and, since  $\text{rank}(S) \leq \text{rank}(S_1) + \text{rank}(S_2)$ ,

$$\text{rank}(S_1) \geq \text{rank}(S) - \text{rank}(S_2) \geq n - m$$

so,  $\text{rank}(S_1) = n - m$ . Clearly, the nullspace of  $S_1$ ,  $\mathcal{N}_1$  say, has dimension  $m$ , and the range,  $\mathcal{R}_1$  say, has dimension  $n - m$ . Note that for  $u \in \mathcal{R}_1$ , it can easily be shown that  $C^T u = 0$  and this agrees with the second equation in (2).

We can now be more precise about the link between  $S_1$  and  $S$  as follows. For  $u$  and  $x$  defined as  $S_1 u = \theta u$  and  $Sx = \theta x$ ,

$$(14) \quad u \in \mathcal{R}_1 \iff x = \begin{pmatrix} u \\ \theta^{-1} S_2 u \end{pmatrix} \in \mathcal{R}$$

and

$$(15) \quad u \in \mathcal{N}_1 \iff \begin{pmatrix} u \\ 0 \end{pmatrix} \in \mathcal{G} .$$

Equivalence (14) follows from (5) and (6), and (15) follows from the fact that

$$S \begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} S_1 u \\ S_2 u \end{pmatrix} = \begin{pmatrix} 0 \\ S_2 u \end{pmatrix} \in \mathcal{N}$$

and thus

$$S^2 \begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} S_1^2 u \\ S_2 S_1 u \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

Since  $S_1$  has an eigenvalue zero,  $M$ -orthogonal Arnoldi (and  $B$ -orthogonal Arnoldi) may compute the zero eigenvalue in practice. Nevertheless, if  $y_1 \in \mathcal{R}_1$ ,

$$\mathcal{K}_k(S_1, y_1) \subset \mathcal{R}_1$$

and so in exact arithmetic, the zero eigenvalue cannot be computed.  $y_1 \in \mathcal{R}_1$  can be achieved by  $y_1 \leftarrow S_1 y_1$ . To apply  $B$ -orthogonal Arnoldi to  $S$ , one should start with  $v_1$ , given by (9). This is achieved by  $v_1 \leftarrow S v_1$ , since

$$S v_1 = S \begin{pmatrix} y_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} S_1 y_1 \\ S_2 y_1 \end{pmatrix} .$$

Note that then in exact arithmetic  $\mathcal{K}_k(S, S v_1) \subset \mathcal{R} + \mathcal{N}$ . Of course, in practice, round-off errors cannot be avoided and in the following two sections we derive and analyse a new approach to mitigate the effect of round-off for this problem. Ericsson [5] even suggests to start with  $v_1 \leftarrow S^2 v_1 \in \mathcal{R}$ .

### 3. THE $B$ -ORTHOGONAL ARNOLDI METHOD : PRACTICE

In this section, we describe the Ericsson approach [5] and present a new approach which under certain conditions significantly reduces the effects of errors in  $\mathcal{G}$  and  $\mathcal{N}$ . The approach is based on a combination of the implicitly restarted Arnoldi method of Sorensen [13] with  $B$ -inner product and purification.

**3.1. The Ericsson approach.** Ericsson showed that in finite precision arithmetic  $H_k$  can be corrupted when  $S$  has a generalised nullspace  $\mathcal{G}$ , even when  $v_1 \in \mathcal{R}$ . Although spurious eigenvalues might be computed, the computed eigenvectors can easily be mapped in  $\mathcal{R}$  by applying  $S$  twice instead of once. This procedure leads to uncorrupted eigenvectors. But of course, the spurious eigenvalues remain unaltered.

The application of  $S^2$  to  $x$  can be carried out implicitly by two ‘purification’ steps. From (12)

$$S^2 x = S(V_{k+1} \bar{H}_k z) ,$$

and one further step of the Arnoldi process is needed to calculate  $v_{k+2}$  and the last column of  $\bar{H}_{k+1}$ , in order to apply  $S^2$  efficiently to all eigenvectors. Hence, following the recurrence relation (11),

$$(16) \quad S^2 x = S V_{k+1} \bar{H}_k z = V_{k+2} \bar{H}_{k+1} \bar{H}_k z .$$

The Ericsson algorithm is thus as follows.

**Algorithm 2.** *The Ericsson purification for the defective case*

Given the initial vector  $v_1 \leftarrow S^2 v_1$ .

Compute in  $k$  steps of  $B$ -orthogonal Arnoldi the matrices  $V_{k+1}$  and  $\bar{H}_k$ .

Compute the eigenpairs  $(\theta_j, z_j)$  of  $H_k$ .

Perform one more step of Arnoldi : compute  $v_{k+2}$  and the last column of  $\bar{H}_{k+1}$ .

Compute the eigenpairs  $(\lambda_j, x_j)$  of  $Ax = \lambda Bx$  as  $\lambda_j = \theta_j^{-1}$  and  $x_j = V_{k+2} \bar{H}_{k+1} \bar{H}_k z_j$ .

**3.2. A new approach.** In order to present the details of our method, we need first to understand the steps of the implicitly restarted Arnoldi method by Sorensen [13], and to do this we present the following theorem, which is a compilation of results from [13], but is stated and proved here for completeness.

**Theorem 3.** Let  $v_1, \dots, v_{k+2}$ ,  $\bar{H}_{k+1} = \begin{bmatrix} H_{k+1} & \\ 0 \cdots 0 & h_{k+2,k+1} \end{bmatrix}$  be computed by  $B$ -orthogonal Arnoldi applied to  $S$ . Consider the QR factorisation  $\bar{H}_{k+1} = \bar{Q}_{k+1}\bar{R}_{k+1}$  with  $\bar{Q}_{k+1} \in \mathbb{C}^{(k+2) \times (k+1)}$  unitary and  $\bar{R}_{k+1} \in \mathbb{C}^{(k+1) \times (k+1)}$  upper triangular. Let  $\bar{Q}_k$  be the  $k+1 \times k$  upper left part of  $\bar{Q}_{k+1}$  and define

$$W_{k+1} = V_{k+2}\bar{Q}_{k+1} \quad \text{and} \quad \bar{G}_k = R_{k+1}\bar{Q}_k \in \mathbb{C}^{(k+1) \times k} .$$

If  $H_{k+1}$  has nonzero eigenvalues, then  $W_{k+1}$  and  $\bar{G}_k$  have the following properties :

1.  $W_{k+1}^H B W_{k+1} = I$ .
2.  $\text{span}(w_1, \dots, w_{k+1}) = \text{span}(Sv_1, \dots, Sv_{k+1})$ .
3.  $\bar{G}_k$  is upper Hessenberg and for  $W_{k+1} = [W_k \ w_{k+1}]$ ,

$$S W_k = W_{k+1} \bar{G}_k$$

with  $w_1 = Sv_1 / \|Sv_1\|_B$ .

*Proof.* (See also Sorensen [13].) The proof of statement 1 is easy and is omitted.  $\bar{H}_{k+1}$  and  $V_{k+2}$  are computed from

$$\begin{aligned} SV_{k+1} &= V_{k+1}H_{k+1} + v_{k+2}h_{k+2,k+1}e_{k+1}^T \\ (17) \quad &= V_{k+2}\bar{H}_{k+1} , \end{aligned}$$

and replacing  $\bar{H}_{k+1}$  by its QR factorisation, we obtain

$$(18) \quad SV_{k+1} = W_{k+1}R_{k+1} .$$

If  $H_{k+1}$  has nonzero eigenvalues, then  $\text{rank}(R_{k+1}) = \text{rank}(\bar{H}_{k+1}) \geq \text{rank}(H_{k+1}) = k+1$ , and thus  $\text{rank}(R_{k+1}) = k+1$ . Hence,

$$(19) \quad W_{k+1} = SV_{k+1}R_{k+1}^{-1}$$

which proves statement 2.

Let  $W_k$  be the first  $k$  columns of  $W_{k+1}$ , then  $W_k = V_{k+1}\bar{Q}_k$ . Multiply (18) on the right by  $\bar{Q}_k$ , then

$$S W_k = W_{k+1} \bar{G}_k$$

with  $\bar{G}_k = R_{k+1}\bar{Q}_k$ . It is well known (see Sorensen [13]) that if  $\bar{H}_{k+1}$  is upper Hessenberg,  $\bar{Q}_k$  and  $\bar{Q}_{k+1}$  are upper Hessenberg too, and therefore  $\bar{G}_k$  is upper Hessenberg. Since  $w_1$  is the first column of (19),

$$w_1 = Sv_1 r_{11}^{-1}$$

and since  $\|w_1\|_B = 1$ ,  $r_{11} = \|Sv_1\|_B$ . □

In other words, the theorem says that  $W_{k+1}$  and  $\bar{G}_k$  can be viewed as having been computed by the  $B$ -orthogonal Arnoldi method with starting vector  $Sv_1 / \|Sv_1\|_B$ , which is exactly what we wish to do, but is accomplished implicitly by the QR factorisation of  $\bar{H}_{k+1}$ .

In Algorithm 3 we write down the different steps that are needed to perform one implicit restart step. Our approach consists of three stages :

- (1) the computation of  $V_{k+2}$ ,

TABLE 1. Schematic representation of the flow of the error components in  $\mathcal{N}$  and  $\mathcal{G}$  by implicit application of  $S$  using first implicitly restarted  $B$ -orthogonal Arnoldi and then purification (Here the notation  $V_{k+2}(\mathcal{N}, \mathcal{G})$  means that  $V_{k+2}$  contains components in  $\mathcal{N}$  and  $\mathcal{G}$ )

stage 1 ( $B$ -orthog. Arn.)	$V_{k+2}(\mathcal{N}, \mathcal{G})$	$\bar{H}_{k+1}(\mathcal{G})$
implicit $S$ :	$\downarrow \quad \downarrow \quad \downarrow$	$\downarrow$
stage 2 (implicit restart)	$W_{k+1}(\{0\}, \mathcal{N})$	$\bar{G}_k(\quad)$
implicit $S$ :	$\downarrow \quad \downarrow$	
stage 3 (purification)	$x_j(\quad \{0\})$	

- (2) an implicit application of  $S$  to  $V_{k+1}$  (achieved by the QR step for  $\bar{H}_{k+1}$ ), which has the effect of removing the  $\mathcal{N}$  portion of  $V_{k+1}$  and mapping the  $\mathcal{G}$  portion into  $\mathcal{N}$ ,
- (3) the purification of  $W_k z$  that achieves the removal of the remaining  $\mathcal{N}$  component (which is a second implicit application of  $S$ ).

Table 1 gives a schematic representation of the impact of these implicit applications of  $S$  on the  $\mathcal{N}$  and  $\mathcal{G}$  components of  $V_{k+2}$ ,  $W_{k+1}$  and  $W_k z$ , and on the Arnoldi upper Hessenberg matrices  $\bar{H}_{k+1}$  and  $\bar{G}_k$ . We assume that  $V_{k+2}$  has components in  $\mathcal{N}$  and  $\mathcal{G}$  which we represent by  $V_{k+2}(\mathcal{N}, \mathcal{G})$  and we trace the propagation of these components through the 3 stages of Algorithm 3. We see that  $\bar{G}_k$  and the purified  $x_j$  do not depend on  $\mathcal{N}$  and  $\mathcal{G}$  components.

**Algorithm 3.** *Implicitly restarted  $B$ -orthogonal Arnoldi with purification*

Compute in  $k + 1$  steps of  $B$ -orthogonal Arnoldi the matrices  $V_{k+2}$  and  $\bar{H}_{k+1}$ .  
 Compute the QR factorisation  $\bar{H}_{k+1} = \bar{Q}_{k+1} R_{k+1}$ .  
 Perform one restart :  $W_{k+1} = V_{k+2} \bar{Q}_{k+1}$ ,  $\bar{G}_k = R_{k+1} \bar{Q}_k$ .  
 Let  $G_k$  be the  $k \times k$  upper part of  $\bar{G}_k$ .  
 Compute the eigenpairs  $(\theta_j, z_j)$ ,  $j = 1, \dots, k$  of  $G_k$ .  
 Compute the eigenpairs  $(\lambda_j, x_j)$  of  $Ax = \lambda Bx$  as  $\lambda_j = \theta_j^{-1}$  and  $x_j = W_{k+1} \bar{G}_k z_j$ .

**3.3. Numerical example.** In this section, we compare the Ericsson approach for the defective case and the implicitly restarted Arnoldi method with purification for a small example. Consider the matrices  $A$  and  $B$  generated using the following Matlab code.

```

rand('seed', 1); n=200; m=100;
C = [rand(m,m) ; zeros(n-m,m)];
K = rand(n);
A = [K C ; C' zeros(m,m)];
B = [eye(n,n) zeros(n,m) ; zeros(m,n+m)];

```

The rightmost eigenvalues of  $Ax = \lambda Bx$  are 49.9129, 2.7017,  $2.2521 \pm 1.4533i$  and 2.0517, and all eigenvalues satisfy

$$(20) \quad -3 \leq \operatorname{Re}(\lambda) \leq 50 .$$

TABLE 2. The growth of  $\mathcal{N}$  and  $\mathcal{G}$  components of the vectors  $v_j$ , computed by  $B$ -orthogonal Arnoldi, for the example in Section 3.3

$j$	$\ P_{\mathcal{N}}v_j\ _2$	$\ P_{\mathcal{G}}v_j\ _2$
1	$3.4 \cdot 10^{-11}$	$2.57 \cdot 10^{-12}$
2	$5.70 \cdot 10^{-10}$	$1.52 \cdot 10^{-11}$
3	$1.57 \cdot 10^{-8}$	$3.10 \cdot 10^{-10}$
4	$3.88 \cdot 10^{-7}$	$6.24 \cdot 10^{-9}$
5	$1.06 \cdot 10^{-5}$	$1.43 \cdot 10^{-7}$
6	$2.78 \cdot 10^{-4}$	$3.26 \cdot 10^{-6}$
7	$6.73 \cdot 10^{-3}$	$6.94 \cdot 10^{-5}$
8	$1.80 \cdot 10^{-1}$	$1.66 \cdot 10^{-3}$
9	$4.34 \cdot 10^0$	$3.61 \cdot 10^{-2}$
10	$7.70 \cdot 10^1$	$5.88 \cdot 10^{-1}$
11	$1.20 \cdot 10^2$	$8.05 \cdot 10^{-1}$

To compute a few eigenvalues, we used  $S = (A - 60B)^{-1}B$ . It is easy to check that for this example,  $S$  has the block form

$$S = \begin{bmatrix} 0 & 0 & 0 \\ 0 & S_{22} & 0 \\ S_{31} & S_{32} & 0 \end{bmatrix}$$

where  $S_{22} \in \mathbb{R}^{n-m \times n-m}$  and  $S_{31} \in \mathbb{R}^{m \times m}$  are nonsingular. Then

$$\mathcal{G} := \left\{ \begin{pmatrix} u \\ 0 \\ 0 \end{pmatrix} : u \in \mathbb{C}^m \right\}, \quad \mathcal{N} := \left\{ \begin{pmatrix} 0 \\ 0 \\ p \end{pmatrix} : p \in \mathbb{C}^m \right\} \text{ and}$$

$$\mathcal{R} := \left\{ \begin{pmatrix} 0 \\ u \\ S_{32}S_{22}^{-1}u \end{pmatrix} : u \in \mathbb{C}^{n-m} \right\}.$$

Hence, for any vector  $y$ , the size of  $\mathcal{N}$  and  $\mathcal{G}$  components can be measured by calculating  $\|P_{\mathcal{N}}y\|_2$  and  $\|P_{\mathcal{G}}y\|_2$  where

$$P_{\mathcal{N}} = [0 \quad -S_{32}S_{22}^{-1} \quad I_m] \in \mathbb{R}^{m \times n+m},$$

$$P_{\mathcal{G}} = [I \quad 0 \quad 0] \in \mathbb{R}^{m \times n+m}.$$

In practice we first normalise so that  $\|P_{\mathcal{N}}\|_2 = \|P_{\mathcal{G}}\|_2 = 1$ . For all runs we selected the initial vector as

$$v_1 = S^2[1, \dots, 1]^T \in \mathcal{R}.$$

One would expect that the eigenvalue 49.9 would be found in a few Arnoldi steps, since it is clearly much closer to the shift,  $\alpha$ , than any other eigenvalues. Besides the rightmost eigenvalue, we find that a number of spurious eigenvalues are also computed that do not satisfy (20).

We first illustrate the growth of the  $\mathcal{N}$  and  $\mathcal{G}$  components of the Arnoldi vectors in finite precision arithmetic by the measures  $\|P_{\mathcal{N}}v_j\|_2$  and  $\|P_{\mathcal{G}}v_j\|_2$  in Table 2. As recommended in [5] we choose a starting vector in  $\mathcal{R}$ , but even then the corrupting effect of round-off is clearly seen. Recall that in exact arithmetic, if  $v_1 \in \mathcal{R}$ , then  $v_1, \dots, v_{k+1} \in \mathcal{R}$ .

TABLE 3. Numerical results for the Example in Section 3.3 for  $B$ -orthogonal Arnoldi with double purification step (Algorithm 2)

$k = 10$					
$j$	$\lambda_j$	$\ P_{\mathcal{N}}(V_k z_j)\ $	$\ P_{\mathcal{G}}(V_k z_j)\ $	$\ P_{\mathcal{N}}x_j\ $	$\ P_{\mathcal{G}}x_j\ $
1	49.9	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$
2	$2.2 + 1.2i$	$3 \cdot 10^{-3}$	$2 \cdot 10^{-5}$	$2 \cdot 10^{-12}$	$4 \cdot 10^{-14}$
$\vdots$					
9	$-1.9 - 1.0i$	$4 \cdot 10^{-3}$	$3 \cdot 10^{-5}$	$7 \cdot 10^{-12}$	$2 \cdot 10^{-13}$
10	-32.2	$4 \cdot 10^{-3}$	$3 \cdot 10^{-5}$	$9 \cdot 10^{-11}$	$2 \cdot 10^{-13}$
$k = 15$					
$j$	$\lambda_j$	$\ P_{\mathcal{N}}(V_k z_j)\ $	$\ P_{\mathcal{G}}(V_k z_j)\ $	$\ P_{\mathcal{N}}x_j\ $	$\ P_{\mathcal{G}}x_j\ $
1	49.9	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$
2	$3.0 + 1.1i$	$3 \cdot 10^{-3}$	$5 \cdot 10^{-14}$	$2 \cdot 10^{-12}$	$5 \cdot 10^{-14}$
$\vdots$					
14	$-2.3 - 1.0i$	$4 \cdot 10^{-3}$	$5 \cdot 10^{-14}$	$5 \cdot 10^{-12}$	$9 \cdot 10^{-14}$
15	$-1 \cdot 10^{12}$	$3 \cdot 10^{-3}$	$5 \cdot 10^{-14}$	$5 \cdot 10^{-4}$	$3 \cdot 10^{-14}$
$k = 20$					
$j$	$\lambda_j$	$\ P_{\mathcal{N}}(V_k z_j)\ $	$\ P_{\mathcal{G}}(V_k z_j)\ $	$\ P_{\mathcal{N}}x_j\ $	$\ P_{\mathcal{G}}x_j\ $
1	49.9	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$
2	$3.0 + 1.2i$	$4 \cdot 10^{-3}$	$5 \cdot 10^{-14}$	$1 \cdot 10^{-8}$	$5 \cdot 10^{-14}$
$\vdots$					
19	$-1 \cdot 10^6$	$4 \cdot 10^{-3}$	$5 \cdot 10^{-14}$	$5 \cdot 10^{-7}$	$4 \cdot 10^{-14}$
20	$-1 \cdot 10^{12}$	$4 \cdot 10^{-3}$	$5 \cdot 10^{-9}$	$5 \cdot 10^{-3}$	$6 \cdot 10^{-14}$

Let us now discuss the results for Algorithm 2, which is the  $B$ -orthogonal Arnoldi method with two purification steps using (16). Consider first the results in Table 3 for  $k = 10$ . We see that  $\lambda_{1,2,9}$  satisfy (20) and in fact correspond to finite eigenvalues of  $Ax = \lambda Bx$ . As we mentioned before, the eigenvalues obtained from Algorithm 2 can be spurious since they are precisely those of the upper Hessenberg matrix  $H_k$  computed by  $B$ -orthogonal Arnoldi without purification. However, the purified eigenvectors lie in the correct space. We see that  $\lambda_{10} = -32.2$  is spurious since  $\lambda_{10} \ll -3$ , but  $\|P_{\mathcal{N}}x_{10}\| \simeq 10^{-10}$  and  $\|P_{\mathcal{G}}x_{10}\| \simeq 10^{-13}$ . Note that the unpurified eigenvectors  $V_k z_j$  have significant components in  $\mathcal{N}$  and  $\mathcal{G}$ . The practical effect of the implicit application of  $S^2$  in the purification step given by (16) in the presence of large error components in the Arnoldi vectors is striking (cf. Table 1 in [10]). However, this purification can fail. For example, with  $k = 20$ , the purified  $x_{20}$  still contains a significant component in  $\mathcal{N}$ , which is represented by  $\|P_{\mathcal{N}}x_{20}\| \simeq 10^{-3}$ . We shall see in Section 4 that this is due to the fact that  $|\lambda_{20}| (\simeq 10^{12})$  is large. Similar conclusions hold for  $k = 15$ .

Now consider the results for IRA plus purification (Algorithm 3) in Table 4. First consider  $k = 10$ . After  $k + 1 = 11$  steps of the  $B$ -orthogonal Arnoldi method, we have  $\|P_{\mathcal{N}}V_{12}\|_2 \simeq 2 \cdot 10^3$  and  $\|P_{\mathcal{G}}V_{12}\|_2 \simeq 1 \cdot 10^0$ . After one implicit restart,  $\|P_{\mathcal{N}}W_{11}\|_2 \simeq 2 \cdot 10^3$  and is essentially unchanged, but  $\|P_{\mathcal{G}}W_{11}\|_2 \simeq 3 \cdot 10^{-9}$ . By the implicit operation with  $S$ , the  $\mathcal{N}$  component of  $V_{k+2}$  is wiped out and the  $\mathcal{G}$

TABLE 4. Numerical results for the example in Section 3.3 for implicitly restarted Arnoldi with  $B$ -orthogonalisation and purification

$k = 10$			
$\ P_{\mathcal{N}}V_{k+2}\ _2 \simeq 2 \cdot 10^3$		$\ P_{\mathcal{G}}V_{k+2}\ _2 \simeq 1 \cdot 10^0$	
$\ P_{\mathcal{N}}W_{k+1}\ _2 \simeq 2 \cdot 10^3$		$\ P_{\mathcal{G}}W_{k+1}\ _2 \simeq 3 \cdot 10^{-9}$	
$\ R_{k+1}^{-1}\ _2 \simeq 9 \cdot 10^2$			
$j$	$\lambda_j$	$\ P_{\mathcal{N}}x_j\ $	$\ P_{\mathcal{G}}x_j\ $
1	49.9	$7 \cdot 10^{-14}$	$5 \cdot 10^{-14}$
2	$2.2 + 1.3i$	$2 \cdot 10^{-11}$	$6 \cdot 10^{-14}$
$\vdots$			
10	$-2.3 - 1.1i$	$5 \cdot 10^{-11}$	$1 \cdot 10^{-13}$
$k = 15$			
$\ P_{\mathcal{N}}V_{k+2}\ _2 \simeq 1 \cdot 10^9$		$\ P_{\mathcal{G}}V_{k+2}\ _2 \simeq 1 \cdot 10^0$	
$\ P_{\mathcal{N}}W_{k+1}\ _2 \simeq 1 \cdot 10^9$		$\ P_{\mathcal{G}}W_{k+1}\ _2 \simeq 2 \cdot 10^{-2}$	
$\ R_{k+1}^{-1}\ _2 \simeq 6 \cdot 10^9$			
$j$	$\lambda_j$	$\ P_{\mathcal{N}}x_j\ $	$\ P_{\mathcal{G}}x_j\ $
1	49.9	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$
2	$2.9 + 1.2i$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-14}$
$\vdots$			
15	$-2.3 - 1.1i$	$1 \cdot 10^{-4}$	$8 \cdot 10^{-14}$
$k = 20$			
$\ P_{\mathcal{N}}V_{k+2}\ _2 \simeq 7 \cdot 10^{10}$		$\ P_{\mathcal{G}}V_{k+2}\ _2 \simeq 1 \cdot 10^1$	
$\ P_{\mathcal{N}}W_{k+1}\ _2 \simeq 7 \cdot 10^{10}$		$\ P_{\mathcal{G}}W_{k+1}\ _2 \simeq 1 \cdot 10^1$	
$\ R_{k+1}^{-1}\ _2 \simeq 4 \cdot 10^{11}$			
$j$	$\lambda_j$	$\ P_{\mathcal{N}}x_j\ $	$\ P_{\mathcal{G}}x_j\ $
1	49.9	$7 \cdot 10^{-13}$	$5 \cdot 10^{-14}$
2	$2.9 + 1.2i$	$1 \cdot 10^{-7}$	$5 \cdot 10^{-14}$
$\vdots$			
19	$-2.3 - 1.1i$	$8 \cdot 10^{-7}$	$9 \cdot 10^{-14}$
20	$-2.2 \cdot 10^9$	$4 \cdot 10^{-3}$	$6 \cdot 10^{-14}$

component is mapped to the  $\mathcal{N}$  component of  $W_{k+1}$ . All the computed eigenvalues satisfy (20). The purified eigenvectors have small  $\mathcal{N}$  and  $\mathcal{G}$  components. Even though  $W_{k+1}$  has small  $\mathcal{G}$  components  $G_k$  is not affected by components in  $\mathcal{N}$  and  $\mathcal{G}$ , and the purified  $x_j$  has small  $\mathcal{N}$  and  $\mathcal{G}$  portions in spite of large  $\mathcal{N}$  and  $\mathcal{G}$  components in  $V_{k+1}$ .

For  $k = 15$ , the eigenvalues are not spurious since they satisfy (20), but the eigenvectors have components in the  $\mathcal{N}$  space of the order of  $10^{-4}$ . This follows from the fact that the implicit restart did not remove the  $\mathcal{G}$  components in  $W_{k+1}$  well. Spurious eigenvalues did not appear, which shows, at least for this example, the superiority of Algorithm 3 over Algorithm 2 for the calculation of eigenvalues.

Now consider the performance of IRA plus purification for  $k = 20$ . For  $k = 20$ , the IRA-step was not able to dramatically reduce the  $\mathcal{G}$ -component of  $W_{k+1}$ . As a consequence,  $G_k$  produces a spurious eigenvalue  $\lambda_{20}$ . Also the purified eigenvector  $x_{20}$  has a ‘large’ component in  $\mathcal{N}$ . We shall see in Section 4 that the failure of IRA is due to the fact that  $\|R_{k+1}^{-1}\|$  is large. On the other hand, remark that the purification step in Stage 3 of Algorithm 3 did succeed in the sense that the  $\mathcal{G}$  component of  $x_j$  is wiped out.

#### 4. ROUNDING ERROR ANALYSIS

It is clear that in exact arithmetic, the implicit restart and the purification operations implicitly perform multiplication by  $S$ . In this section we provide a detailed analysis of the effect of round-off error in several of the important operations when  $S$  satisfies Theorem 1 and assuming that the relative round-off errors on the calculation of  $Sv_j$ , the Gram-Schmidt orthogonalisation and QR factorisation are much smaller than 1. To be precise we shall discuss the following results :

- (a) the  $\mathcal{N}$  and  $\mathcal{G}$  component in  $v_j$  may increase as  $j$  increases ;
- (b) the purification operation (12) produces an approximate eigenvector with no  $\mathcal{G}$  component but which may have a component in  $\mathcal{N}$  arising from a  $\mathcal{G}$  component in the unpurified vector. Similarly (16) produces an approximate eigenvector with no  $\mathcal{N}$  and  $\mathcal{G}$  component. Both purification steps may fail when  $|\theta^{-1}| \gg \epsilon_M^{-1}$  where  $\epsilon_M$  denotes the machine precision.
- (c) one implicit restart in  $B$ -orthogonal Arnoldi produces a  $\bar{G}_k$  that is not corrupted by  $\mathcal{N}$  or  $\mathcal{G}$  components, and a  $W_{k+1}$  which has no  $\mathcal{G}$  component. This IRA-step may fail when  $\|R_{k+1}^{-1}\| \gg \epsilon_M^{-1}$ .

The combination of the results in (a) and (b) produce the rounding error analysis for Ericsson’s method (Algorithm 2), which we state as follows.

*Rounding Error Conclusion for Algorithm 2 :*

- i. The  $B$ -orthogonal Arnoldi method can produce spurious eigenvalues in  $H_k$ .
- ii. The eigenvectors purified using (16) have no significant components in  $\mathcal{N}$  and  $\mathcal{G}$  if the corresponding eigenvalues of  $H_k$  are much larger than the machine precision.

The combination of the results in (b) and (c) produce the rounding error analysis for our proposed method (Algorithm 3) :

*Rounding Error Conclusion for Algorithm 3 :*

- i. If  $\|R_{k+1}^{-1}\| \ll \epsilon_M^{-1}$ , one implicit restart in  $B$ -orthogonal Arnoldi with purification using (12) produces a  $\bar{G}_k$  uncorrupted by  $\mathcal{G}$  or  $\mathcal{N}$  components with the consequence that no spurious eigenvalues are found.
- ii. In addition, the eigenvectors purified using (12) have no significant components in  $\mathcal{N}$  and  $\mathcal{G}$  if the corresponding eigenvalues are much larger than the machine precision.

(Note that, as in all analyses of round-off error, statements like *removes the  $\mathcal{N}$  components in the computed eigenvector* should really be interpreted as *removes the significant parts of round-off error in the  $\mathcal{N}$  components in the computed eigenvector*.)

The numerical results in our example agree with these statements, and we now proceed to justify them. We introduce the projectors  $\mathcal{P}_{\mathcal{R}}$ ,  $\mathcal{P}_{\mathcal{N}}$ , and  $\mathcal{P}_{\mathcal{G}}$  that map a

vector into  $\mathcal{R}$ ,  $\mathcal{N}$ , and  $\mathcal{G}$  respectively, and so an arbitrary vector  $x$  can be decomposed uniquely as  $x = \mathcal{P}_{\mathcal{R}}x + \mathcal{P}_{\mathcal{N}}x + \mathcal{P}_{\mathcal{G}}x$ . Note that

$$(21) \quad \mathcal{P}_{\mathcal{N}}S = S\mathcal{P}_{\mathcal{G}} \quad \text{and} \quad \mathcal{P}_{\mathcal{G}}S = 0 .$$

Throughout this section, round-off errors are denoted by Greek characters, the machine precision is denoted by  $\epsilon_M$ .

*Basic linear algebra operations in finite precision arithmetic.* The basic operations that we use are matrix-vector multiplication, the solution of linear systems by Gaussian elimination, the Gram-Schmidt orthogonalisation and the Householder QR-factorisation. From classical round-off error analysis, we know that  $c = x^T y$  with  $x, y \in \mathbb{R}^l$ , computed in finite precision arithmetic by the classical dot-product algorithm (see Algorithm 1.1.1 in Golub and Van Loan [7]), results in

$$c = x^T y + \xi \quad \text{with} \quad |\xi| \leq l\|x\|_2\|y\|_2\epsilon_M + O(\epsilon_M^2) ,$$

see Formula (2.4.10) in Golub and Van Loan [7]. A natural extension is the calculation of  $y = Fx$  with  $F \in \mathbb{R}^{m \times l}$ ,  $x \in \mathbb{R}^l$  and  $y \in \mathbb{R}^m$  by the dot-product matrix vector multiplication Algorithm 1.1.3 in [7], such that

$$y = Fx + \xi \quad \text{with} \quad \|\xi\|_2 \leq l\|F\|_F\|x\|_2\epsilon_M + O(\epsilon_M^2) \leq l^{3/2}\|F\|_2\|x\|_2\epsilon_M + O(\epsilon_M^2) .$$

This formula is valid for full matrices  $F$ . If  $F$  is sparse, the factor  $l$  in the bound for  $\|\xi\|_2$  should be replaced by the maximal number of nonzero elements in a row of  $F$ . An extension to the matrix-matrix product  $X = ZY$  with  $Z \in \mathbb{R}^{m \times l}$ ,  $Y \in \mathbb{R}^{l \times p}$  and  $X \in \mathbb{R}^{m \times p}$  by the dot-product matrix-matrix multiplication Algorithm 1.1.6 in [7], is

$$X = ZY + \Xi \quad \text{with} \quad \|\Xi\|_2 \leq l\|Z\|_F\|Y\|_F\epsilon_M + O(\epsilon_M^2) \leq l^{3/2}\|Z\|_2\|Y\|_F\epsilon_M + O(\epsilon_M^2) .$$

Consider now the errors generated by Gram-Schmidt orthogonalisation and the solution of linear systems. We shall not give rigorous upper bounds for these operations, since this is not the goal of this paper, but merely assume that these operations produce small relative round-off errors. For the analysis of the error propagation in the solution of linear systems by Gaussian elimination with partial pivoting, we refer to Golub and Van Loan [7, Equation (3.5.2)]. Björck [1] has shown that the modified Gram-Schmidt orthogonalisation of a set of vectors  $W_k = V_k R_k$  produces a  $V_k$  with

$$V_k^T V_k = I + \Gamma \quad \text{with} \quad \|\Gamma\|_2 \leq \gamma\epsilon_M \kappa_2(W_k) , \quad \gamma = O(1) .$$

To increase the numerical stability of the orthogonalisation process, one often uses modified Gram-Schmidt with reorthogonalisation [4] or even orthogonal transformations [14]. The Householder orthogonalisation of  $W_k$  produces a  $V_k$  with  $\|\Gamma_k\|_2 \leq \gamma\epsilon_M$ . We furthermore assume that  $\|W_k - V_k R_k\|_2 \leq \omega\epsilon_M\|W_k\|_2$  with  $\omega = O(1)$ . We shall not give rigorous upper bounds for the  $B$ -orthogonalisation of a set  $W_k$ , but merely assume that after this operation,

$$V_k^T B V_k = I + \Gamma \quad \text{with} \quad \|\Gamma\|_2 \leq \gamma\epsilon_M \quad \text{and} \quad \gamma = O(1) ,$$

$$\|W_k - V_k R_k\|_2 \leq \omega\epsilon_M\|W_k\|_2 \quad \text{with} \quad \omega = O(1) .$$

Finally, to simplify the error bounds, we shall use an upper bound for  $\|\bar{H}_k\|_2$  where  $\bar{H}_k$  is defined by (11). Since  $B$  is positive semidefinite, there always exists  $P \in \mathbb{C}^{N \times N}$  such that  $B = P^T P$ . Since

$$\bar{H}_k = V_{k+1}^T B A^{-1} B V_k = (P V_{k+1})^T P A^{-1} P^T (P V_k)$$

and  $V_{k+1}^T B V_{k+1} = (P V_{k+1})^T (P V_{k+1}) = I$  implies that  $\|P V_{k+1}\|_2 = 1$ , it follows that

$$(22) \quad \|\bar{H}_k\|_2 \leq \|\tilde{S}\|_2 \quad \text{with} \quad \tilde{S} = P A^{-1} P^T .$$

Note that  $S$  and  $\tilde{S}$  have the same eigenvalues, but different eigenvectors. Also note that  $\|\bar{H}_k\|_F \leq \sqrt{k} \|\bar{H}_k\|_2 \leq \sqrt{k} \|\tilde{S}\|_2$ .

We shall now proceed with the justification of the three statements above.

**4.1. Proof of (a).** We shall not derive rigorous upper bounds for the error growth in the Arnoldi process, since this is not the objective of the paper, but rather express the errors of the purification and the implicit restart in terms of errors obtained by the Arnoldi process (cf. analysis in [10]). This is also motivated by the fact that these errors depend on many factors like the sparsity of  $A$  and  $B$ , the matrix-vector multiplication used, the linear system solver, and the orthogonalisation process. The (computed) Arnoldi vector  $v_{j+1}$  satisfies the following relations :

$$(23a) \quad h_{j+1,j} v_{j+1} = S v_j - \sum_{i=1}^j h_{ij} v_i + \psi_j ,$$

$$(23b) \quad h_{ij} = v_i^T B S v_j + \delta_{ij} ,$$

$$(23c) \quad v_i^T B v_j = 1 + \gamma_{ij} , \quad j = 1, \dots, i+1 .$$

The rounding errors  $\psi_j \in \mathbb{R}^N$  and  $\delta_{ij}, \gamma_{ij} \in \mathbb{R}$  arise from the matrix-vector multiplication by  $B$ , the solution of linear systems with  $A$  and the Gram-Schmidt orthogonalisation of  $S v_j$ . In block form, the equations (23) are written as

$$(24a) \quad V_{k+1} \bar{H}_k = S V_k + \Psi_{k+1} ,$$

$$(24b) \quad V_{k+1}^T B V_{k+1} = I + \Gamma_{k+1} ,$$

$$(24c) \quad \bar{H}_k = V_{k+1}^T B S V_k + \Delta_k .$$

The norms  $\|\Psi_{k+1}\|_2$ ,  $\|\Gamma_{k+1}\|_2$  and  $\|\Delta_k\|_2$  characterise the round-off errors in the Arnoldi process. In the following, we assume that  $\|\Delta_k\|_2 \ll \|\bar{H}_k\|_2 \leq \|\tilde{S}\|_2$ ,  $\|\Psi_{k+1}\|_2 \ll \|\tilde{S}\|_2$  and  $\|\Gamma_{k+1}\|_2 \ll 1$ . This automatically assumes a proper implementation of the  $B$  inner product, such that  $x^T B y$  is unaffected by any  $\mathcal{N}$  component in  $x$  and  $y$ .

The portion of  $v_{j+1}$  in  $\mathcal{N}$  and  $\mathcal{G}$  is found by applying the projectors  $\mathcal{P}_{\mathcal{N}}$  and  $\mathcal{P}_{\mathcal{G}}$  to (23a) and using (21) :

$$h_{j+1,j} \mathcal{P}_{\mathcal{N}} v_{j+1} = S \mathcal{P}_{\mathcal{G}} v_j - \sum_{i=1}^j h_{ij} \mathcal{P}_{\mathcal{N}} v_i + \mathcal{P}_{\mathcal{N}} \psi_j ,$$

$$h_{j+1,j} \mathcal{P}_{\mathcal{G}} v_{j+1} = \sum_{i=1}^j h_{ij} \mathcal{P}_{\mathcal{G}} v_i + \mathcal{P}_{\mathcal{G}} \psi_j .$$

Whatever the sizes of  $\|\mathcal{P}_{\mathcal{N}} \psi_j\|$  and  $\|\mathcal{P}_{\mathcal{G}} \psi_j\|$ , there is no reason to assume that  $\|\mathcal{P}_{\mathcal{N}} v_{j+1}\|$  and  $\|\mathcal{P}_{\mathcal{G}} v_{j+1}\|$  do not grow.

**4.2. Proof of (b), single purification by (12).** The purified  $x_j$  computed by  $x_j = V_{k+1} \bar{H}_k z_j$  with  $\|z_j\|_2 = 1$  satisfies

$$(25a) \quad \mathcal{P}_{\mathcal{N}} x_j / \theta_j = \theta_j^{-1} (S \mathcal{P}_{\mathcal{G}} V_k z_j + \mathcal{P}_{\mathcal{N}} \xi_j) ,$$

$$(25b) \quad \mathcal{P}_{\mathcal{G}} x_j / \theta_j = \theta_j^{-1} \mathcal{P}_{\mathcal{G}} \xi_j ,$$

with

$$(26) \quad \|\xi_j\|_2 \leq 3k^{3/2}\|V_{k+1}\|_F\|\tilde{S}\|_2\epsilon_M + \|\psi_{k+1}\|_2 + O(\epsilon_M^2) .$$

Observe that  $\|\xi_j\|$  consists of small terms compared to  $x_j$ . Note that if  $\|z_j\|_2 = 1$ , then  $\|V_k z_j\|_B \simeq 1$ , and  $\|x_j\|_B \simeq \|V_{k+1}\bar{H}_k z_j\|_B \simeq \theta_j$ . Thus, by dividing the left- and right-hand side in (25) by  $\theta_j$ , we obtain the relative contributions of the  $\mathcal{N}$  and  $\mathcal{G}$  spaces in  $x_j$ . Clearly, if  $\theta_j$  is small, the right-hand sides in (25) are large, while the left-hand sides remain of order 1. In this case, purification can fail. On the other hand, if  $|\theta_j^{-1}|\|\xi_j\|$  is small compared to 1, the  $\mathcal{N}$  component in  $x_j$  mainly arises from the  $\mathcal{G}$  component in  $V_k z_j$  and the  $\mathcal{G}$  component in  $x_j$  is wiped out.

*Proof.* The purified  $x_j := SV_k z_j$  computed by (12), satisfies

$$(27) \quad x_j = V_{k+1}\bar{H}_k z_j + \pi_j$$

with  $V_k$  and  $\bar{H}_k$  satisfying (24), and where  $\pi_j$  is the round-off error arising from the computation of (12). Hence,  $x_j = V_{k+1}(\bar{H}_k z_j + \pi'_j) + \pi''_j$  with  $\pi_j = V_{k+1}\pi'_j + \pi''_j$ , where  $\pi'_j$  and  $\pi''_j$  are the round-off errors for the calculation of  $\bar{H}_k z_j$  and  $V_{k+1}(\bar{H}_k z_j)$  respectively. With  $\|z_j\|_2 = 1$ , we then find using (22)

$$\begin{aligned} \|\pi'_j\|_2 &\leq k\|\bar{H}_k\|_F\|z_j\|_2\epsilon_M + O(\epsilon_M^2) \leq k^{3/2}\|\tilde{S}\|_2\epsilon_M + O(\epsilon_M^2) , \\ \|\pi''_j\|_2 &\leq (k+1)\|V_{k+1}\|_F\|\bar{H}_k z_j\|_2\epsilon_M + O(\epsilon_M^2) \\ &\leq (k+1)\|V_{k+1}\|_F\|\tilde{S}\|_2\epsilon_M + O(\epsilon_M^2) . \end{aligned}$$

Thus,

$$\|\pi_j\|_2 \leq (k^{3/2} + k + 1)\|V_{k+1}\|_F\|\tilde{S}\|_2\epsilon_M + O(\epsilon_M^2) .$$

Observing that  $k^{3/2} + k + 1 \leq 3k^{3/2}$  for  $k \geq 1$ , we get the first term in (26). At first sight, it is not obvious why (27) provides the necessary cancellation of spurious  $\mathcal{G}$  components, but this is in fact easily shown : from (24a), it follows that

$$(28) \quad SV_k z_j = V_{k+1}\bar{H}_k z_j - \Psi_{k+1} z_j ,$$

and by combining (27) and (28), we obtain  $x_j = SV_k z_j + \xi_j$  with  $\xi_j = \pi_j + \Psi_{k+1} z_j$  and  $\|\xi_j\|_2$  bounded by (26). Hence, using (21),

$$\begin{aligned} \mathcal{P}_N x_j &= \mathcal{P}_N SV_k z_j + \mathcal{P}_N \xi_j = S\mathcal{P}_G V_k z_j + \mathcal{P}_N \xi_j \\ \mathcal{P}_G x_j &= \mathcal{P}_G SV_k z_j + \mathcal{P}_G \xi_j = \mathcal{P}_G \xi_j . \end{aligned}$$

From these equations, (25) follow. □

**4.3. Proof of (b), double purification with (16).** *The purified  $x_j$  computed by  $x_j = V_{k+2}\bar{H}_{k+1}\bar{H}_k z_j$  with  $\|z_j\|_2 = 1$  satisfies*

$$(29a) \quad \|\mathcal{P}_N x_j / \theta_j^2\|_2 \leq |\theta_j^{-2}|(\|\mathcal{P}_N \Psi_{k+2}\|_2\|\tilde{S}\|_2 + \|\mathcal{P}_G \Psi_{k+1}\|_2\|\tilde{S}\|_2 + \|\mathcal{P}_N \pi_j\|_2),$$

$$(29b) \quad \|\mathcal{P}_G x_j / \theta_j^2\|_2 \leq |\theta_j^{-2}|(\|\mathcal{P}_G \Psi_{k+2}\|_2\|\tilde{S}\|_2 + \|\mathcal{P}_G \pi_j\|_2)$$

with

$$(30) \quad \|\pi_j\|_2 \leq 4(k+1)^{3/2}\|V_{k+2}\|_F\|\tilde{S}\|_2^2\epsilon_M + O(\epsilon_M^2) .$$

Note that if  $\|z_j\|_2 = 1$ , then  $\|V_k z_j\|_B \simeq 1$ , and  $\|x_j\|_B \simeq \|V_{k+2}\bar{H}_{k+1}\bar{H}_k z_j\|_B \simeq \theta_j^2$ . Thus, by dividing the left- and right-hand side in (25) by  $\theta_j^2$ , we obtain the relative contributions of the  $\mathcal{N}$  and  $\mathcal{G}$  spaces in  $x_j$ . Clearly, if  $\theta_j$  is small, the purification

does not work. Observe that both the  $\mathcal{N}$  and  $\mathcal{G}$  components in  $V_k z_j$  are wiped out if  $\theta_j^{-2}$  is small compared to 1.

*Proof.* We first observe that  $x_j := S^2 V_k z_j$  is computed in finite precision from (16) such that

$$x_j = V_{k+2} \bar{H}_{k+1} \bar{H}_k z_j + \pi_j .$$

Similar to the proof of (26), we find that for  $\|z_j\|_2 = 1$ ,

$$\|\pi_j\|_2 \leq (k^{3/2} + (k+1)^{3/2} + (k+2)) \|V_{k+2}\|_F \|\tilde{S}\|_2^2 \epsilon_M + O(\epsilon_M^2) .$$

Observing that  $k^{3/2} + (k+1)^{3/2} + k + 2 \leq 4(k+1)^{3/2}$ , (30) follows. From (24a) it follows that

$$\begin{aligned} S^2 V_k z_j &= S V_{k+1} \bar{H}_k z_j - S \Psi_{k+1} z_j \\ &= V_{k+2} \bar{H}_{k+1} \bar{H}_k z_j - \Psi_{k+2} \bar{H}_k z_j - S \Psi_{k+1} z_j . \end{aligned}$$

Hence

$$x_j = V_{k+2} \bar{H}_{k+1} \bar{H}_k z_j + \pi_j = S^2 V_k z_j + \Psi_{k+2} \bar{H}_k z_j + S \Psi_{k+1} z_j + \pi_j ,$$

from which (29) follow. □

**4.4. Proof of (c).** *After the implicit restart,*

$$(31a) \quad \|\mathcal{P}_{\mathcal{N}} W_{k+1}\|_2 \leq \|S \mathcal{P}_{\mathcal{G}} V_{k+1}\|_2 \|R_{k+1}^{-1}\|_2 + \|\mathcal{P}_{\mathcal{N}} \Xi_{k+1}\|_2 + O(\epsilon_M^2) ,$$

$$(31b) \quad \|\mathcal{P}_{\mathcal{G}} W_{k+1}\|_2 \leq \|\mathcal{P}_{\mathcal{G}} \Xi_{k+1}\|_2 + O(\epsilon_M^2) ,$$

$$\|\bar{G}_k - W_{k+1}^T B S W_k\|_2 \leq (2(k+2)^{3/2} \|V_{k+2}\|_F \|P\|_2 + \omega + \gamma + (k+2)^2) \|\tilde{S}\|_2 \epsilon_M$$

$$(31c) \quad + \|\Delta_{k+1}\|_2 + O(\epsilon_M^2) ,$$

with  $\omega, \gamma = O(1)$  and

$$\begin{aligned} \|\Xi_{k+1}\|_2 &\leq (k+2)^{3/2} \|V_{k+2}\|_F \epsilon_M \\ &\quad + \{\omega \|V_{k+2}\|_2 \|\tilde{S}\|_2 \epsilon_M + \|\Psi_{k+2}\|_2\} \|R_{k+1}^{-1}\|_2 + O(\epsilon_M^2) . \end{aligned}$$

Observe that the right-hand side of (31c) consists of small terms, such that  $\bar{G}_k \simeq W_{k+1}^T B S W_k$ . The right-hand sides in (31a) and (31b) are small if  $\|R_{k+1}^{-1}\|_2$  is small. If  $\|R_{k+1}^{-1}\|_2 \simeq 1$ ,  $W_{k+1}$  has no significant component in  $\mathcal{G}$  but still contains significant components in  $\mathcal{N}$  coming from the  $\mathcal{G}$  component of  $V_{k+2}$  (see the first term in the right-hand side of (31a)). The action of the  $B$ -inner product in  $W_{k+1}^T B S W_k$  eliminates any contribution from the  $\mathcal{N}$  component of  $W_{k+1}$ . Hence, since  $\bar{G}_k \simeq W_{k+1}^T B S W_k$ ,  $\bar{G}_k$  is free from significant errors in  $\mathcal{N}$  and  $\mathcal{G}$ .

If  $\|R_{k+1}^{-1}\|_2$  is large, then  $\|\Xi_k\|_2 \gg \epsilon_M$  and  $W_{k+1}$  can have large components in  $\mathcal{G}$ . Hence,  $G_k \simeq W_k^T B S W_k$  is corrupted by the  $\mathcal{G}$  components of  $W_k$  which can lead to spurious eigenvalues.

*Proof.* Let the QR-factorisation of  $\bar{H}_{k+1}$  (in finite precision arithmetic) be

$$(32) \quad \bar{H}_{k+1} = \bar{Q}_{k+1}R_{k+1} + \Omega_{k+1}, \quad \bar{Q}_{k+1}^T \bar{Q}_{k+1} = I + \Gamma_{k+1},$$

with  $\bar{Q}_{k+1} \in \mathbb{R}^{k+2 \times k+1}$ ,  $R_{k+1} \in \mathbb{R}^{k+1 \times k+1}$ ,  $\Omega_{k+1} \in \mathbb{R}^{k+2 \times k+1}$ ,  $\Gamma_{k+1} \in \mathbb{R}^{k+1 \times k+1}$ , and

$$(33) \quad W_{k+1} = V_{k+2} \bar{Q}_{k+1} + \Pi_{k+1},$$

with  $\Pi_{k+1} \in \mathbb{R}^{N \times k+1}$ . Following the introduction on basic matrix operations, we assume that

$$\begin{aligned} \|\Omega_{k+1}\|_2 &\leq \omega \|\bar{H}_{k+1}\|_2 \epsilon_M \leq \omega \|\tilde{S}\|_2 \epsilon_M, \quad \omega = O(1), \\ \|\Gamma_{k+1}\|_2 &\leq \gamma \epsilon_M, \\ \|\Pi_{k+1}\|_2 &\leq (k+2) \|V_{k+2}\|_F \|\bar{Q}_{k+1}\|_F \epsilon_M \leq (k+2)^{3/2} \|V_{k+2}\|_F \epsilon_M + O(\epsilon_M^2). \end{aligned}$$

Note that we assume here that  $V_{k+2} \bar{Q}_{k+1}$  is computed as the product of two matrices by the dot-product algorithm. In practice, one would rather use a factorised form in terms of Givens or Householder transformations, which would give another bound.

The extension of (24a) is

$$SV_{k+1} = V_{k+2} \bar{H}_{k+1} - \Psi_{k+2}$$

and by applying (32) and (33) in that order,

$$(34) \quad \begin{aligned} SV_{k+1} &= V_{k+2} \bar{Q}_{k+1} R_{k+1} + V_{k+2} \Omega_{k+1} - \Psi_{k+2} \\ &= W_{k+1} R_{k+1} - \Pi_{k+1} R_{k+1} + V_{k+2} \Omega_{k+1} - \Psi_{k+2}, \end{aligned}$$

so

$$W_{k+1} = SV_{k+1} R_{k+1}^{-1} + \Xi_{k+1}$$

with  $\Xi_{k+1} = \Pi_{k+1} - (V_{k+2} \Omega_{k+1} - \Psi_{k+2}) R_{k+1}^{-1}$ . Thus, using (21),

$$\begin{aligned} \mathcal{P}_N W_{k+1} &= \mathcal{P}_N SV_{k+1} R_{k+1}^{-1} + \mathcal{P}_N \Xi_{k+1} = S \mathcal{P}_G V_{k+1} R_{k+1}^{-1} + \mathcal{P}_N \Xi_{k+1} \\ \mathcal{P}_G W_{k+1} &= \mathcal{P}_G SV_{k+1} R_{k+1}^{-1} + \mathcal{P}_G \Xi_{k+1} = \mathcal{P}_G \Xi_{k+1}. \end{aligned}$$

The analysis for  $\bar{G}_k$  is given next. In the implicit restart step,  $\bar{G}_k$  is computed (in finite precision arithmetic) as

$$\bar{G}_k = R_{k+1} \bar{Q}_k + \Sigma_k,$$

with  $\bar{Q}_k \in \mathbb{R}^{k+1 \times k}$  and

$$\|\Sigma_k\| \leq (k+1) \|R_{k+1}\|_F \|\bar{Q}_k\|_F \epsilon_M \leq (k+1)^{3/2} \|\tilde{S}\|_2 \sqrt{k} \epsilon_M \leq (k+2)^2 \|\tilde{S}\|_2 \epsilon_M,$$

if  $\bar{G}_k$  is computed by the dot-product matrix-vector multiplication instead of using orthogonal transformations. By applying (32), (24c) and (33) in this order,

$$\begin{aligned} \bar{G}_k &= \bar{Q}_{k+1}^T \bar{H}_{k+1} \bar{Q}_k - \Gamma_{k+1} R_{k+1} \bar{Q}_k - \bar{Q}_{k+1}^T \Omega_{k+1} \bar{Q}_k + \Sigma_k \\ &= \bar{Q}_{k+1}^T V_{k+2}^T B S V_{k+1} \bar{Q}_k + \bar{Q}_{k+1}^T \Delta_{k+1} \bar{Q}_k - \Gamma_{k+1} R_{k+1} \bar{Q}_k \\ &\quad - \bar{Q}_{k+1}^T \Omega_{k+1} \bar{Q}_k + \Sigma_k \\ &= W_{k+1}^T B S W_k - \bar{Q}_{k+1}^T V_{k+2}^T B S \Pi_k - \Pi_{k+1}^T B S W_k + \bar{Q}_{k+1}^T \Delta_{k+1} \bar{Q}_k \\ &\quad - \Gamma_{k+1} R_{k+1} \bar{Q}_k - \bar{Q}_{k+1}^T \Omega_{k+1} \bar{Q}_k + \Sigma_k. \end{aligned}$$

$\|\Pi_{k+1}^T BSW_k\|_2$  can be bounded by

$$\begin{aligned} \|\Pi_{k+1}^T BSW_k\|_2 &= \|(P\Pi_{k+1})^T P A^{-1} P^T (PW_k)\|_2 \\ &\leq \|P\Pi_{k+1}\|_2 \|\tilde{S}\|_2 \leq (k+2)^{3/2} \|V_{k+2}\|_F \|P\|_2 \|\tilde{S}\|_2 \epsilon_M + O(\epsilon_M^2). \end{aligned}$$

Note that  $\|PW_k\|_2 = 1$ , since  $W_k^T B W_k = I$ . There is a similar bound for  $\|\bar{Q}_{k+1}^T V_{k+2}^T B S \Pi_k\|_2$ . Hence

$$\begin{aligned} \|\bar{G}_k - W_{k+1}^T BSW_k\|_2 &\leq 2(k+2)^{3/2} \|V_{k+2}\|_F \|P\|_2 \|\tilde{S}\|_2 \epsilon_M + \|\Delta_{k+1}\|_2 + \gamma \|\tilde{S}\|_2 \\ &\quad + \omega \|\tilde{S}\|_2 \epsilon_M + (k+2)^2 \|\tilde{S}\|_2 \epsilon_M + O(\epsilon_M^2). \quad \square \end{aligned}$$

The final, and for this paper the most important, result concerns the round-off error in the implicitly restarted  $B$ -orthogonal Arnoldi method with purification (Algorithm 3), and is given by the ‘Round-off Error Conclusion’ at the beginning of this section. The proof consists of combining the arguments in (b) and (c) and is omitted.

**4.5. A test for spurious eigenvalues.** The preceding error analysis and the numerical results for  $k = 15$  and  $20$  in Section 3.3 show that numerical problems can occur when  $\|R_{k+1}^{-1}\|$  and/or  $|\theta_j|^{-1}$  are very large. We show an important relation between these two quantities in the following theorem. We complete this section with a discussion on the advantages of Algorithm 3 over Algorithm 2.

First note that in exact arithmetic  $R_{k+1} = \bar{Q}_{k+2}^T V_{k+2}^T B S V_{k+1}$  and, since  $S$  is singular, it is certainly possible for  $R_{k+1}$  to have small or zero singular values. Second, there is a connection between  $\|R_{k+1}^{-1}\|_2$  and the small eigenvalues of  $H_{k+1}$ . This is given by the following theorem.

**Theorem 4.** *Suppose that  $H_k$  has  $k$  independent eigenpairs, denoted by  $(\theta_j, z_j)$  with  $\|z_j\|_2 = 1$ ,  $j = 1, \dots, k$ . Define the residual semi-norm*

$$\rho_j = \|S V_k z_j - \theta_j V_k z_j\|_B.$$

Then

$$(35) \quad \|R_k^{-1}\|_2 \geq 1 / \min_j \sqrt{|\theta_j|^2 + \rho_j^2}.$$

*Proof.* Since  $\|R_k^{-1}\|_2 = 1/\sigma_{\min}$  with  $\sigma_{\min}$  the smallest singular value of  $R_k$ , we shall derive an upper bound for  $\sigma_{\min}$ . From the definition of  $\sigma_{\min}$ , one has

$$\sigma_{\min} \leq \|R_k z_j\|_2 = \|\bar{H}_k z_j\|_2, \quad \|z_j\|_2 = 1.$$

Also, from the recurrence relation (10) it follows that

$$S V_k z_j - \theta_j V_k z_j = S V_k z_j - V_k H_k z_j = h_{k+1,k} v_{k+1} e_k^T z_j,$$

and with  $\|v_{k+1}\|_B = 1$ ,  $\rho_j = h_{k+1,k} |e_k^T z_j|$ . Hence

$$\|\bar{H}_k z_j\|_2 = \left\| \begin{pmatrix} \theta_j z_j \\ h_{k+1,k} e_k^T z_j \end{pmatrix} \right\|_2 = (|\theta_j|^2 + \rho_j^2)^{1/2}. \quad \square$$

The relation (35) provides a lower bound for  $\|R_{k+1}^{-1}\|_2$ . It is clear that if the eigenvalues of  $S$  are computed accurately, the smallest eigenvalue of  $H_{k+1}$  can provide a sharp lower bound for  $\|R_{k+1}^{-1}\|_2$ . This also explains that both a small  $\theta_j$  and a large  $\|R_{k+1}^{-1}\|_2$  often occur in tandem. This observation is important because it indicates that failure of IRA and purification often occur together. However, there is a subtle distinction between failure of purification and IRA. Purification

fails only for those eigenvectors  $x_j$  corresponding to small eigenvalues  $\theta_j$ , since the accuracy of the purification only depends on  $\theta_j^{-1}$ . Corruptions emerge in IRA, when  $\max_j(|\theta_j^{-1}|)$  is large, and will affect all the eigenvalues and eigenvectors. This is clearly a disadvantage of the IRA-approach. Nevertheless, we have several reasons to prefer Algorithm 3 over Algorithm 2.

First, if  $\|R_{k+1}^{-1}\|_2$  is small,  $G_k$  does not produce spurious eigenvalues. This is not the case for the Ericsson approach as was shown by the example in Section 3.3, since, in fact,  $H_k$  is not ‘purified’ at all. Second, it is easy to check when spurious eigenvalues might emerge in IRA by monitoring  $\|R_{k+1}^{-1}\|_2$  for succeeding values of  $k$ . Such a test is quite cheap. A similar test for the Ericsson approach is not provided. Of course, it is possible to check possible failure of purification for a specific eigenvector  $x_j$  by looking at  $|\theta_j|^{-1}$ .

The analysis in this section, shows the importance of a small  $\|\Psi_{k+1}\|_2$  in (24a), i.e. assumes accurate matrix-vector operations with  $S$  and Gram-Schmidt orthogonalisation implemented with care. This emphasises the need for a robust reorthogonalisation strategy. Finally, the choice of initial vector  $v_1$  in  $\mathcal{R}$  can help in the reduction of  $\|R_{k+1}^{-1}\|_2$ , since in exact arithmetic, if  $v_1 \in \mathcal{R}$ , then  $V_{k+2} \in \mathcal{R}$ , and thus

$$\|R_{k+1}^{-1}\|_2 = 1/\sigma_{\min}(\bar{Q}_{k+2}^T V_{k+2}^T (BS)|_{\mathcal{R}} V_{k+1}) ,$$

where  $(BS)|_{\mathcal{R}}$  is the restriction of  $BS$  over  $\mathcal{R}$ . Since  $(BS)|_{\mathcal{R}}$  is not singular in  $\mathcal{R}$ ,  $\|R_{k+1}^{-1}\|_2$  should not be large, in general.

## 5. CONCLUSIONS

We end this paper with a summary of the main points and some brief comments on related work.

The main conclusion to draw is that the Algorithm 3 which does two implicit multiplications by  $S$ , is less susceptible to round-off errors than the algorithm in Ericsson [5] for the shift-invert operator  $S$  derived from problems with block-structure (2). Possible severe corruption can be detected by monitoring  $\|R_{k+1}^{-1}\|_2$ . In fact, Algorithm 3 requires two fewer matrix-vector multiplications than the one in [5] (since Algorithm 3 needs no pre-multiplication of the initial vector  $v_1$  by  $S^2$ ) with the only (minor) addition of the  $QR$ -step for  $\bar{H}_{k+1}$  and the formation of  $W_{k+1}$ . Nevertheless, we do recommend the pre-multiplication in order to reduce the risk of large  $\|R_{k+1}^{-1}\|$ .

On an aesthetic level, the purification approach of [10] using (12) rather than direct multiplication by  $S$  was very elegant. We think that the use of implicitly restarted Arnoldi to carry out implicitly a second multiplication by  $S$  is similarly pleasing.

Note that one could avoid the need for the purification step at stage 3 of Algorithm 3 by performing two steps of IRA. In fact, for *any*  $S$  which is singular with generalised null vectors of grade 2 we see that 2 steps of IRA with standard inner product will remove elements from  $\mathcal{N}$  and  $\mathcal{G}$ . This can clearly be extended to cases where grade  $r$  generalised null vectors arise when  $r$  IRA steps would be required. If  $S$  arises from a generalised problem with  $B$  of the form (2) then one fewer IRA step is needed with the addition of a purification step.

It appears that the IRA method may help to stabilise calculations of eigenvalues (which is a known property of this method). The analysis here, where  $S$  is singular and defective helps to illustrate why this might be the case since there is an implicit

subspace iteration step (see statement 2 in Theorem 3), which is another nice feature of this method. In fact, Philippe and Sadkane [11] use a few steps of subspace iteration in an explicit way to purify the Arnoldi vectors in an application coming from Maxwell's equations.

The theoretical analysis of iterative methods for the reduced problem  $S_1 u = \theta u$  helps in an understanding of the performance of iterative methods for  $Sx = \theta x$ . The fact that implicitly restarted  $B$ -orthogonal Arnoldi applied to  $S$  works could, following the analysis in Section 2.2, also be explained by analysing the implicitly restarted  $M$ -orthogonal Arnoldi method applied to  $S_1$ , since one implicit matrix-vector multiplication by  $S_1$  maps the initial vector into  $\mathcal{R}_1$ .

Finally note that the  $B$ -inner product has been used in subspace iteration by Scott [12]. Ericsson [5] and Nour-Omid, Parlett, Ericsson and Jensen [10] mention that it is also a useful tool when  $B$  is nonsingular, but very ill conditioned.

## 6. ACKNOWLEDGEMENTS

The authors are grateful to the British Council and the Belgian National Fund for Scientific Research (NFWO) for supporting the collaboration between the Katholieke Universiteit Leuven and the University of Bath.

We also wish to thank Dan Sorensen for reporting an error in the example in Section 3.3. The authors are grateful to the referee for his suggestions that made the paper much clearer and more complete.

## REFERENCES

1. Å. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalisation*, BIT **7** (1967), 1–21. MR **35**:5126
2. K.A. Cliffe, T.J. Garratt, and A. Spence, *Eigenvalues of the discretized Navier-Stokes equation with application to the detection of Hopf bifurcations*, Advances in Computational Mathematics **1** (1993), 337–356. MR **94k**:65052
3. ———, *Eigenvalues of block matrices arising from problems in fluid mechanics*, SIAM J. Matrix Anal. Applic. **15** (1994), 1310–1318. MR **95g**:65053
4. J.W. Daniel, W.B. Gragg, L. Kaufman, and G.W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization*, Math. Comp. **30** (1976), 772–795. MR **55**:4638
5. T. Ericsson, *A generalised eigenvalue problem and the Lanczos algorithm*, Large Scale Eigenvalue Problems (J. Cullum and R.A. Willoughby, eds.), Elsevier Science Publishers BV, 1986, pp. 95–119. MR **88c**:65035
6. T. Ericsson and A. Ruhe, *The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems*, Math. Comp. **35** (1980), 1251–1268. MR **83b**:65034
7. G. Golub and C. Van Loan, *Matrix computations*, 2nd ed., The Johns Hopkins University Press, 1989. MR **90d**:65055
8. R.G. Grimes, J.G. Lewis, and H.D. Simon, *A shifted block Lanczos algorithm for solving sparse symmetric generalised eigenproblems*, SIAM J. Matrix Anal. Applic. **15** (1994), 228–272. MR **94k**:65054
9. D.S. Malkus, *Eigenproblems associated with the discrete LBB condition for incompressible finite elements*, Internat. J. for Engrg. Sci. **19** (1981), 1299–1310. MR **83k**:73053
10. B. Nour-Omid, B.N. Parlett, T. Ericsson, and P.S. Jensen, *How to implement the spectral transformation*, Math. Comp. **48** (1987), 663–673. MR **88f**:65062
11. B. Philippe and M. Sadkane, *Improving the spectral transformation block Arnoldi method*, Second IMACS Symposium on Iterative Methods in Linear Algebra (P.S. Vassilevski and S.D. Margenov, eds.), IMACS Series in Computational and Applied Mathematics, vol. 3, IMACS Symposium on Iterative Methods in Linear Algebra, 1996, pp. 57–63.

12. D.S. Scott, *The advantages of inverted operators in Rayleigh-Ritz approximations*, SIAM J. Sci. Statist. Comput. **3** (1982), 68–75. MR **83e**:65072
13. D.C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Applic. **13** (1992), 357–385. MR **92i**:65076
14. H.F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Statist. Comput. **9** (1988), 152–163. MR **88m**:65056

LMS NUMERICAL TECHNOLOGIES, INTERLEUVENLAAN 70, 3001 HEVERLEE, BELGIUM  
*E-mail address:* `km@lmsnit.be`

SCHOOL OF MATHEMATICAL SCIENCES, UNIVERSITY OF BATH, CLAVERTON DOWN, BATH BA2  
7AY, UNITED KINGDOM  
*E-mail address:* `A.Spence@maths.bath.ac.uk`