

# The Arnoldi method in the light of Homotopic Deviation theory

Françoise Chaitin-Chatelin<sup>1</sup>

CERFACS Technical Report TR/PA/03/15

<sup>1</sup>Université Toulouse 1 and CERFACS, 42 Ave. G. Coriolis, 31057 Toulouse  
Cedex 1, France. E-mail: chatelin@cerfacs.fr

## Abstract

This paper aims at providing an algorithmic understanding of the “convergence” of Krylov-type methods which relies on asymptotic properties at 0 and  $\infty$ . The classical normwise (or analytic) perturbation approach corresponds to the limit towards 0. We complement this analysis by the structural perturbation approach provided by the limit to  $\infty$  in the Homotopic Deviation theory. We easily get back the classical results when the homotopy parameter  $h$  tends to 0. But we get new results by letting  $|h| \rightarrow \infty$ , leading to new insights on long standing algorithmic issues.

**Key words:** Homotopic Deviation, critical point, kernel point, Hessenberg form, irreducible, derogatory, incomplete Arnoldi decomposition, Arnoldi residual, restarted Arnoldi, inner-outer iterations, Cauchy interlace theorem.

# 1. Introduction

In this paper, we use the conceptual framework of Homotopic Deviation theory [4, 10, 13] to analyse two important questions which arise in the context of Krylov-type methods, namely:

- i*) the arithmetical behaviour of the Arnoldi residual, and
- ii*) the extreme robustness of such methods to large perturbations.

The first question, although easily overlooked by software developers, is important in our opinion because it signals a weak spot in our algorithmic understanding. It was first reported by Bennani in 1991 and not much progress has been made since, [6].

The second question emerged gradually during the 1990s with the notion of inexact methods. The partial answers given so far are not satisfactory because they rely on analytic perturbation theory under (sometimes hidden) assumptions of convergence. We believe, on the contrary, that Krylov-type methods are best understood by a structural perturbation approach [3, 12, 7].

This is the reason why we revisit the incomplete Arnoldi decomposition by means of the tools provided by the theory of Homotopic Deviation presented in [4, 5, 9, 10]. Classical perturbation results are obtained when the homotopy parameter  $h$  is small enough and tends to 0. New results are obtained by letting the homotopy parameter be unbounded. This allows to capture global as well as local effects [4, 13].

An important feature of the deviation matrix involved in an incomplete Arnoldi decomposition is that it is *nilpotent* [9]. Therefore the main assumption in [4] is not satisfied, and the results in [4] cannot be transferred readily. The necessary adaptation of the theory to the Arnoldi context will be done in Section 5. It emphasizes the algorithmic role of the irreducible Hessenberg form in Krylov methods.

Before that, the incomplete Arnoldi decomposition is briefly presented in Section 2, the spectral structure of an irreducible Hessenberg matrix is analysed in Section 3, and Homotopic Deviation theory is recalled in Section 4.

## 2. The incomplete Arnoldi decomposition

### 2.1. Definition

Let  $A \in \mathbb{C}^{n \times n}$ . For  $1 < k < n$ , the incomplete (or inexact) Arnoldi decomposition for  $A$  can be written as

$$(A - hU)V = VH_k,$$

where  $V = [v_1, \dots, v_k] \in \mathbb{C}^{n \times k}$  is the Krylov basis at step  $k$ ,  $H_k \in \mathbb{C}^{k \times k}$  is in Hessenberg form,  $U = v_{k+1} v_k^H$  is nilpotent:  $v_k^H v_{k+1} = 0$  implies  $U^2 = 0$ . And  $h = h_{k+1, k}$  denotes the  $(k+1, k)$  element in  $H_{k+1}$ .

We remark that  $h$  is real positive with a (classical or modified) Gram-Schmidt orthogonalisation strategy. With a Householder strategy,  $h$  can be complex.

When  $k = n$ , the exact decomposition ( $h = 0$ ) is completed with  $AV = VH_n$ ,  $V \in \mathbb{C}^{n \times n}$ , where  $H_n = V^H AV$  is one possible Hessenberg form for  $A$ . We denote by  $\sigma(A)$  the spectrum (set of eigenvalues) of the matrix  $A$ ,  $\rho(A)$  is the spectral radius of  $A$ , and  $\|\cdot\|$  represents the euclidean norm (vector or induced).  $\text{lin}(v_i)$  denotes the linear span of the vectors  $v_i$ .  $\text{re}(A) = \mathbb{C} - \sigma(A)$  is the resolvent set.

### 2.2. Relation between $\sigma(A)$ and $\sigma(H_k)$ , $1 < k < n$

The reader is referred to [9, 12] and to [16], chapter 2, p. 41–42, where it is shown that

- $\sigma(H_k) \subset \sigma(A - hU)$ ,  $1 < k < n$ ,
- $|h|$  represents the homotopic distance between  $\sigma(A)$  and  $\sigma(H_k)$ ,
- the  $n$  eigenvalues of  $A - hU$  lie on the homotopic  $\frac{1}{|h|}$ -level curve  $\Gamma$  defined by :

$$\Gamma = \{z \in \mathbb{C} - \sigma(A); \rho(U(A - zI)^{-1}) = \frac{1}{|h|}\}.$$

As a consequence,  $|h|$  measures the backward error for the incomplete Arnoldi decomposition, which consists in replacing  $A$  by  $H_k$  [1].

### 2.3. Spectral consequences of $A = VH_nV^H$ ( $k = n$ )

This paragraph summarizes results given in [3, 12, 16]. Irreducible Hessenberg matrices are nonderogatory (one Jordan block per distinct eigenvalue, [1], chapter 6). Therefore the Arnoldi decomposition on a derogatory matrix  $A$  yields a reducible Hessenberg matrix  $H_n$ .

The algorithmic assumption that  $H_n$  is irreducible implies that:

- 1) if  $H_n$  is simple, then  $A$  is simple with  $n$  *distinct* eigenvalues. Multiple eigenvalues escape.
- 2) if  $H_n$  is defective, then  $A$  is nonderogatory: multiple Jordan blocks for the same eigenvalue are out of reach.

However, these limitations are only valid in exact arithmetic. They are bypassed by finite precision as experience tells us.

In algorithmic practice, the goal is to force near reducibility to occur as soon as possible (that is for  $k$  small with respect to  $n$ ). Understanding the process of numerical (or happy) breakdown — the finite precision counterpart of exact reducibility — is therefore crucial to the design of efficient stopping criteria [3, 12].

**Methodological remark 2.1.** We wish to comment again [1, 3], at this point, that Krylov-type methods are *finite* methods. They require at most  $n$  steps to deliver the exact matrix  $H_n$  which is sought for (from  $H_n$ , eigenvalues or solutions as the case may be, are derived for  $A$ ).

The current trend amongst Numerical Analysts to talk about the “convergence” of such methods is to be taken as a loose metaphor, meaningful only when  $n$  is very large with respect to  $k$ .

The fact that Krylov-type methods are intrinsically finite algorithms is well illustrated by their robustness to large perturbations. Such a robustness is not exhibited by truly asymptotic methods such as Newton-type methods [7, 8]. As one proceeds towards convergence, the admissible perturbations must be of decreasing norm in Newton-type methods.

### 2.4. The Arnoldi residual, $1 < k < n$

$(\lambda, y)$  denotes an eigenpair for  $H = H_k$ ,  $1 < k < n$ . This yields

$$Hy = \lambda y, \quad y \in \mathbb{C}^k,$$

$$(A - \lambda I)Vy = V(H - \lambda I)y + hUVy, \quad Vy \in \mathbb{C}^n.$$

The pair  $(\lambda, Vy)$  is a pseudoeigenpair for  $A$  corresponding to the residual

$$r = (A - \lambda I)Vy = hUVy = hv_{k+1}(e_k^T y). \quad (1)$$

We set  $y_k = e_k^T y$ , the  $k$ th (and last) component for  $y$ . Equality (1) is valid in exact arithmetic, but not always in finite precision. Therefore we introduce the following distinction:

$r_D = (A - \lambda I)Vy$  is the *direct residual* for  $A$ ,

$r_A = (hy_k)v_{k+1}$  is the *Arnoldi residual*.

The backward error for  $A$  at  $(\lambda, Vy)$  is

$$BE(\lambda, Vy) = \frac{\|r_D\|}{\|A\| \|y\|} = \frac{|hy_k|}{\|A\| \|y\|}. \quad (2)$$

It is a fact of experience [6, chapter 6, p. 89–90], [12], that the mathematical equality (2), which derives from (1), may not hold for iterations which follow the one for which the backward error reaches machine precision. Therefore the Arnoldi residual, of norm  $|hy_k|$ , is not a reliable indicator for “convergence”, once machine precision is reached for  $\frac{|hy_k|}{\|A\| \|y\|}$ . The fact that numerically subtle events take place in the vicinity of “convergence” has been known for more than 2 decades [15, chapter 13].

One immediately remarks that the residual in (1) has an absolute formulation, suitable for a mathematical forward analysis of convergence. By comparison, the normalised residual in (2) is suited for a backward analysis of algorithmic convergence in finite precision.

In this paper, we choose the theory of Homotopic Deviation to take a fresh look at the phenomenon of near reducibility.

### 3. Spectral structure of an irreducible Hessenberg matrix

#### 3.1. An inductive analysis for $1 < k < n$

$H \in \mathbb{C}^{k \times k}$  is a Hessenberg matrix of order  $k$  assumed to be *irreducible*. We consider the Hessenberg matrix of order  $k + 1$  defined by

$$H^+ = \left( \begin{array}{c|c} H & u \\ \mathbf{0} & h \end{array} \middle| \begin{array}{c} a \\ a \end{array} \right)$$

where  $u \in \mathbb{C}^k$ ,  $h$  and  $a \in \mathbb{C}$ . What can be said about the eigenstructure of  $H^+$ ? We consider the eigenpair  $(\mu, x)$  for  $H^+$ , with  $x = (y^T, \alpha)^T \in \mathbb{C}^{k+1}$ .  $H^+x = \mu x$  implies

$$\begin{cases} Hy + \alpha u = \mu y, \\ hy_k + a\alpha = \mu\alpha. \end{cases}$$

By assumption,  $H^+$  is irreducible for  $h \neq 0$ , and  $x$  is the only eigendirection associated with  $\mu$ . We discuss whether  $\alpha \neq 0$  or not. If  $\alpha \neq 0$ , we can set  $\alpha = 1$  by proper normalisation of  $x$ . Note that  $\alpha = e_{k+1}^T x = x_{k+1}$ .

1)  $\alpha = 0$  implies

$$\begin{cases} (H - \mu I_k)y = 0, \\ hy_k = 0. \end{cases}$$

$y$  is the (unique) eigendirection for  $H$  associated with  $\mu$ . The condition  $hy_k = 0$  implies either  $h = 0$  ( $H^+$  reducible) or  $y_k = 0$ . We show now that  $y_k = 0$  contradicts the assumption that  $H$  is irreducible.

We set  $y = (y^{-T}, 0)^T$  with  $y^- \in \mathbb{C}^{k-1}$ , and

$$H = \left( \begin{array}{c|c} H^- & v \\ \mathbf{0} & l \end{array} \middle| \begin{array}{c} v \\ b \end{array} \right)$$

with  $H^-$  of order  $k - 1$ ,  $v \in \mathbb{C}^{k-1}$ ,  $l$  and  $b \in \mathbb{C}$ .  $Hy = \mu y$ , together with  $y_k = 0$ , implies that

$$Hy = \mu y \iff \begin{cases} H^- y^- = \mu y^-, \\ ly_{k-1} = 0 \end{cases} \text{ with } y_{k-1} = e_{k-1}^T y^-.$$

Therefore  $ly_{k-1} = 0$  with  $l \neq 0$ , hence  $y_{k-1} = 0$ . Finite induction on  $k$  leads to  $y = 0$ , which is impossible.

If  $\alpha = 0$ , the only possibility is that  $h = 0$  :  $H^+$  is reducible.

2)  $\alpha = 1$  implies

$$\begin{cases} (H - \mu I_k)y = -u, \\ hy_k = \mu - a. \end{cases} \quad (3)$$

When nonzero, the quantity  $hy_k$  measures the forward error for  $H^+$  on the eigenvalue  $\mu \in \sigma(H^+)$  when it is approximated by  $a$ .

When  $h \neq 0$ ,  $hy_k = 0$  implies  $y_k = 0$ . And the consequences of  $|y_k| \sim 0$  will be discussed later (Section 7).

If the matrix  $A$  is assumed nonderogatory, then  $H_k$  is irreducible for  $k = 2, \dots, n-1$ . Therefore  $H^+ (= H_{k+1})$  is irreducible and we only have to consider the case  $\alpha = 1$ .

We conclude that when  $(\mu, x)$  is an exact eigenpair for  $H^+$ , then

*i)*  $x = (y^T 1)^T$ , where  $y$  corresponds to one step of inverse iteration on  $H$  with  $\mu \in \sigma(H^+)$  as an approximate eigenvalue for  $H$ , and  $-u$  being the residual  $(H - \mu I_k)y$ , that is,

$$y = -(H - \mu I_k)^{-1}u,$$

*ii)*  $hy_k$  equals the forward error  $\mu - a$ . It can be zero either for  $y_k = x_k = e_k^T x = 0$  or for  $h = 0$ ,

*iii)*  $|h_{k+2k+1}| = \|(A - \mu I)Vx\|$  is the Arnoldi residual at step  $k+1$ , since  $x_{k+1} = e_{k+1}^T x = 1$ .

These observations show how the quantities  $h, a$  and  $u$  which represent new information about  $A$  not present in  $H$  are processed algorithmically to update the eigendecomposition of  $H$  into that of  $H^+$ , during the Arnoldi decomposition at step  $k+1$ .

### 3.2. $h$ as a homotopy parameter

Consider the matrix

$$B = \left( \begin{array}{c|c} H & u \\ \hline \mathbf{0} & a \end{array} \right)$$

of order  $k+1$ , with spectrum  $\sigma(B) = \sigma(H) \cup \{a\}$ , and  $1 < k < n$ . The matrix

$$H^+ = \left( \begin{array}{c|c} H & u \\ \hline \mathbf{0} & h \end{array} \right)$$

can be written as  $H^+ = B + hE = B(h)$  with  $E = e_{k+1} e_k^T$ . Note that  $E^2 = 0$  because  $e_k^T e_{k+1} = 0$ :  $E$  is *nilpotent*.

To study the dependence of the spectrum of  $H^+$  on the parameter  $h$ , the framework of Homotopic Deviation  $(B, E)$ , where  $E$  is rank one and nilpotent, is most natural [4, 10]. The homotopy parameter  $h$  will be considered *complex*. Of particular interest will be the limits of the eigenvalues  $\lambda(h)$  as  $|h| \rightarrow \infty$ .

**Remark 3.1** In view of Section 3.1, a hasty conclusion is that the case  $h \rightarrow 0$  is the only situation with any numerical significance. It will become clear why this view is short sighted, and unable to capture such important features of Krylov-type methods as their robustness to large perturbations.

To be explained convincingly, this robustness requires a backward analysis for  $|h|$  large, as counter-intuitive as it may seem. . .

For another very different example where the limit to  $\infty$  for the homotopy parameter receives a natural interpretation, the reader is referred to the Acoustics problem treated in [13]. The complex homotopy parameter is taken to be the inverse of the complex impedance parameter which tends to 0.

## 4. Summary of Homotopic Deviation theory

### 4.1. The scope

The theory of Homotopic Deviation analyses, from a computational point of view, the vector spectral field  $t \in \mathbb{C} \rightarrow \lambda_i(t) \in \mathbb{C}$ ,  $i = 1, \dots, n$ , where  $\lambda_i(t)$  is an eigenvalue of the matrix family  $A(t) = A + tE$ . Here,  $A$  and  $E$  are given matrices in  $\mathbb{C}^{n \times n}$  and  $t$  is a *complex* parameter,  $t \neq 0$ .

This theory is a general model to perform a backward information analysis in Inexact Computing, where direct information on  $A$  is not accessible by computation. One only has an algorithmic access to information related to the family  $A(t) = A + tE$ , where  $E$  is a given matrix called *deviation*, and  $t$  is a complex parameter which can vary continuously in  $\mathbb{C}$ ,  $t \neq 0$ , [10, 11].

The name ‘‘Homotopic Deviation’’ comes from the facts that

- i)*  $A$  is modified by  $tE$  which retains the same structure as  $t \in \mathbb{C}$ , and

ii) there is no restriction on the norm  $|t| \|E\|$ . This norm can be as *large* as one wishes.

The classical homotopic theory is obtained by taking  $t \in \mathbb{R}$ ,  $0 \leq t \leq 1$ . The importance of having a *complex* homotopy parameter will become clear in the next paragraph. We restrict the use of the word ‘‘perturbation’’ to the most studied situation where  $|t| \|E\|$  is bounded. For example, classical perturbation theory corresponds to the limit as  $|t| \rightarrow 0$  [1, 2].

## 4.2. Computational connection between $t$ and $z = \lambda(t)$

The relationship between  $t$  and  $\lambda(t)$  relies on the multiplicative representation:

$$A + tE - zI = (I + tE(A - zI)^{-1})(A - zI)$$

which is valid for  $z \notin \sigma(A)$ .

$z \notin \sigma(A)$  is an eigenvalue for  $A(t)$  iff  $A + tE - zI$  is singular, or equivalently, iff  $I - tF_z$  is singular, where  $F_z = -E(A - zI)^{-1}$  has the eigenvalue  $\mu_z \in \mathbb{C}$ .

Therefore  $z \notin \sigma(A)$  is an eigenvalue  $\lambda(t)$  of  $A(t)$  for some  $t \neq 0$  iff there exists an eigenvalue  $\mu_z \neq 0$  for  $F_z$  such that

$$t\mu_z = 1. \tag{4}$$

**It is clear from (4) that the possibility for  $\mu_z$  to be complex implies the necessity of considering  $t$  complex to satisfy (4).**

When  $E$  is full rank,  $r = \text{rank } E = \text{rank } F_z = n$ , then  $\mu_z$  is nonzero for any  $z \notin \sigma(A)$ . Therefore the correspondence (4) is well defined. However, when  $r = \text{rank } E = \text{rank } F_z < n$ ,  $F_z$  has at least  $n - r$  zero eigenvalues, and the correspondence (4) may not be well defined for all  $z \notin \sigma(A)$ .

Indeed, if for some  $z = \xi$ ,  $F_\xi$  is nilpotent, that is

$$\rho(F_z) = \max_{\mu_z \in \sigma(F_z)} |\mu_z| = 0,$$

then  $t$  is unbounded, and the correspondence (4) is singular at  $z = \xi$ .

The values  $\xi \notin \sigma(A)$  such that  $F_\xi$  is nilpotent are called the *critical points* of  $(A, E)$ . Their existence is studied in [4] under the assumption that

$0 \in \sigma(E)$  is semisimple. It is shown, in particular, that when  $r = 1$ , but  $E^2 \neq 0$ , there exist at most  $n - 1$  critical points.

The computational importance of critical points has several reasons [4]. One of these is the fact that

$$\xi \text{ critical} \implies F_\xi \text{ nilpotent, } F_\xi^{\delta+1} = 0, F_\xi^\delta \neq 0, 1 \leq \delta \leq r.$$

Therefore, there exist no  $t$  in  $\mathbb{C}$  such that  $\xi \notin \sigma(A)$  is an eigenvalue of  $A(t)$ . On the contrary, for any  $t$  in  $\mathbb{C}$ ,

$$(A + tE - \xi I)^{-1} = (A - \xi I)^{-1} \sum_{k=0}^{\delta} (tF_\xi)^k$$

is a polynomial in  $t$  of degree  $\delta$  between 1 and  $r$ .

### 4.3. The limit points for $\sigma(A(t))$ as $|t| \rightarrow \infty$

The correspondence (4) suggests to look for  $\lim_{|t| \rightarrow \infty} \lambda(t)$ . Can we characterise those of the eigenvalues of  $A(t)$  which remain bounded as  $|t| \rightarrow \infty$  when  $r < n$ , as it appears possible by letting  $\mu_z \rightarrow 0$  in (4)?

This question has been addressed in [4], with  $E = UV^H$ ,  $U, V \in \mathbb{C}^{n \times r}$  of rank  $r$ , under the assumption that  $G = V^H U$  has full rank  $r$ .

The points  $\xi \in \mathbb{C}$  at finite distance such that  $\lim_{|t| \rightarrow \infty} \lambda(t) = \xi$  are called the *limit points* of  $(A, E)$ .

At a limit point  $\xi$  in  $\text{re}(A)$ ,  $0 \in \sigma(F_\xi)$  has an ascent  $\geq 2$ . We note for future reference, that for  $z \notin \sigma(A)$ , the eigenvalues of  $F_z$  which are not necessarily 0 are that of  $M_z = -V^H(A - zI)^{-1}U$ .

When  $G$  is full rank, 0 is a semisimple eigenvalue for  $E$  with multiplicity  $n - r$ . Let  $P$  be the corresponding eigenprojection for  $E$  on  $\text{Ker } E$ . The limit points for  $(A, E)$  are the  $n - r$  eigenvalues of the matrix  $\Pi$  of order  $n - r$ , which represents the Petrov approximation  $PAP$  of  $A$ , restricted to  $\text{Ker } E$ , [4].

**Remark 4.1** In general, the eigenprojection  $P$  need not be orthogonal. When  $P$  is an oblique (resp. orthogonal) projection,  $\Pi$  is the Petrov (resp. Galerkin) approximation of  $A$  associated with  $P$  [1].

The situation when  $G$  is *rank deficient* ( $\text{rank } G < r$ ) is studied in [5] for the general case  $r = \text{rank } E$ ,  $1 \leq r < n$ . In this paper, we restrict our investigation to the case  $r = 1 = \text{rank } E$ , with  $E^2 = 0$ , which covers the situation resulting from an analysis of the incomplete Arnoldi decomposition.

**5.  $E = uv^H$ , and  $v^H u = 0$ ,  $u, v \in \mathbb{C}^n$**

**5.1.  $E = e_n e_{n-1}^T$**

For  $u, v \in \mathbb{C}^n$  such that  $v^H u = 0$ , consider the unitary basis

$$Q = \left[ X, \frac{v}{\|v\|}, \frac{u}{\|u\|} \right] \text{ in } \mathbb{C}^n,$$

with  $X^H X = I_{n-2}$ ,  $v^H X = u^H X = 0$ . Set  $E = uv^H$  with  $v^H u = 0$ . The matrix  $A(t) = A + tE$  is unitarily equivalent to  $C(t) = C + tD$ , with

$$D = \|u\| \|v\| e_n e_{n-1}^T = Q^H E Q.$$

We can, therefore, without loss of generality, restrict our attention to the deviation  $E = e_n e_{n-1}^T$ . From now on, in this Section,  $E$  is assumed to be  $e_n e_{n-1}^T$ .

$E$  is nilpotent, with only eigenvalue 0 with multiplicity  $n$  and structure  $(0^1)^{n-2}(0^2)$ . The Jordan chain associated with 0 double defective is  $(e_n, e_{n-1})$ :

$$E e_n = 0 \text{ and } E e_{n-1} = e_n.$$

**5.2. The four sets of interest for  $(A, E)$ ,  $E = e_n e_{n-1}^T$**

Let  $P$  be the *orthogonal* projection on  $W_{n-2} = \text{lin}(e_1, \dots, e_{n-2})$  which represents the eigenspace for  $E$  associated with 0 of ascent 1, multiplicity  $n-2$ .

$A_{n-2} = P A P$  represents the section (principal submatrix) of  $A$  of order  $n-2$ . We define the partitioning  $(n-2, 2)$  of  $A$  as

$$A = \left( \begin{array}{c|c} A_{n-2} & R \\ \hline S & A_2 \end{array} \right)$$

with  $R, S^T \in \mathbb{C}^{(n-2) \times 2}$  and  $A_2$  of order 2. We assume that  $\sigma(A) \cap \sigma(A_{n-2}) = \emptyset$ . And we consider the family  $A(t) = A + tE$ ,  $t \in \mathbb{C}$ .

The overview presented above tells us that four sets in  $\mathbb{C}$  are useful to study the homotopic deviation process  $(A, E)$ , where  $r = 1$  and  $E^2 = 0$ . These are

- i*) the set  $\sigma(A)$  of  $n$  eigenvalues for  $A$ ,
- ii*) the set  $\text{Lim}$  of *limit* points for  $\sigma(A(t))$  which remain at finite distance when  $|t| \rightarrow \infty$  such that  $\sigma_\infty(A, E) = \lim \sigma(A(t)) = \{\infty, \text{Lim}\}$ ,
- iii*) the set  $K_c(A, E)$  of *critical* points  $z$  which are such that  $\mu_z = 0$ , hence  $F_z^2 = 0$ ,
- iv*) the set  $\sigma(A_{n-2})$  of  $n - 2$  Ritz values for  $A$  associated with the eigenprojection  $P$  for  $E$ .

The general theory [5] applied to  $E$  **nilpotent** with  $r = 1$  entails that generically

$$\text{Lim} \cap \text{re}(A) = K_c(A, E)$$

contains at most  $n - 2$  critical points. Nongenerically, it is possible that  $K_c(A, E)$  is the continuous set  $\text{re}(A)$ , and  $\sigma(A) = \sigma_\infty(A, E)$  is invariant under  $t \in \mathbb{C}$ . We confirm some of these results by direct proof.

**Proposition 5.1** *Any point in the limit at finite distance of  $\sigma(A(t))$  which is not an eigenvalue of  $A$  is critical.*

**Proof** Let  $\xi = \lim_{|t| \rightarrow \infty} \lambda(t)$  such that  $\xi \notin \sigma(A)$ . Because  $r = 1$ , the relationship  $t\mu_z = 1$  for  $z = \lambda(t)$  forces  $|\mu_z|$  to be small for  $|t|$  large. Therefore  $\xi$  is the limit point of a family of spectral orbits  $\Sigma(|t|)$ ,  $|t| \rightarrow \infty$ . They enclose  $\xi$  which corresponds to a global minimum for  $z \rightarrow |\mu_z|$ , at the value 0:  $\rho(F_\xi) = 0$ , and  $F_\xi^2 = 0$ .  $\square$

**Proposition 5.2** *If  $A_2$  is not a lower triangle, exactly 2 eigenvalues of  $A(t)$  escape to  $\infty$ . The remaining  $n - 2$  converge to the eigenvalues of  $\Omega$ , where*

$$\Omega = A_{n-2} - \frac{1}{a_{n-1n}} u v^T,$$

with  $u = (a_{1n}, \dots, a_{n-2n})^T$  and  $v^T = (a_{n-11}, \dots, a_{n-1n-2})$ .

**Proof** We apply Lidskii's theory [5, 17] to our particular framework, where  $E$  has one nontrivial Jordan block (of size 2). When  $a_{n-1n} \neq 0$ ,  $\text{Lim} = \sigma(\Omega)$  consists of  $n - 2$  limit points (counting algebraic multiplicities).

Exactly 2 eigenvalues escape to  $\infty$  at the rate  $O(t^{1/2})$ , along the axis defined by the angle  $\theta$ , where  $2\theta = \text{Arg } a_{n-1} t$ .  $\square$

### 5.3. The structure of $F_z$ , for $z \notin \sigma(A)$

Because  $-F_z = e_n e_{n-1}^T (A - zI)^{-1}$ ,  $e_n$  is the eigenvector associated with  $\mu_z = -e_{n-1}^T (A - zI)^{-1} e_n$ , which is nonzero when  $z$  is not critical. In the generic case,  $F_z$  is semisimple: it has  $n$  independent eigenvectors.

When  $z = \xi \notin \sigma(A)$  is critical, however, the structure of  $F_z$  changes from semi-simple ( $z \neq \xi$ ) to defective and nilpotent:  $F_\xi^2 = 0$ : the eigenvector  $e_n$  is linked with another vector  $\alpha$  by the Jordan chain of length 2:  $F_\xi \alpha = e_n$ . This creates, at the critical points, a computational dependency which is not present at a generic  $z \notin \sigma(A)$ . This dependency is easy to explicit in the case corresponding to the incomplete Arnoldi decomposition described in Section 3, paragraph 3.2. This is the subject of Section 6.

## 6. The Arnoldi algorithm: $H^+ = B + hE$ of order $k + 1$ , $1 < k < n$

We go back to  $H^+ = B(h) = B + hE$  of order  $k + 1$ ,  $3 \leq k + 1 \leq n$ . We have

$$B = \left( \begin{array}{c|c} H & u \\ \mathbf{0} & a \end{array} \right)$$

where  $H$  is an *irreducible Hessenberg* matrix, and

$$(B - zI_{k+1})^{-1} = \left( \begin{array}{c|c} (H - zI_k)^{-1} & w_z \\ \mathbf{0} & (a - z)^{-1} \end{array} \right)$$

with

$$w_z = -\frac{1}{a - z} (H - zI_k)^{-1} u \in \mathbb{C}^k.$$

$\sigma(B) = \sigma(H) \cup \{a\}$ , we assume that  $a \notin \sigma(H)$ , that is  $B$  is nonderogatory.

### 6.1. The sets of interest for $(B, E)$

The section of  $B$  of order  $(k + 1) - 2 = k - 1$  is given by the irreducible Hessenberg matrix  $H^-$ . By assumption  $h^- = h_{k k-1} \neq 0$  then  $\sigma(H^-) \cap$

$\sigma(H) = \emptyset$ . Furthermore we assume that  $\sigma(B) \cap \sigma(H^-) = \emptyset$ , that is  $a \notin \sigma(H^-)$ .

The four sets of paragraph 5.2 become respectively:

- $\sigma(B) = \sigma(H) \cup \{a\}$ , the spectrum of  $B$ ,
- $\sigma_\infty = \sigma_\infty(B, E) = \{\infty, \text{Lim}\}$ ,
- $K_c(B, E)$ , the set of critical points,
- $\sigma^- = \sigma(H^-)$ .

We set  $\sigma^- = \sigma(H^-)$ . We also write  $\sigma^+ = \sigma(H^+)$  for the spectrum of  $H^+ = B + hE$ , for any  $h \neq 0$ ,  $h \in \mathbb{C}$ .

We set  $\beta_z = (B - zI_{k+1})^{-1} e_{k+1}$  which represents the last column of  $(B - zI_{k+1})^{-1}$ .  $\mu_z = -e_k^T \beta_z$  represents the only possibly non zero eigenvalue of  $F_z = -e_{k+1} e_k^T (B - zI_{k+1})^{-1}$ .

**Proposition 6.1** *If  $z \in \sigma(H^+) = \sigma^+$ , the vector  $\beta_z$  is an eigendirection for  $H^+$  associated with the eigenvalue  $z$ , normalised such that  $-h e_k^T \beta_z = h \mu_z = 1$ .*

**Proof** Clear by (3) in Section 3.1, where the quantities  $(\mu, x)$  are replaced by  $(z, \frac{1}{a-z} x = \beta_z)$ . The Arnoldi residual at step  $k + 1$  associated with the eigenpair  $(z, x)$  for  $H^+$  is  $r_A = (A - zI)Vx$ , and  $\|r_A\| = |h_{k+2k+1}|$ .  $\square$

We know from the general theory that  $z \in \text{re}(B)$  is an eigenvalue of  $H^+$  iff  $h\mu_z = 1$ . Proposition 5.1 applies: if  $z \in \text{Lim} \cap \text{re}(B)$ , then  $\mu_z = 0$ , that is  $z$  is critical:  $\text{Lim} \cap \text{re}(B) = K_c(B, E)$ . We write  $u = (\tilde{u}^T, u_k)^T$  with  $u_k = e_k^T u$ ,  $\tilde{u} \in \mathbb{C}^{k-1}$ .

As a corollary of Proposition 5.2, we get the

**Theorem 6.2** *If  $u_k = e_k^T B e_{k+1} \neq 0$ , exactly 2 eigenvalues  $\lambda(h)$  escape to  $\infty$ . The  $k - 1$  others converge to  $\text{Lim} = \sigma(\Omega)$ , with*

$$\Omega = H^- - \frac{h^-}{u_k} \tilde{u} e_{k-1}^T,$$

such that

$$\|H^- - \Omega\| = \left| \frac{h^-}{u_k} \right| \|\tilde{u}\|.$$

**Proof** Clear. If

$$C = |h^-| \frac{\|\tilde{u}\|}{|u_k|}$$

is small, then  $\text{Lim}$  is in the  $C$ -pseudospectrum of  $H^-$ .  $\square$

We observe that for  $u \neq 0$ ,  $\frac{\|\tilde{u}\|}{|u_k|} = \tan \psi$ , where  $\psi$  is the acute angle between the directions spanned by  $u$  and  $e_k$ . The condition  $u_k \neq 0$  is equivalent to  $0 \leq \psi < \pi/2$ , and  $\|\tilde{u}\| = 0 \iff \psi = 0 \iff \Omega = H^-$ , since  $h^- \neq 0$ .

The computational significance of Theorem 6.2 should not be underestimated. It shows that, from an algorithmic point of view, the spectral information about  $A$  given by  $H^-$  when  $|h|$  is large, can be *as meaningful as* the information given by  $H^+$  for  $|h|$  small. The robustness of the Arnoldi decomposition to large deviations stems from this powerful dual point of view.

## 6.2. Maximal range connection at a critical point

We set

$$\begin{aligned} \alpha_z^T &= e_k^T (B - zI_{k+1})^{-1} \\ &= (e_k^T (H - zI_k)^{-1}, e_k^T w_z) \end{aligned}$$

which represents the  $k$ th row of  $(B - zI_{k+1})^{-1}$ . At the critical points  $\xi$  in  $K_c(B, E) = \text{Lim} \cap \text{re}(B)$ ,  $F_\xi$  becomes nilpotent and the last component of  $\alpha_\xi$  vanishes, by the

**Lemma 6.3** *For  $\xi$  critical in  $K_c(B, E) = \text{Lim} \cap \text{re}(B)$ ,  $\alpha_\xi^T e_{k+1} = e_k^T \beta_\xi = 0$ . For  $z$  in  $\sigma^-$ ,  $\alpha_z^T e_k = 0$ .*

**Proof**

i)  $\alpha_\xi^T e_{k+1} = e_k^T \beta_\xi = -\mu_\xi = 0$  when  $\xi$  is critical:  $F_\xi e_{k+1} = 0$  and  $F_\xi^2 = 0$ .

ii)  $\alpha_z^T e_k = e_k^T (H - zI_k)^{-1} e_k$ , a quantity which is proportional to  $\det(H^- - zI_{k-1}) = 0$  for  $z \in \sigma(H^-)$ .  $\square$

We look at  $\text{Ker } F_z$  for  $F_z = -e_{k+1} e_k^T (B - zI_{k+1})^{-1}$ .

$$\text{Ker } F_z = \{x; e_k^T (B - zI_{k+1})^{-1} x = 0\} = \{x; \alpha_z^T x = 0\}.$$

**Lemma 6.4** *The irreducible Hessenberg form for the matrix  $H$  implies that  $F_z e_1 \neq 0$  for any  $z \notin \sigma(B)$ .*

**Proof** Because  $H$  is irreducible, and  $z \neq a$ , then for any  $z \notin \sigma(B)$ ,  $e_1$  cannot belong to  $\text{Ker } F_z$ . The proof relies on  $\text{Ker } F_z = \{\alpha_z\}^\perp$ . We assume that  $e_1 \in \text{Ker } F_z$ , that is,  $\alpha_z^T e_1 = 0$ , and get a contradiction. We define  $x = (B - zI_{k+1})^{-1} e_1$  for  $z \notin \sigma(B)$  which satisfies  $e_k^T x = x_k = 0$ . And we look at the system

$$Bx - zx = e_1$$

with  $x_j = e_j^T x$  for  $j = 1, \dots, k+1$ . The  $(k+1)^{\text{st}}$  equation yields  $(a-z)x_{k+1} = 0$ , hence  $x_{k+1} = 0$  since  $z \neq a$ . The  $k^{\text{th}}$  equation yields  $h_{kk-1}x_{k-1} = 0$ , thus  $x_{k-1} = 0$  because  $h_{kk-1} \neq 0$ . By induction on  $j$  until  $j = 2$ , we get  $x = 0$ , and the first equation is impossible. Therefore the assumption  $\alpha_z^T e_1 = 0$  for  $z \notin \sigma(B)$  leads to a contradiction.  $F_z e_1 \neq 0$  means that  $F_z e_1$  is proportional to  $e_{k+1}$ , with a coefficient  $-\alpha_z^T e_1$  which can never vanish for  $z \in \text{re}(B)$ .  $\square$

**Proposition 6.5** *For  $\xi$  critical,  $F_\xi^2 = 0$ . The Jordan chain associated with  $0 \in \sigma(F_\xi)$  double defective is  $(e_{k+1}, e_1)$ .*

**Proof** Direct consequence of Lemma 6.3 and 6.4.  $\square$

As  $z$  varies in  $\mathbb{C} - \sigma(B)$ , one gets the two possibilities:

- i) generically, for  $z$  not critical,  $e_{k+1}$  is an eigenvector for  $\mu_z \neq 0$ , which does not belong to  $\text{Ker } F_z = \{\alpha_z\}^\perp$ .  $F_z$  is semi-simple and all the  $k+1$  eigenvectors are linearly independent.
- ii) for  $z = \xi$  critical,  $e_{k+1} \in \text{Ker } F_\xi$  is the eigenvector associated with  $(0^2)$ , and  $e_1$ , such that  $F_\xi e_1$  is colinear with  $e_{k+1}$ , defines the generalised

eigendirection. The nilpotency of  $F_\xi$  expresses a structural connection between  $e_{k+1}$  and  $e_1$ .

We find enlightening to contrast the computational connections between the canonical basis vectors created respectively by the nilpotency of  $E$  and that of  $F_\xi$  at  $\xi \in K_c(B, E)$ . In both cases, the degree of nilpotency is 2. But for  $E$ , the connection between  $e_k$  and  $e_{k+1}$  has the *minimal* extent 1, whereas for  $F_\xi$ , the connection between  $e_1$  and  $e_{k+1}$  has the *maximal* extent  $k$ , when  $\xi$  is critical.

We find also important to relate this to the role of additive versus multiplicative representation in the broader framework of Qualitative Computing [11]. Applied to Numbers, the difference between these representations yields the crucial, if underestimated, *Newcomb-Borel paradox* [11].

In the present context of matrices, there are also two representations at work for  $H^+$ :

*i*) the additive representation:  $H^+ = B(h) = B + hE$ ,

*ii*) the multiplicative representation, for  $z \notin \sigma(B)$ ,

$$H^+ - zI_{k+1} = (I_{k+1} - hF_z)(B - zI_{k+1}). \quad (5)$$

In case *i*), the deviation matrix  $E = e_{k+1} e_k^T$  is given. Its nilpotency  $E^2 = 0$  expresses the short range connection between  $(e_k, e_{k+1})$ . Such a local knowledge is sufficient to derive the critical set  $\text{Lim} \cap \text{re}(B) = K_c(B, E)$ .

In case *ii*),  $F_z$  derives from the multiplicative representation (5) required to use algorithmically the Neumann series expansion to get  $(H^+ - zI_{k+1})^{-1}$ , whenever possible. Such a representation can be more complex than the previous additive representation. Indeed, at critical points, it turns the connection  $(e_k, e_{k+1})$  of extent 1 into the connection  $(e_1, e_{k+1})$  of extent  $k$ . This maximal range connection will be used in Section 7 to explain the reason for the success of the restart procedure for Arnoldi.

### 6.3. Criticality implies convergence in two steps at most

For  $\xi$  critical,  $(H^+ - \xi I_{k+1})^{-1}$  exists, for any  $h \in \mathbb{C}$ , and can be written as:

$$(H^+ - \xi I_{k+1})^{-1} = (B - \xi I_{k+1})^{-1} (I_{k+1} + hF_\xi) \quad (6)$$

since  $F_\xi^2 = 0$ .

The linear system

$$(H^+ - \xi I_{k+1}) a(h) = b \quad (7)$$

has a unique solution  $a(h)$  linear in  $h$ , for any  $h \in \mathbb{C}$ . Moreover, for any right hand side  $b \in \text{Ker } F_\xi$  the solution  $a(h) = a = (B - \xi I_{k+1})^{-1} b$  is independent of  $h$ .

We know that  $e_1^T b \neq 0$  implies  $b \notin \text{Ker } F_\xi$  by Lemma 6.4. Therefore, the condition  $e_1^T b \neq 0$  ensures that the solution  $a(h)$  of (7) depends linearly on  $h$ . On the contrary, a right hand side  $b$  such that  $\alpha_\xi^T b = 0$  guarantees that the solution is independent of  $h$  ( $b \in \text{Ker } F_\xi$ ).

#### 6.4. Small range connection for $z$ non critical

We suppose that  $z$  is not critical ( $\mu_z \neq 0$ ) and that  $z \notin \sigma(H^+)$ , that is  $h\mu_z \neq 1$ . We consider the resolution of (7), where  $\xi$  is replaced by  $z$ .  $F_z$  is semi simple and  $(I - hF_z)^{-1}$  is analytic around 0 and around  $\infty$  [5].

**Proposition 6.6** *For  $z \notin \sigma(H^+)$  such that  $\mu_z \neq 0$*

$$\begin{aligned} (H^+ - zI_{k+1})^{-1} &= (B - zI_{k+1})^{-1} - \frac{h}{1 - h\mu_z} \beta_z \alpha_z^T \\ &= (B - zI_{k+1})^{-1} \left( I_{k+1} + \frac{h}{1 - h\mu_z} F_z \right). \end{aligned}$$

**Proof** Simple calculation. Compare with (6).  $\square$

$E_z = \beta_z \alpha_z^T$  is the rank one matrix  $-(B - zI_{k+1})^{-1} F_z$ , such that  $\text{Im } E_z = \{\beta_z\}$  and  $\text{Ker } E_z = \text{Ker } F_z = \{\alpha_z\}^\perp$ . We observe that  $\alpha_z^T \beta_z = e_k^T (B - zI_{k+1})^{-2} e_{k+1}$ .

The case  $\mu_z = 0$  gives back (6) ( $z$  critical).

As  $|h| \rightarrow \infty$ , the coefficient  $-h/(1 - h\mu_z)$  tends to  $1/\mu_z$  defined iff  $\mu_z \neq 0$ .  $I_{k+1} - \frac{1}{\mu_z} F_z$  is a projection on  $\{\alpha_z\}^\perp$  along  $\{e_{k+1}\}$ . For  $z \in \sigma(H^-)$ ,  $E_z e_k = \beta_z (\alpha_z^T e_k) = 0$  by Lemma 6.3.

## 7. Algorithmic consequences

### 7.1. Approximation of eigenvalues

The dynamics of the approximation of  $\sigma(A)$  by  $\sigma(H)$  as  $k$  increases can be seen as a finite process, which is analysed as

- *exact* when  $k = n$  (that is  $h = 0$ ),
- *inexact* when  $1 < k < n$  (that is  $h \neq 0$ ).

The case  $k = n$  is clear. Let us look at the incomplete Arnoldi decomposition corresponding to  $1 < k < n$  with  $h \neq 0$ . At step  $k$ , there are three sets of interest:  $\sigma^- = \sigma(H^-)$ ,  $\text{Lim} = \sigma(\Omega)$  and  $\sigma^+ = \sigma(H^+)$ . When  $z \in \sigma(H^+)$  is *nearly critical* ( $|\mu_z|$  small and  $|h|$  large) and when the quantity

$$C = \frac{\|\tilde{u}\|}{|u_k|} |h^-|$$

is small, then  $z$ , which is close to an eigenvalue of  $\Omega$  in  $\text{Lim}$ , lies in the  $C$ -pseudospectrum of  $H^-$ . Despite the fact that  $|h|$  is large,  $z$  can be almost an eigenvalue for  $H^-$  when  $|h^-|$  is small enough. “Convergence”, that is near-reducibility, can happen for  $|h|$  large, in addition to the classical case  $|h|$  small.

Let  $z$  be destined to approximate the eigenvalue  $\lambda$  for the original matrix  $A$ . When  $z \in \sigma(H^+)$  is almost an eigenvalue for  $H^-$ , then all the relevant information (for  $h \neq 0$ ) about  $\lambda$  contained in  $H^+$  of order  $k + 1$ , for  $k < n$ , has already been extracted by  $H^-$  which is the Hessenberg form of order  $k - 1$ , obtained by the Arnoldi process, which is specified by  $A$  and a starting vector  $u$  defining the Krylov subspace  $K(A, u) = \text{lin}(u, Au, \dots, A^{k-1}u)$ .

We leave for a future report the analysis of the restart procedure. This consists in updating the starting vector  $u$  for  $K(A, u)$  to enrich it in approximate eigendirections for the eigenvalues of  $A$  which are sought for. This can be seen as an outer iteration on  $u$  based on an inner step which is finite ( $k < n$ ).

## 7.2. On the pseudoeigenpairs for $H_l$ , $l \geq k$ , deriving from an exact eigenpair for $H^-$

Let  $(\xi, p)$  be an exact eigenpair for  $H^-$ :

$$H^- p = \xi p, \quad p \neq 0, \quad p \in \mathbb{C}^{k-1}.$$

We set  $p_{k-1} = e_{k-1}^T p$ . And we define in  $\mathbb{C}^l$ ,  $l \geq k$ , the augmented vector  $\hat{\chi}_l = (p^T, 0)^T$ .

**Proposition 7.1** *The pair  $(\xi, \hat{\chi}_l)$  is a pseudoeigenpair for  $H_l$ ,  $l \geq k$ , corresponding to the residual vector  $(h^- p_{k-1}) e_k$  in  $\mathbb{C}^l$ .*

**Proof** Simple algebraic verification.  $h^- = h_{k, k-1}$  (denoted  $l$  in Section 3.1).  $\square$

**Corollary 7.2** *The pair  $(\xi, \hat{\chi}_l)$  cannot be improved by inverse iteration using the Hessenberg form  $H_l$  for  $l \geq k + 1$ .*

**Proof**

- i)* Set  $l = k + 1$ . By Proposition 6.6, the linear system  $(H^+ - \xi I_{k+1}) a(h) = e_k$  has a solution independent of  $h$ , since  $a_\xi^T e_k = 0$  for  $\xi \in \sigma(H^-)$ , by Lemma 6.3. This shows that  $\hat{\chi}_{k+1}$  cannot be improved.
- ii)* For  $l > k + 1$ , this is true also by induction.  $\square$

Corollary 7.2 is illustrated by the numerical experiment performed in [12].

## 7.3. The Arnoldi residual in finite precision

Corollary 7.2 is the key to understand the numerical behaviour of the Arnoldi residual in finite precision. When an eigenpair is optimal at step  $k_0$  (backward error  $\frac{|hy_k|}{\|y\| \|A\|}$  for  $k = k_0$  of the order of machine precision), it cannot be improved at step  $k_0 + 2$  and beyond. Insisting on increasing  $k$  unduly (beyond  $k_0 + 2$ ) only creates a numerical artefact, where the computed value  $\tilde{y}_k$  is much smaller than the exact value  $y_k$ , [6], p. 89–90. The appropriate alternative is to restart the procedure.

## 7.4. Algebraic justification for restart

We observe that  $(h^- p_{k-1}) v_k = (h^- p_{k-1}) V e_k$  is the Arnoldi residual for  $A$  at  $(\xi, V_{k-1} p)$  computed at the iteration  $k - 1$ , with  $h^- = h_{k k-1}$ . This corresponds, in Section 3, in replacing  $H$  (resp.  $k$ ) by  $H^-$  (resp.  $k - 1$ ).

The exact eigenpair  $(\xi, p)$  for  $H^-$ , or  $(\xi, V_{k-1} p)$  for  $A$ , generates a sequence of pseudoeigenpairs  $(\xi, \hat{\chi}_l)$  for  $H_l$ ,  $l \geq k$ , corresponding to the constant residual norm  $|h^- p_{k-1}|$ , independently of  $h (= h_{k+1 k}) \in \mathbb{C}$ .

Corollary 7.2 tells us that the successive pseudoeigenvectors  $\hat{\chi}_l$  cannot be improved by using the Hessenberg form for  $l \geq k + 1$ . Moreover, if

$$C = \frac{\|\tilde{u}\|}{|u_k|} |h^-|$$

is of the order of machine precision,  $|h|$  is large. The only way to progress is to restart to get a residual  $r$  such that  $e_1^T r \neq 0$ .

Looking for  $r$  with a nonzero first component is in essence what is performed during a *restart* of the incomplete Arnoldi decomposition.

It is remarkable that the above justification for the restart procedure, which is incorporated in most Krylov-type methods, is valid for any  $h$ . It does not require the assumption that  $h$  is small. On the contrary, finite precision effects such as the approximation  $\sigma(H^-) \stackrel{\approx}{\sim} \sigma(H^+)$  are often better explained by considering  $|h|$  large. The notion of near criticality for certain eigenvalues in  $\sigma(H^+)$  is a key element of the explanation.

This should be contrasted with the classical justification which is based on a notion of “convergence” as  $h \rightarrow 0$  [1, 2]. A convincing explanation for the convergence of Krylov methods in finite precision appears to require the use of the two *complementary* tools given below:

- i*) the *classical* notion of convergence as  $h \rightarrow 0$  which applies as long as the arithmetic can be regarded as exact,
- ii*) the new notion of *criticality* as  $|h| \rightarrow \infty$ , which takes care of the effects of finite precision when they cannot be ignored.

The concept of criticality is especially useful when the assumption that  $A$  is nonderogatory is unrealistic.

## 7.5. Inexact Krylov methods

Inexact Krylov methods use an inexact matrix-vector product, which amounts to replace  $A$  by  $A + \Delta A$ . Convergence can be maintained under relative perturbations as large as  $\frac{\|\Delta A\|}{\|A\|} \sim 1$ , provided that each global sweep starts with the information given by a residual computed to working accuracy [7, 14]. This fact is again well explained algebraically by Corollary 7.2.

This Section has illustrated how the classical analysis of Arnoldi by perturbation theory ( $h$  small  $\rightarrow 0$ ) can be fruitfully complemented by an algebraic approach where  $h$  is not small, and on the the contrary, can be unbounded. This allows to explain algorithmic issues which seem to lie out of the reach of the explanatory power of perturbation theory (when  $h \rightarrow 0$ ).

## 8. An interlude about hermitian matrices

A hermitian Hessenberg matrix is tridiagonal. It is tempting to ask the question: is there an interesting counterpart to Homotopic Deviation theory when one takes  $A$  and  $E$  to be hermitian with  $E$  positive semidefinite?  $0 \in \sigma(E)$  is semi-simple.

### 8.1. Criticality implies $r \leq \lfloor \frac{n}{2} \rfloor$ and $\delta = 1$

The framework of [4] applies without difficulty with  $E = UU^H$  ( $\text{rank } U^H U = r$ ) and becomes simpler. When  $A$  and  $E = UU^H$  are **hermitian**, so are  $\Pi = PAP_{\uparrow \text{Ker } E}$  and  $M_z = -U^H(A - zI)^{-1}U$  for  $z$  real ( $z = \bar{z}$  and  $(M_z)^H = M_{\bar{z}}$ ). Therefore the kernel points  $\xi$ , which are the eigenvalues of  $\Pi$ , are *real*. At such a real point  $\xi \in \mathbb{R}$ ,  $M_\xi$  is semi-simple and 0 is an eigenvalue for  $M_\xi$  (resp.  $F_\xi$ ) of ascent 1 (resp. 2). Therefore a kernel point can be critical ( $M_\xi$  nilpotent) only if  $r \leq \lfloor \frac{n}{2} \rfloor$ , hence  $\delta = 1$ .

When  $r > \lfloor \frac{n}{2} \rfloor$ , critical points cannot exist. The notion of kernel point, however, retains its full power.

For example, one easily gets an analogue of Proposition 6.1 for a tridiag-

onal symmetric matrix  $T$  by choosing  $E$  of rank 2:

$$E = [e_n e_{n-1}] \begin{bmatrix} e_{n-1}^T \\ e_n^T \end{bmatrix},$$

and letting  $|h| \rightarrow \infty$ .

Set

$$T = \left( \begin{array}{c|cc} T^- & & \mathbf{0} \\ \hline & l & \\ \hline l & b & h \\ \hline \mathbf{0} & h & a \end{array} \right) \in \mathbb{R}^{n \times n}.$$

As  $|h| \rightarrow \infty$ ,  $n - 2$  eigenvalues of  $T$  tend to  $\sigma(T^-)$ , and 2 eigenvalues escape to infinity.

## 8.2. The Cauchy interlace theorem revisited

This famous theorem interlaces the eigenvalues of a hermitian matrix  $A$  of order  $n$ , with the eigenvalues of any of its principal submatrices (or sections) [15, chapter 10].

For example,  $\sigma(A) = \{\lambda_i\}_1^n$  and  $\sigma(A_{n-1}) = \{\mu_i\}_1^{n-1}$ , where  $A_{n-1}$  is the section of  $A$  of order  $n - 1$ , are intertwined on the real line as follows:

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \dots \leq \mu_{n-1} \leq \lambda_n.$$

This mutual ordering by the two spectra  $\sigma(A)$  and  $\sigma(A_{n-1})$  can be easily interpreted in the context of the Homotopic Deviation  $(A, E)$  with  $E = e_n e_n^T$  of rank 1.

$\sigma(A_{n-1})$  is obtained in the limit, as  $|t| \rightarrow \infty$ , of  $n - 1$  eigenvalues  $\lambda_j(t)$  of  $A(t) = A + tE$ . The remaining  $n$ th eigenvalue escapes to infinity. When  $t$  is chosen real,  $A(t) = A + tE$  is hermitian and the eigenvalues  $\lambda_j(t)$  stay real as they converge to  $\sigma(A_{n-1})$ . When  $t$  is chosen complex, on the contrary, the  $\lambda_j(t)$  go off the real line and later bifurcate to converge to the real values  $\mu_i$ , as  $|t| \rightarrow \infty$ .

By varying the rank of  $E$ , one can obtain any section of  $A$  as  $\Pi = PAP_{\uparrow \text{Ker } E}$ , where  $P$  is the orthogonal eigenprojection on  $\text{Ker } E$ . It appears that Homotopic Deviation theory can be seen as a generalisation to a non hermitian context of the Cauchy interlace theorem.

## 9. Numerical illustrations

### 9.1. Example n° 1

We first illustrate the Homotopic Deviation  $H^+ = B(h) = B + hE$ , of order  $k + 1 = 9$ , see Section 6.  $H - I$  of order  $k = 8$  is taken to be Venice, the companion matrix defined in [10], p. 11: its characteristic polynomial is  $(x - 1)^3(x - 3)^4(x - 7)$ . Therefore the spectrum  $\sigma(H) = \{2, 4, 8\}$  has the structure  $(2^3)(4^4)(8^1)$ .

$E = e_9 e_8^T$  and  $B$  of order 9 is obtained by bordering  $H$  with  $a = 9$ ,  $u = (1, 2, 3, \dots, 8)^T$ . The spectrum of  $B$  is that of  $H$  plus 9. The projection  $P$  is on  $\text{lin}(e_1, \dots, e_7)$ , and

$$\Pi = PBP_{\upharpoonright \text{Ker } E} = H^- = \begin{pmatrix} 1 & & & & \mathbf{0} \\ 1 & \ddots & & & \\ & \ddots & \ddots & & \\ \mathbf{0} & & & 1 & 1 \end{pmatrix},$$

a transposed Jordan block of size 7.

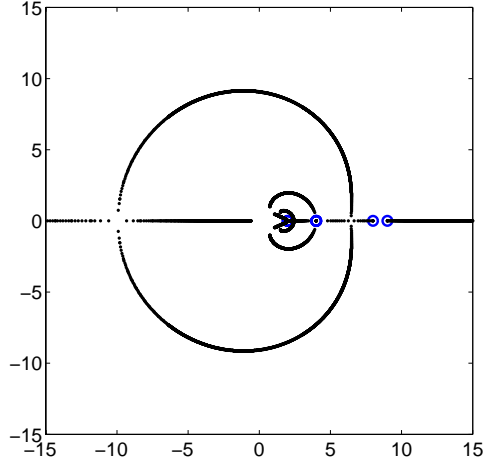
For  $h \in \mathbb{C}$ , the 9 maps  $h \rightarrow \lambda_i(h) \in \sigma(H^+)$  ( $i = 1, \dots, 9$ ) represent the spectral rays for the spectral field associated with  $(B, E)$ .

Figure 9.1 displays the 9 spectral rays (computed by QR) for  $h \in [0, 5 \times 10^3] \subset \mathbb{R}^+$ . The eigenvalues of  $B$  corresponding to  $h = 0$  are circled (bold/blue).

It is expected that exactly 2 spectral rays escape to infinity ( $\pm\infty$ ) because  $e_8^T u = 8 \neq 0$ . Indeed, this is the case with two eigenvalues diverging: the ray originating at 8 or 4 (resp. 9) escapes to  $-\infty$  (resp.  $+\infty$ ). The remaining 7 rays converge to  $\text{Lim} = \sigma(\Omega)$  with  $\Omega = H^- - \frac{1}{8} \tilde{u} e_7^T$ .

We observe that because one ray originating from 4 coalesces twice with the one originating from 8, for  $h > 0$ , there are two successive ambiguities in deciding which of the two rays tends to  $\infty$ .

Numerical experiments have been performed by Morad Ahmadnasab (Cerfacs) to examine computationally the nongeneric situation where  $u_8 = 0$  or  $u_8$  small. Preliminary results indicate that finite precision computation tends to reproduce the mathematical reality much more faithfully than we are used to. This unexpected phenomenon is thoroughly explored for possible confirmation.



**Figure 9.1** *Homotopic Deviation in Example n° 1*  
 $0 \leq h \leq 5 \times 10^3$ ,  $h \in \mathbb{R}^+$ ,  $e_8^T u \neq 0$ .

## 9.2. Example n° 2

In this example, we illustrate the phenomenon behind the Cauchy interlace theorem.

We take  $n = 7$ ,  $A$  is the symmetric tridiagonal Wilkinson matrix  $W_7$  defined by

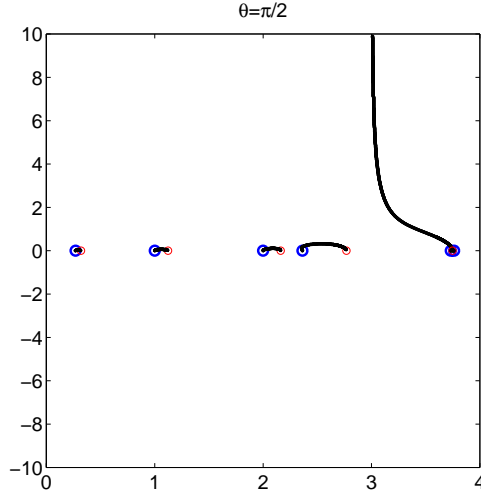
$$\begin{cases} w_{ii+1} = w_{i+1i} = 1, & i = 1 \text{ to } 6, \\ w_{ii} = 4 - i, & i = 1 \text{ to } 7. \end{cases}$$

Then we choose  $E = e_7 e_7^T$  and  $A(t) = A + tE$ .

There are  $6 = 7 - 1$  kernel points which are the eigenvalues of the section  $\Pi$  of order 6:  $\Pi = P W_7 P_{\uparrow \text{Ker } E}$  where  $P$  is the orthogonal projection on  $\text{lin}(e_1, \dots, e_6) = \text{Ker } E$ .

As  $|t| \rightarrow \infty$ , we expect that 6 eigenvalues  $\lambda(t)$  of  $A(t)$  will converge to  $\sigma(\Pi)$ . Because  $r = 1$ , we know that the kernel points  $\xi$  in  $\sigma(\Pi)$  are also critical ( $F_\xi$  is nilpotent,  $F_\xi^2 = 0$ ).

In order to make the convergence phenomenon more visible, we choose  $t$  pure imaginary:  $t = i|t|$ . Figure 9.2 displays the 7 spectral rays for  $0 \leq |t| \leq 10$ . The 7 (resp. 6) eigenvalues of  $A = W_7$  (resp.  $\Pi$ ) are circled by a bold/blue (resp. light/red) circle. It appears that the 3 (resp. 2) rightmost eigenvalues of  $A$  (resp.  $\Pi$ ) are very close. The family of Wilkinson matrices is notorious for such a spectral behaviour: this is the reason for the particular



**Figure 9.2** *Homotopic Deviation in Example n° 2*  
 $t = i|t|, 0 \leq |t| \leq 10$

choice  $A = W_7$ .

The spectral ray which originates in the rightmost eigenvalue of  $A$  escapes to infinity in the direction of the imaginary axis.

The remaining 6 spectral rays converge to the 6 eigenvalues  $\xi_i$  of  $\Pi$ , as is more visible on the zoom displayed by Figure 9.3.

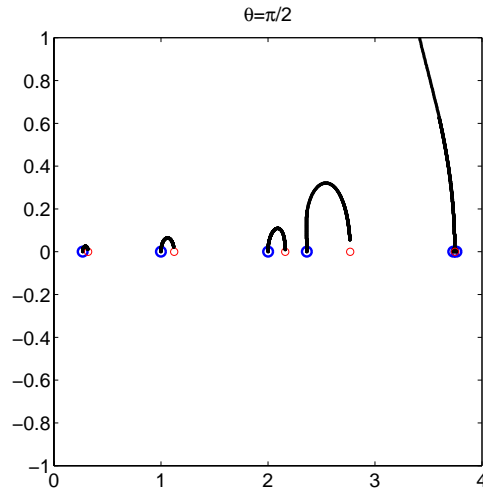
As a consequence of criticality at  $\xi \in \sigma(\Pi)$ ,

$$(A + tE - \xi I)^{-1} = (A - \xi I)^{-1} (I + tF_\xi)$$

exists for any  $t \in \mathbb{C}$  in *exact arithmetic*. This is well satisfied in finite precision at the 4 leftmost eigenvalues of  $\Pi$ . This is not the case, however, for the 2 rightmost ones. Such points are too close to eigenvalues of  $A$ .

## 10. Conclusion

Homotopic Deviation provides a theoretical framework to analyse the Arnoldi algorithm at step  $k$ , by expliciting the dependence on  $h$  of the spectra for the three successive imbedded Hessenberg matrices  $H^-$ ,  $H$  and  $H^+$  of respective order  $k - 1$ ,  $k$  and  $k + 1$ ,  $1 < k < n$ . The spectrum  $\sigma(H^+)$  can be related to  $\sigma(H)$  as  $h \rightarrow 0$ , and to  $\sigma(H^-)$  (or, more precisely, to  $\sigma(\Omega)$ ) as  $|h| \rightarrow \infty$ .



**Figure 9.3** Zoom for Figure 9.2

The first relation is well known: this gives the conventional approach to the convergence of Krylov methods.

The second relation is perhaps more surprising; it expresses analyticity in  $h$  around  $\infty$ , whereas the first relation expresses analyticity in  $h$  around 0. It allows to explain algebraically the remarkable robustness of Krylov methods to large perturbations. It helps explaining the nagging behaviour of the Arnoldi residual in finite precision. And it enables us to make a compelling case for restart based on algebra alone.

**Acknowledgements:** The author is grateful to Morad Ahmadnasab and Daniel Loghin (Cerfacs) for performing the Matlab computation described in Section 9. The author expresses her heartfelt thanks to her overskilled typist.

## References

- [1] F. Chatelin, *Valeurs propres de matrices*, Masson, Paris, 1988.
- [2] F. Chatelin, *Eigenvalues of matrices*, Wiley, Chichester, 1993, enlarged translation of [1].
- [3] F. Chaitin-Chatelin, *Comprendre les méthodes de Krylov en précision*

- finie: le programme du Groupe Qualitative Computing au CERFACS. Cerfacs Rep. TR/PA/00/11, Cerfacs, 2000.
- [4] F. Chaitin-Chatelin, About Singularities in Inexact Computing. Cerfacs Rep. TR/PA/02/106, Cerfacs, 2002.
  - [5] F. Chaitin-Chatelin, Computing beyond analyticity. Matrix Algorithms in Inexact and Uncertain Computing. Cerfacs Rep. TR/PA/03/110, Cerfacs, 2003.
  - [6] F. Chaitin-Chatelin, V. Frayssé, *Lectures on Finite Precision Computations*, SIAM Publications, Philadelphia, 1996.
  - [7] F. Chaitin-Chatelin, T. Meškauskas, Inner-outer iterations for mode solvers in structural mechanics: application to the Code-Aster. Contract Rep. FR/PA/01/85, Cerfacs, 2001.
  - [8] F. Chaitin-Chatelin, T. Meškauskas, M. van Gijzen, Inner-outer iterations for power method with Chebyshev acceleration in neutronics. Contract Rep. CR/PA/02/56, Cerfacs, 2002.
  - [9] F. Chaitin-Chatelin, V. Toumazou, E. Traviesas, “Accuracy assessment for eigencomputations: variety of backward errors and pseudospectra,” *Lin. Alg. Appl.* **309**, 73–83, 2000. Also available as Cerfacs Rep. TR/PA/99/03.
  - [10] F. Chaitin-Chatelin, E. Traviesas, Homotopic perturbation — Unfolding the field of singularities of a matrix by a complex parameter: a global geometric approach. Cerfacs Rep. TR/PA/01/84, Cerfacs, 2001.
  - [11] F. Chaitin-Chatelin, E. Traviesas, Qualitative Computing. Cerfacs Rep. TR/PA/02/58, Cerfacs, 2002, to appear as a chapter of *Handbook for Computation*, B. Einarsson ed., SIAM Publ., Philadelphia.
  - [12] F. Chaitin-Chatelin, E. Traviesas, L. Plantié, “Understanding Krylov methods in finite precision,” in *Numerical Analysis and its Applications*, NAA 2000 (L. Vulkov, J. Wasiewski, P. Yalamov eds.), Springer Verlag Lecture Notes in CS, vol. **1988**, pp. 187–197, 2000. Also available as Cerfacs Rep. TR/PA/00/40.

- [13] F. Chaitin-Chatelin, M. van Gijzen, Homotopic Deviation: an example in Acoustics. Cerfacs Rep. TR/PA/03/04, Cerfacs, 2003.
- [14] V. Frayssé, The power of backward error analysis. Habilitation, INP Toulouse 2000, Cerfacs Rep. TH/PA/00/65, Cerfacs, 2000.
- [15] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, 1980.
- [16] E. Traviesas, Sur le déploiement du champ spectral d'une matrice. Thèse d'Université, Toulouse 1, 2000, Cerfacs Report TH/PA/00/30, Cerfacs, 2000.
- [17] J. Moro, J. Burke, M. Overton, "On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure," *SIAM J. Matrix. Anal. Appl.* **18**, 793–817, 1997.

The Cerfacs reports are available from <http://aton.cerfacs.fr/algor/reports/>