The dynamics of matrix coupling with an application to Krylov methods

Françoise Chaitin-Chatelin *

CERFACS Technical Report TR/PA/04/29

Abstract

Given the matrices A and E in $\mathbb{C}^{n\times n}$, we consider, for the family $A(t) = A+tE,\ \ t \in \mathbb{C}$, questions such as i) existence and analyticity of $t \mapsto R(t,z) = (A(t)-zI))^{-1}$, and ii) limit as $|t| \to \infty$ of $\sigma(A(t))$, the spectrum of $A(t)$. The answer depends on the Jordan structure of $0 \in \sigma(E)$, more precisely on the existence of trivial Jordan blocks (of size 1). The results of the theory of Homotopic Deviation are then used to analyse the convergence of Krylov methods in finite precision.

Keywords : Sherman-Morrison formula, Jordan structure, frontier point, critical point, limit point, Ritz value, eigenprojection, analyticity, singularity, backward analysis, Krylov method.

1 Introduction

A and E are given matrices in $\mathbb{C}^{n\times n}$, which are coupled by the complex parameter t to form $A(t) = A + tE$. $\sigma(A)$ (resp. $re(A) = \mathbb{C} - \sigma(A)$) denotes the spectrum (resp. resolvent set) of A. We study the two maps:

$$t \in \mathbb{C} \mapsto R(t,z) = (A(t) - zI)^{-1},$$

for z given in $re(A)$, and

$$t \in \mathbb{C} \mapsto \sigma(A(t)).$$

*Université Toulouse 1 and CERFACS,42 avenue G. Coriolis 31057 Toulouse Cedex 1, France. E-mail: chatelin@cerfacs.fr

Such a framework is useful to perform a **backward analysis** for computational methods which are **inexact**: one has access to properties of $A(t)$ by means of the resolvent matrix $R(0,z) = (A - zI)^{-1},\ z \in re(A)$, only. In this context, the question of the behavior of $R(t,z)$ and $\sigma(A(t))$ as $|t| \to \infty$ arises naturally [6]. Such a study is also of interest for engineering when the parameter t has a physical meaning and can be naturally unbounded [10].

Various approaches are useful, ranging from analytic/algebraic spectral theory [1, 2, 3, 6, 10] to linear control system theory [12]. The theory surveyed here is **Homotopic Deviation** [4, 5, 11] which specifically looks beyond analyticity for $|t|$ large. The case of interest corresponds to a singular matrix E. The tools are elementary linear algebra based on the Sherman-Morrison formula and on the Jordan structure of $0 \in \sigma(E)$, as well as the more advanced Lidskii's perturbation theory [17].

1.1 Presentation of the paper

The paper is organized as follows. The mathematical setting is given in the rest of Section 1. Then Section 2 analyses the convergence rates for the two analytic developments for $R(t,z)$ around 0 and ∞. A similar analysis for $\sigma(A(t))$ is performed in Section 3. This results in a complete homotopic **backward analysis** for the eigenproblem for A, in terms of $t \in \mathbb{C}$, the homotopy parameter. The theory is used in Section 4 to explain the extreme robustness of inexact Krylov methods to very large perturbations [5, 15].

1.2 Notation

We set

$$F_z = -E(A - zI)^{-1},\ z \in re(A)$$

Formally

$$R(t,z) = R(0,z)(I - tF_z)^{-1}.$$

exists for $t \neq \dfrac{1}{\mu_z},\ 0 \neq \mu_z \in \sigma(F_z)$ and is computable as

$$R(t,z) = R(0,z)\sum_{k=0}^{\infty}(tF_z)^k \text{ for } |t| < \frac{1}{\rho(F_z)},\ \rho(F_z) = \max|\mu_z|.$$

When $\operatorname{rank} E = n,\ 0 \notin \sigma(F_z)$, and the eigenvalues of F_z are denoted by $\mu_{iz},\ i = 1, \cdots, n$. Therefore $R(t,z)$ is defined for almost all $t \in \mathbb{C},\ t \neq t_i$, with $t_i = \dfrac{1}{\mu_{iz}},\ i = 1, \cdots, n$. Consequently z is an eigenvalue of the n matrices $A(t_i),\ i = 1, \cdots, n$. What happens in the limit $|t| \to \infty$?

We set $s = 1/t,\ t \neq 0$.

$$I - tF_z = (sF_z^{-1} - I)tF_z,$$

and

$$(I - tF_z)^{-1} = -sF_z^{-1}(I - sF_z^{-1})^{-1} \to 0 \text{ as } s \to 0$$

Therefore

$$\lim_{|t|\to\infty} R(t, z) = 0$$

Similarly

$$A(t) = A + tE = t(sA + E) = \frac{1}{s}(E + sA).$$

An eigenvalue $\lambda(t)$ of $A(t)$ is such that

$$\lambda(t) = \frac{\nu(s)}{s} \text{ with } \nu(s) \in \sigma(E + sA).$$

Clearly, by continuity,

$$\nu(s) \to \nu \in \sigma(E) \text{ as } s \to 0,$$

and $\nu \neq 0$ implies $|\lambda(t)| \to \infty$. Therefore, when E is *regular* and $|t| \to \infty$, the limit of the resolvent matrix $R(t, z)$ (resp. the spectrum $\sigma(A(t))$ is 0 (resp. at ∞). To get a richer situation where the limit resolvent may be nonzero, and eigenvalues may stay at finite distance, we assume that $E \neq 0$ is *singular*, or *rank deficient*, $r = \text{rank } E,\ 1 \leq r < n$. We set $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$; card $\hat{\mathbb{C}} =$ card $\mathbb{C} = c$ denotes the cardinal of the (complex) continuum.

1.3 $E = UV^H$ with $U,\ V \in \mathbb{C}^{n\times r}$ of rank r $1 \leq r < n,\ E \neq 0$

Any singular matrix $E \neq 0$ of rank r can be written under the form

$$E = UV^H, \text{ with } U, V \in \mathbb{C}^{n\times r} \text{ of rank } r,\ 1 \leq r < n,$$

where $U,\ V$ of rank r represent a basis for Im E, Im E^H respectively [12]. F_z has now rank r, so that at most r eigenvalues $\mu_{iz},\ i = 1, \cdots, r$ are nonzero. They are the r eigenvalues of

$$M_z = -V^H(A - zI)^{-1}U \in \mathbb{C}^{r\times r},\ z \in re(A).$$

By applying the *Sherman-Morrison formula* [12] we have that

$$R(t, z) = R(0, z)[I_n - tU(I_r - tM_z)^{-1}V^H R(0, z)] \tag{1}$$

exists for $t \neq \frac{1}{\mu_z}$, $0 \neq \mu_z \in \sigma(M_z)$. For $z \in re(A)$, $R(t,z)$ is not defined when $t \in \mathbb{C}$ satisfies $t\mu_z = 1$, $0 \neq \mu_z \in \sigma(M_z)$. If M_z is regular, this is equivalent to $t \in \sigma(M_z^{-1})$.

Therefore $z \in re(A)$ is an eigenvalue of $A + tE$ iff $t\mu_z = 1$. This means that any z in $re(A)$ is an inexact eigenvalue for A at homotopic distance $|t|$, that is z is an exact eigenvalue of the r matrices $A(t_i) = A + t_i E$ with $t_i = \frac{1}{\mu_{iz}} \in \mathbb{C}$, $i = 1, \cdots, r$, when M_z is of rank r.

When $r > 1$, the homotopic distance is not uniquely defined.

The matrix M_z of order $r < n$ will play the key role in the analysis of our problem, similar to the role of the transfer matrix in linear control theory [12].

1.4 The limit of $R(t,z)$ when $|t| \to \infty$, for $z \in re(A)$

We suppose that $|t| > 1/\min|\mu_z|$ for M_z of rank r.

Proposition 1.1 *For $1 \leq r < n$, z given in $re(A)$ such that rank $M_z = r$, $\lim_{|t|\to\infty} R(t,z)$ exists and is given by*

$$R(\infty, z) = R(0,z)[I_n + UM_z^{-1}V^H R(0,z)]$$

Proof. By assumption, M_z^{-1} exists. $I_r - tM_z = (sM_z^{-1} - I_r)tM_z$,

$$(I_r - tM_z)^{-1} = -sM_z^{-1}(I_r - sM_z^{-1})^{-1},$$

$$-tU(I_r - tM_z)^{-1} = UM_z^{-1}(I_r - sM_z^{-1})^{-1} \to UM_z^{-1}.$$

The rest follows from(1). $P_{r_z} = I_n + UM_z^{-1}V^H R(0,z)$ is the eigenprojection for $F_z = -UV^H R(0,z)$ associated with the semi simple eigenvalue $0 \in \sigma(F_z)$ of multiplicitiy $n - r$. $\square$

When M_z^{-1} exists, the asymptotic resolvent $R(\infty, z)$ exists and is computable in *closed form* as $R(0,z)P_{r_z}$. This shows the dual role played by the two quantities $|t_1| = 1/\max|\mu_z| = 1/\rho(M_z)$ and $|t_r| = 1/\min|\mu_z| = \rho(M_z^{-1})$.

1) $|t_1|$ defines the largest analyticity disk for $R(t,z)$: it rules the convergence of the initial analytic development

$$R(t,z) = R(0,z)[I_n - tU\sum_{k=0}^{\infty}(tM_z)^k V^H R(0,z)] \tag{2}$$

based on M_z and valid for $|t| < |t_1|$ (around 0).

The series expansion(2) becomes *finite* when M_z is nilpotent ($\rho(M_z) = 0$).

2) $|t_r|$ defines the smallest value for $|t|$ beyond which $R(t,z)$ is analytic in $s = 1/t$: it rules the convergence of the asymptotic analytic development:

$$\begin{aligned} R(t,z) &= R(0,z)[I_n + UM_z^{-1}\textstyle\sum_{k=0}^{\infty}(sM_z^{-1})^k V^H R(0,z)] \\ &= R(\infty,z) + R(0,z)UM_z^{-1}\textstyle\sum_{k=1}^{\infty}(tM_z)^{-k}V^H R(0,z), \end{aligned} \tag{3}$$

based on M_z^{-1} and valid for $|t| > |t_r|,\ s = 1/t$, (around ∞).

Observe that M_z^{-1} cannot be nilpotent (because it is invertible).

1.5 Frontier of existence for $R(\infty, z) = \lim\limits_{|t|\to\infty} R(t,z)$ $z \in re(A)$

In general, $\lim\limits_{z\to\lambda} \rho(M_z) = \infty$ for $\lambda \in \sigma(A)$. If $\lambda \in \sigma(A)$ is such that $\lim\limits_{z\to\lambda} M_z = M_\lambda$ is defined (hence $\rho(M_\lambda) < \infty$, see [3]) we say that λ is *normwise-unobservable* by the deviation process (A, E) [11]. An eigenvalue λ such that $\lim\limits_{z\to\lambda} \sigma(M_z) = \sigma_\lambda$ exists, in particular $\rho(M_z) \to \rho_\lambda < \infty$ is *spectrally unobservable* [11], in short σ-unobservable.

Definition 1.1 *The frontier points form the set $F(A,E) = \{z \in re(A);\ 0 \in \sigma(M_z)\}$ of points in $re(A)$ for which $R(\infty, z)$ does not exist. The critical points form the set $\mathcal{C}(A,E)$ of frontier points such that $\rho(M_z) = 0$.*

The inclusion $\mathcal{C}(A,E) \subset F(A,E))$ becomes an equality when $r = 1$. In general, if M_z is not rank defective for all z in $re(A)$, $F(A,E)$ is a finite set of isolated points in $re(A) \subset \mathbb{C}$. We shall see below that when $0 \in \sigma(E)$ is semi-simple, then card $F(A,E) \leq (n-1)r$. [11].

An exceptional case when card $F(A,E) = c$ or 0 is provided by the particular matrix $A = \lambda I$, which entails $M_z = \dfrac{1}{z-\lambda}V^H U$. Clearly, M_z is regular (resp. singular) for $z \neq \lambda$ when 0 is semi simple (resp. defective).

Similarly, it will be shown that $\mathcal{C}(A,E)$ is a finite set of at most $n-1$ points, unless the map $t \mapsto \sigma(A(t))$ is constant for $t \in \mathbb{C}$, and $\mathcal{C}(A,E) = F(A,E) = re(A)$. This situation requires E to be nilpotent [11].

2 Convergence rates for the two analytic developments for $R(t,z)$ as functions of $z \in re(A)$.

As z varies in $re(A)$, the convergence rate for (2) (resp. (3) is described by the map : $\varphi_1 : z \mapsto \rho(M_z)$ (resp. $\varphi_2 : z \mapsto \rho(M_z^{-1})$).

2.1 The spectral portrait φ_1

The map φ_1 is the homotopic analogue of the popular normwise spectral portrait map : $z \mapsto ||(A - zI)^{-1}||$, [6]. In φ_1, the matrix $(A - zI)^{-1}$ of order n is replaced by M_z of order $r < n$, and $|| \cdot ||$ by $\rho(\cdot)$.

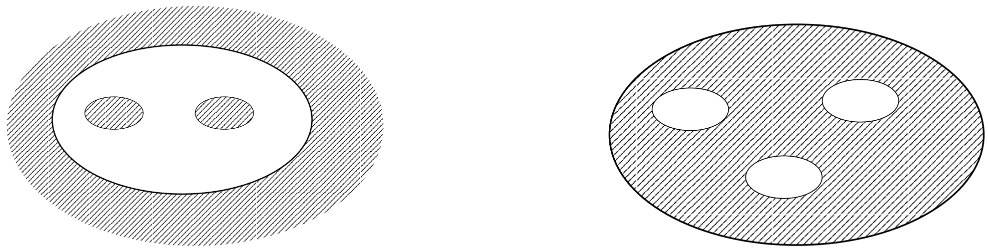
An important consequence is that φ_1 can localize the critical points ($\rho = 0$) when they are isolated, whereas the normwise spectral portrait cannot, see specifically the paragraph 2.3.

The map $\varphi_1 : z \mapsto \rho(M_z)$ is subharmonic with singularities at the σ-observable eigenvalues of A ($\rho = \infty$) and the critical points ($\rho = 0$). We assume that there exist σ-observable eigenvalues. Subharmonicity in $\mathbb{C}$ is the 2D-analogue of monotonicity in $\mathbb{R}$. It allows to order the $\varepsilon-$level sets, $\varepsilon > 0$ by inclusion. As z varies outside the disk $\{z;\ |z| \leq \rho(A)\}$, $\rho(M_z)$ decreases from $+\infty$ to 0 ($\rho(M_z) \to 0$ as $|z| \to \infty$). Therefore the set $\Gamma_0^\alpha = \{z \in \mathbb{C};\ \rho(M_z) = \alpha\}$ consists of a finite number of closed curves. For α small enough, there exists one single exterior curve around all the others which enclose local minima or isolated critical points.

The associated domain of convergence for(2) is the unbounded region outside the outer curve and inside the inner curves. See Figure 1, a) on the left. See also [7, 8].

2.2 The frontier portrait φ_2

The map $\varphi_2 : z \mapsto \rho(M_z^{-1}) = \rho_2$ is also subharmonic with singularities ($\rho = \infty$) at points in $F(A,E)$. We assume that $A \neq \lambda I$, and that $F(A,E)$ is a non empty finite set. When $|z|$ increases away from $F(A,E)$, $\rho(M_z^{-1})$ decreases to a local minimum to increase again ($\rho(M_z^{-1}) \to \infty$ as $|z| \to \infty$). For $\beta \geq \beta_\star > 0$, the set $\Gamma_\infty^\beta = \{z \in \mathbb{C};\ \rho(M_z^{-1}) = \beta\}$ consists of a finite number of closed curves. And for β large enough, there exists one single exterior curve around the others which enclose the points in $F(A,E)$. We observe that in exact arithmetic, it is conceivable that $\rho(M_z^{-1})$ can be 0 at σ-observable eigenvalues



a) Γ_0^α for (2) b) Γ_∞^β for (3)

Figure 1: Analytic representations for $R(t,z)$, $\alpha \le \beta$

of A, for which M_z is not defined, hence $\mu_{min} = \infty(\frac{1}{\mu_{min}} = 0)$ where μ_{min} is an eigenvalue for M_z of minimal modulus.

The associated domain of convergence for (3) is the bounded region inside the outer curve and outside the inner ones. See Figure 1, b) on the right and [10]. The shaded areas represent the respective analyticity domains for $R(t,z)$ around 0 ($|t| < \frac{1}{\alpha}$) and ∞ ($|t| > \beta$), with α small or β large, $\alpha \le \beta$.

2.3 The critical points

When they exist, the critical points in $\mathcal{C}(A,E) \subset F(A,E)$ are singularities for φ_1 (at 0) and for φ_2 (at ∞).

At an isolated critical point, there is an abrupt change in the representation of $R(t,z)$. The symmetry of the dual analytic representation, valid locally for $|t|$ small (around 0) or large (around ∞) is broken in favour of 0.

The finite representation:

$$R(t,z) = R(0,z)[I_n - tU\sum_{k=0}^{r-1}(tM_z)^k V^H R(0,z)] \tag{4}$$

as a polynomial in t of degree $\le r$, is valid for t everywhere in $\mathbb{C}$. The limit as $|t| \to \infty$ is not defined.

If M_z is nilpotent for any z in $re(A)$, $\sigma(A)$ is unobservable but $R(0,z)$ is *not* defined for $z \in \sigma(A)$.

2.4 The case $r = 1$

The matrix M_z of order r reduces to the scalar μ_z. And $\mu_z\mu_z^{-1} = 1$. Therefore $\mathcal{C}(A,E) = F(A,E)$, and we can choose $\alpha = \beta = 1$. The unique set $\Gamma_0^1 = \Gamma_\infty^1$ reduces to the set Γ studied in [7].

There are at most $n-1$ critical points [4, 11] unless $\sigma(A(t))$ is invariant under $t \in \mathbb{C}$. In this case $\mathcal{C}(A, E) = re(A)$ and can be extended to $\mathbb{C}$ by continuity of $z \mapsto \rho(M_z) = 0$.

The symmetry between 0 and ∞ expressed by $s = 1/t$ is also carried by $\rho(M_z^{-1}) = 1/\rho(M_z)$. Convergence at 0 (resp. ∞) for (2) is equivalent to divergence at 0 (resp. ∞) for (3) for any z not critical ($\rho(M_z) > 0$). Such an exact symmetry does not hold for $r > 1$ since any z in $re(A)$, which is not a frontier point, is simultaneously an eigenvalue for r matrices $A(t)$, instead of just one. We shall continue this analysis in Section 3, after the comparison of the normwise versus homotopic level sets to follow.

2.5 Normwise versus homotopic level sets for $||\cdot||$, φ_1, φ_2.

A classical normwise backward analysis yields the well-known identity for $\varepsilon > 0$:

$$R_\varepsilon^N = \{z \in re(A);\ ||(A-zI)^{-1}|| \geq \frac{1}{\varepsilon}\} = \{z \in \sigma(A+E) \cap re(A),\ ||E|| \leq \varepsilon\} = S_\varepsilon^N,$$

where the sets cannot be empty [6]. N stands for normwise.

The homotopic analogue of R_ε^N is given by $R_\varepsilon = \{z \in re(A);\ \rho(M_z) \geq \frac{1}{\varepsilon}\}$ which can be empty for $\varepsilon > 0$ if all the eigenvalues of A are σ-unobservable by (A, E). Such a situation corresponds to $\rho(M_z) = 0$ for any $z \in re(A)$.

The analogue of S_ε^N consists of the z in $re(A)$ which are eigenvalues of $A + tE$, at distance $|t| \leq \varepsilon$. Because there can be r such matrices for any given z in $re(A)$, the homotopic distance is not uniquely defined.

For example, one can choose a distance which is a) minimal or b) maximal. This corresponds to :

a) $|t| = \dfrac{1}{|\mu_{max}|}$: $A(t)$ is the *closest* matrix having z as its eigenvalue. Then $S_\varepsilon^a = R_\varepsilon$ [9]. This is the only possibility when $r = 1$.

b) $|t| = \dfrac{1}{|\mu_{min}|}$: $A(t)$ is the *farthest* matrix, then $S_\varepsilon^b \subset R_\varepsilon$. The maximal distance induces the level set for $\varphi_2 : \rho(M_z^{-1}) \leq \varepsilon$ [10].

3 The spectrum $\sigma(A(t))$ as $|t| \to \infty$

Because E is singular, it is possible that some eigenvalues $\lambda(t)$ of $A(t)$ remain at finite distance when $|t| \to \infty$ [4].

Observing the evolution $t \mapsto \lambda(t)$ as $t \in \mathbb{C}$ leads to the distinction between invariant and evolving eigenvalues for A, according to the :

Definition 3.1 *$\lambda \in \sigma(A)$ is an evolving (resp. invariant) eigenvalue iff $\lambda(t) \neq \lambda$ for almost all $t \neq 0$ (resp. $\lambda(t) = \lambda$ for all $t \in \mathbb{C}$). We write $\sigma(A) = \sigma^i \cup \sigma^e$ where σ^i (resp. σ^e) consists of invariant (resp. evolving) eigenvalues.*

Note that, in case of a multiple λ, one copy of λ may be invariant while another is evolving [10].

3.1 Notation

$0 \in \sigma(E)$ has algebraic (resp. geometric) multiplicity m (resp. $g = n - r$). The general case is $g < m \leq n$ (0 defective). There are g', $0 \leq g' < g$, trivial Jordan blocks of size 1 associated with 0. The corresponding eigenvectors span $K' \subset \text{Ker } E$ when $g' \geq 1$; $M = \text{Ker } E^m$ is the invariant subspace for 0. Let P (resp. P') be the spectral (resp. eigen)projection on M (resp. K'). Π (resp. Π') of order m (resp. g') represents the Galerkin approximation PAP (resp. $P'AP'$) restricted to M (resp. K'). The spectrum $\sigma(\Pi)$ (resp. $\sigma(\Pi')$) consists of the associated Ritz values. If 0 is semi-simple, $g' = g = m = n - r < n$, $K' = \text{Ker } E = M$, $P' = P$ and $\Pi' = \Pi$.

The Galerkin approximation $P'AP'$ and its restriction Π' to K' will play an essential role for the analysis of $\lim \sigma(A(t))$ as $|t| \to \infty$.

We define in $\hat{\mathbb{C}}$ the set $\sigma_\infty(A, E) = \lim_{|t|\to\infty} \sigma(A(t)) = \{\infty, \text{ Lim }\}$, which represents the possible limits for $\lambda(t) \in \sigma(A(t))$ as $|t| \to \infty$. Either $|\lambda(t)| \to \infty$, or $\lambda(t) \to z \in \text{Lim } \subset \mathbb{C}$.

We set $l_\star = \text{card Lim }$, $0 \leq l_\star \leq n$, where the points in Lim are counted according to their algebraic multiplicity as eigenvalues of $A(t)$, $|t|$ large.

It is clear that all invariant eigenvalues in σ^i belong to Lim .

3.2 Backward analysis for the eigenproblem on A

For any given $z \in \mathbb{C}$, we investigate the various ways in which z can bee seen as an exact eigenvalue for $A + tE$, $t \in \hat{\mathbb{C}}$. Therefore, we introduce the

Definition 3.2 *The set $\mathcal{N}_z = \{0 \neq t \in \hat{\mathbb{C}},\ z \in \sigma(A + tE)\}$ is the nodal set for $z \in \mathbb{C}$.*

We define $k_z = \text{card } \mathcal{N}_z$. We distinguish whether k_z is finite or $k_z = c$. When k_z is finite, each $t_i = \dfrac{1}{\mu_{iz}}$ in $\mathcal{N}_z$ is counted according to the algebraic multiplicity of μ_{iz} in $\sigma(M_z)$.

Proposition 3.1 *When $z \in re(A)$, $k_z = r$. When $z = \lambda \in \sigma(A)$, $k_\lambda = c$ when λ is invariant. When λ is evolving, k_λ is finite.*

Proof. Clear by $t\mu_z = 1$. $\square$

Proposition 3.1 specifies the number k_z of ways by which any z in $\mathbb{C}$ can be seen as an *inexact* eigenvalue for A (that is, an exact eigenvalue for $A+tE$, with $0 \neq t \in \mathbb{C}$ or $|t| = \infty$). Such a number is finite when z is not an invariant eigenvalue $\lambda = z$ in σ^i. When this is the case, the backward analysis delivers an ambiguous answer : $k_\lambda = c$: λ is an exact eigenvalue for $A + tE$ for any $t \in \mathbb{C}$.

3.3 Properties of Lim

We suppose first that $0 \in \sigma(E)$ is defective : $0 \leq g' < g = n - r < m \leq n$.

Lim can be partitioned into the invariant spectrum σ^i and $\mathrm{Lim}^{\,e} = \{z \in \mathbb{C},\ z = \lim_{|t|\to\infty} \lambda(t)$ with $\lambda(t) \neq \lambda(0) = \lambda$ for almost all $t\}$, which consists of the limits in $\mathbb{C}$ of evolving eigenvalues originating from σ^e. Clearly $\sigma^i \cap \mathrm{Lim}^e$ need not be empty.

Lemma 3.2 *If there exists an eigenvector u for A associated with λ such that $u \in Ker\, E$, then λ is invariant in σ^i.*

Proof. $(A + tE)u = \lambda u$ for any $t \in \mathbb{C}$, since u is an eigenvector for A such that $Eu = 0$. Observe that the lemma provides a sufficient condition only for $\lambda \in \sigma^i$ [11]. When λ is multiple, it is possible that another copy is evolving. $\square$

We follow the study of Lim presented in [11]. It relies on the relation $A + tE = t(E + sA) = \frac{1}{s}(E + sA)$ for $s = 1/t$, and on the spectral properties of $\frac{1}{s}(E + sA)$ when $s \to 0$, analyzed by Lidskii's theory [11]. The reader is refered to [11], Section 4, for the notations used below related to $E = XJX^{-1}$: $\tilde{X} = [Z, X'],\ \tilde{Y} = [W, Y'], \tilde{\Pi} = \begin{pmatrix} \Gamma & R \\ L & \Pi' \end{pmatrix} = \tilde{Y}^T B \tilde{X}$, with $B = X^{-1}AX$.

Under the assumption (G) that Γ has rank f, the matrix $\Omega = \Pi' - L\Gamma^{-1}R$ is the *Schur complement* of Γ in $\tilde{\Pi}$. The stronger assumption $(\hat{G})$ on Γ is defined in [11].

Theorem 3.3 *i) When (G) holds with $g' \geq 1$, then Lim $\supset \sigma(\Omega)$*
ii) When (G) is replaced by $(\hat{G})$, then Lim $= \sigma(\Omega)$.

Proof. Point i) is Proposition 4.2 in [11]. For Point ii) the reader is refered to [11], and to [17], theorem 2.1. $\square$

We observe that $(\hat{G})$ reduces to (G) when the non trivial Jordan blocks are of the same size. We shall use this observation in Section 4.

In general, one has the

Proposition 3.4 *When the critical set is discrete, then*

$$\mathcal{C}(A,E) \subset Lim \cap re(A) \subset F(A,E)$$

with equalities when $r = 1$.

When the critical set is continuous, $F(A,E) = \mathcal{C}(A,E)$ *is the continuous set* $re(A)$*, and* $Lim = \sigma(A)$.

Proof. See Theorem 5.5 in [11]. □

Corollary 3.5 *When* 0 *is defective and* $(\hat{G})$ *holds with* $g' \geq 1$*, the critical set* $\mathcal{C}(A,E)$ *is either discrete in* $re(A)$ *with at most* $g' \leq n-r-1$ *points, or it is continuous.*

Proof. Clear from $g' < g = n - r$, and Theorem 3.3. □

An immediate consequence is that when $g' = 0$ (no trivial Jordan blocks) the three sets $\sigma(\Pi')$, Lim and $\mathcal{C}(A,E)$ are *empty* under $(\hat{G})$.

The situation simplifies when $0 \in \sigma(E)$ is semi-simple : first the conditions (G) and $(\hat{G})$ vanish; second, the critical and frontier sets are always discrete with respectively at most $n-r$ and $(n-1)r$ points. Lim contains exactly $n-r$ points which are the Ritz values in $\sigma(\Pi)$. Lim can never contain n points : r eigenvalues necessarily escape to ∞.

Proposition 3.6 *If* $0 \in \sigma(E)$ *is semi-simple, then* $Lim = \sigma(\Pi)$ *and* $\mathcal{C}(A,E) \subset Lim \cap re(A) \subset F(A,E)$*, with* $g' = g = n - r = m = l_\star < n$.

See [4, 10, 11]. A numerical example in Computational Acoustics is treated in [10], where $s = \zeta$ is the complex impedance, and $t = 1/\zeta$ is the admittance. The boundary condition for the acoustic wave is Neumann (resp. Dirichlet) for $\zeta = \infty$ (resp 0).

3.4 Convergence of the eigenvectors

By Theorem 3.3, the assumption $(\hat{G})$ guarantees that exactly g' eigenvalues tend to $\sigma(\Omega)$, the remaining $n-g'$ ones diverging to ∞. If we assume moreover that Ω has distinct **simple** eigenvalues, then for s small, $E(s) = E + sA$ has exactly g' simple eigenvalues. The associated eigenvectors are the eigenvectors for $A(t)$ associated with the converging $\lambda(t)$. They converge in $O(s)$ to g' vectors specified in the

Theorem 3.7 *Under the three assumptions $g' \geq 1$, $(\hat{G})$ and Ω has simple eigenvalues, exactly g' simple eigenvalues $\lambda(t)$ for $A+tE$ converge to a limit point $\xi \in \sigma(\Omega)$ as $|t| \to \infty$.*

The corresponding eigenvectors $x(t)$ converge in $O(1/t)$ to $\varphi \in Ker\, E$ with $\varphi = (I-\Sigma)X'b$, where $\Omega b = \xi b$, and $\Sigma = Z\Gamma^{-1}W^T B$.

Proof. This is a particular case ($j = q$) of theorem 2.2 in [17], which proves that $\varphi = [Z, X'] \begin{pmatrix} c \\ b \end{pmatrix}$ where $(c\ b)^T$ is a nonzero solution of

$$\begin{pmatrix} \Gamma & R \\ L & \Pi' - \xi I_{g'} \end{pmatrix} \begin{pmatrix} c \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This system is equivalent to

$$\begin{pmatrix} \Gamma & 0 \\ 0 & \Omega - \xi I_{g'} \end{pmatrix} \begin{pmatrix} I_f & \Gamma^{-1}R \\ 0 & I_{g'} \end{pmatrix} \begin{pmatrix} c \\ b \end{pmatrix} = 0,$$

that is

$$\begin{cases} (\Omega - \xi I_{g'})b = 0, & b \neq 0 \\ c + \Gamma^{-1}Rb = 0 \end{cases}$$

because Γ has rank f. This yields

$$\begin{aligned} \varphi = [Z, X'] \begin{pmatrix} -\Gamma^{-1}Rb \\ b \end{pmatrix} &= X'b - Z\Gamma^{-1}W^T BX'b \\ &= (I-\Sigma)X'b \in \text{Ker } E. \end{aligned}$$

We observe that $P'\varphi = X'b \in K'$ since $Y'^T Z = 0$. But φ does not belong to K', unless R or $L = 0$, hence $c = 0$. $\square$

The matrix Σ is a projection with rank 1 [11].

We recall [11] that $I - \Sigma$ expresses the complexity introduced by the presence of non trivial Jordan blocks. Indeed, $\Sigma = 0$ when $0 \in \sigma(E)$ is semi-simple.

We know that $\Omega - \xi I_{g'}$, of order g', is singular iff M_ξ, of order r, is also singular. Under (Σ), there is a 1 to 1 correspondence between $a \in \mathbb{C}^r$, eigenvector for M_ξ and $b \in \mathbb{C}^g$, eigenvector for $\Omega = \Pi$ at $\xi (g' = g)$.

The vector $v = Ua \in \text{Im } E$ is such that $v = (B - \xi I)u$ is the residual for (ξ, u), $u \in \text{Ker } E$, an eigenpair for Π. From this follows the preservation of *geometric* multiplicities : dim Ker $(\Pi - \xi I) =$ dim Ker M_ξ.

What is the situation when 0 is defective? The eigenvector a for M_ξ defines $w = Ua \in \text{Im } E$, such that $w = (B - \xi I)u$ with $u \in \text{Ker } E$. The eigenvector

b for Ω defines $\varphi \in \text{Ker } E$ and the residual $v = (B - \xi I)\varphi$ such that $P'v = 0$. (ξ, u) with $u = X'b$ is an eigenpair for $P'B(I - \Sigma)P'_{\restriction K'}$.

The vector v belongs to Ker P' of dimension $n - g'$, whereas w belongs to Im E of dimension $r = n - g < n - g'$. The identification $v = w$ is not possible, unless v has no component on T, the subspace of dimension f spanned by the invariant vectors *ending* the f Jordan chains of dimension > 1.

This happens to be true, according to the

Lemma 3.8 *The residual vector $v = (B - \xi I)\varphi$ has no component in T.*

Proof. Because $\varphi \in \text{Ker } E = K' \oplus S$, it suffices to prove that $W^T B\varphi = 0$, where W is a basis for T. $W^T B\varphi = W^T B(I - \Sigma)X'b = Rb - (\Gamma\Gamma^{-1})Rb = 0$. Therefore $v \in \text{Im } E$. □

Corollary 3.9 *When $F(A, E)$ is finite the equality*

$1 = $ *dim Ker* $(\Omega - \xi I) =$ *dim Ker* M_ξ *holds for* $\xi \in re(A)$ *under the assumptions of Theorem 3.7.*

Proof. Clear from Lemma 3.8. There is a 1 to 1 correspondence between b and a through $v = w$. □

When the eigenvalues of Ω are all *simple*, and when $F(A, E)$ is finite, we get back the preservation of geometric multiplicities which is the rule under (Σ). In the general case, when dealing with the convergence of *eigenvectors*, it seems difficult to bypass the first assumption(ξ simple), which is required to make use of the implicit function theorem [17], p.803. The convergence of *eigenvalues* is less demanding. We know that $(\hat{G})$ can be weakened into (G) to get $\sigma(\Omega) \subset \text{Lim}$ [11].

3.5 Limits of the backward analysis

The connection between z and t which holds when z is interpreted as an eigenvalue of $A + tE$ is expressed by $t\mu_z = 1$.

This relation is well defined for t and μ_z nonzero. The limits of the backward analysis correspond to $(t = 0,\ |\mu_z| = \infty)$ or $(|t| = \infty,\ \mu_z = 0)$.

1) λ is an exact eigenvalue for A : $t = 0$ requires that M_λ is not defined. This is the case for *observable* eigenvalues (μ_λ is not defined).

Normwise-unobservable eigenvalues (M_λ exists) are seen by the process as inexact eigenvalue at a positive distance $\geq \dfrac{1}{\rho(M_\lambda)}$, instead of at a distance exactly zero.

2) $z \in re(A)$ is a critical point such that $\rho(M_z) = 0$, therefore $|t| = \infty$ is the only possibility. An *isolated* critical point is an inexact eigenvalue at infinite distance, in agreement with the representation of $R(t, z)$ as a polynomial in t. Such a z is the limit of $\lambda(t)$ as $|t| \to \infty$.

However, when the set of critical points is $\mathbb{C}$, that is, when M_z is nilpotent for any z in $re(A)$, $\text{Lim} = \sigma(A)$ and only the eigenvalues themselves are (trivial) limits, not arbitrary critical points in $re(A)$.

4 Convergence of Krylov methods in finite precision

We approach this question by considering an iterative Krylov method as an inner-outer iteration.

The outer loop modifies the starting vector v_1 for the construction of the Krylov basis. The inner loop is a direct method which is an incomplete Arnoldi decomposition of size k, $k < n$ [13, 11]. The dynamics of this 2-level algorithm is studied by a homotopic deviation on the matrix of order $k+1$

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

such that $H_{k+1} = \left(\begin{array}{c|c} H_k & u \\ \hline 0\ h_{k+1\ k} & a \end{array}\right)$ is the computed Hessenberg form of order $k+1$. The homotopy parameter is $h = h_{k+1\ k}$, and the deviation matrix is $E = e_{k+1} e_k^T$: $B(h) = B + hE = H_{k+1}$. E is nilpotent ($E^2 = 0$) with rank 1, and $\sigma(E) = \{(0^1)^{k-1},\ (0^2)\}$. For k fixed, $1 < k < n$, we set $H^- = H_{k-1}$, $H = H_k$, $H^+ = H_{k+1}$: these are the three successive Hessenberg matrices constructed by the Arnoldi decomposition, of order $k-1$, k and $k+1$. And we define $u = (\tilde{u}^T, u_k)^T$, $h^- = h_{k\ k-1}$

We assume that $H_k = H$ is *irreducible*, therefore $\sigma(H^-) \cap \sigma(H) = \emptyset$ and $h_{k\ k-1} \neq 0$ in particular. $\sigma(B) = \sigma(H) \cup \{a\}$.

4.1 Theoretical consequences

With the notation of Section 3, $0 \in \sigma(E)$ has the multiplicities $g' = k - 1 < g = k < m = k + 1$. Therefore $g' \geq 1$ for $k \geq 2$. The eigenspace K' is $K' = \text{lin}\,(e_1, \cdots, e_{k-1})$, and P' is the orthogonal projection on K', $P = I_{k+1}$. Thus $\Pi' = H_{k-1} = H^-$, and $\Omega = H^- - \frac{h_{k\ k-1}}{u_k} \tilde{u} e_{k-1}^T$ for $u_k \neq 0$. The matrix M_z reduces to the scalar $\mu_z = -e_k^T (B - zI_{k+1})^{-1} e_{k+1}$, for $z \notin \sigma(B)$. Finally, because $r = 1$, $\mathcal{C}(B, E) = F(B, E)$ in $re(B)$. We survey the results established in [5].

1) About critical and limit points.

Theory tells us that $(\hat{G}) = (G)$ and the generic case corresponds to $u_k \neq 0$. Therefore Lim $= \sigma(\Omega)$, and Lim $\cap\, re(B) = \mathcal{C}(B, E)$ contains at most $k-1$ critical points in $re(B)$. Exactly two eigenvalues of H^+ escape to ∞ as $|h| \to \infty$.

2) Rational/linear representation of $(H^+ - zI_{k+1})^{-1}$ for $z \notin \sigma(H^+)$.

Given any z in $re(B)$, we consider the resolvent $R(z) = (B - zI_{k+1})^{-1}$. We define $\beta_z = (B - zI_{k+1})^{-1} e_{k+1}$ its last column, and $\alpha_z^T = e_k^T (B - zI_{k+1})^{-1}$ its k^{th} row.

Provided that $h\mu_z \neq 1$, z is not an eigenvalue for H^+. One has the following representation in h :

$$(H^+ - zI_{k+1})^{-1} = R(z) + \frac{h}{h\mu_z - 1} \beta_z \alpha_z^T, \ h\mu_z \neq 1.$$

The representation is rational in h for $\mu_z \neq 0$, and linear for $\mu_z = 0$ (z critical).

We consider now the equation :

$$(H^+ - zI_{k+1}) g = f,$$

and its solution $g = g(h, z) = (H^+ - zI_{k+1})^{-1} f$.

We set $g_0(z) = R(z) f$, $g_1(z) = (\alpha_z^T f) \beta_z$. It is clear that

$$g(h, z) = g_0(z) + \frac{h}{h\mu_z - 1} g_1(z), \text{ for } h\mu_z \neq 1.$$

What happens if $h\mu_z = 1$? z is an eigenvalue for H^+ with associated eigenvector β_z, colinear with $g_1(z)$.

3) Remarkable identities for α_z^T and β_z, $z \notin \sigma(H^+)$.

The last two components of these vectors have a simple explicit expression, respectively given by :

$\alpha_z^T e_k = \dfrac{\pi^-(z)}{\pi(z)}$, where $\pi(z) = \det(H - zI_k)$ and $\pi^-(z) = \det(H^- - zI_{k-1})$,

$\alpha_z^T e_{k+1} = e_k^T \beta_z = -\mu_z$, and $e_{k+1}^T \beta_z = \dfrac{1}{a - z}$

When z is critical, $z \in \sigma(\Omega) \cap \mathrm{re}(B)$ and $\mu_z = 0$. Therefore the rank 1 matrix $\beta_z \alpha_z^T$ has its k^{th} row, and its last column equal to 0. This has the following consequences on $g_1(z) = (\alpha_z^T f)\beta_z$, for z critical: the k^{th} component $e_k^T g_1(z) = 0$, and the scalar $\alpha_z^T f$ is independent of the last component of f

4) On the pseudo eigenpairs for H_l, $l \geq k$ deriving from an exact eigenpair for H^-.

Let (ξ, p) be an exact eigenpair for H^- : $H^- p = \xi p$ for $p \in \mathbb{C}^{k-1}$. We consider the augmented vector $\hat{\psi}_l = (p^T, 0)^T$ in $\mathbb{C}^l$, $l \geq k$, and define $h^- = h_{k\ k-1}$, $p_{k-1} = e_{k-1}^T p$.

The pair $(\xi, \hat{\psi}_l)$ is a pseudo eigenpair for H_l, $l \geq k$ corresponding to the residual vector $(h^- p_{k-1}) e_k$ in $\mathbb{C}^l$. The pair $(\xi, \hat{\psi}_l)$ *cannot be improved* by inverse iteration using the Hessenberg form H_l, for $l \geq k+1$ ($g_1(\xi) = 0$ for any f colinear with e_k). This explains why the true residual for ξ increases after the iteration $k+1$, when ξ has been computed at iteration $k-1$. See [13] for a numerical illustration. When this happens, the only solution is to *restart* with an improved starting vector v_1.

5) The four spectra $\sigma(H^-)$, $\sigma(H)$, $\sigma(H^+)$ and $\sigma(\Omega)$.

Classical "convergence" takes place when the three spectra $\sigma(H^-)$, $\sigma(H)$ and $\sigma(H^+)$ have a number of points close to each other. If, in addition, certain eigenvalues of Ω are nearby, this gives a reason why convergence may be better explained with $|h|$ large rather than small.

This happens if $\Omega = H^-$, that is $\tilde{u} = 0$. This is almost true when $||\Omega - H^-|| = |h^-| \dfrac{||\tilde{u}||}{|u_k|}$ is small. Observe that $\dfrac{||\tilde{u}||}{|u_k|} = \tan \psi$, where ψ is the acute angle between the directions spanned by $\tilde{u}$ and e_k. $u_k \neq 0$ iff $0 \leq \psi < \frac{\pi}{2}$.

4.2 Algorithmic consequences in finite precision

In exact arithmetic, the algorithmic analysis of the inner loop is easy under the assumption of irreducibility : either v_1 is an invariant vector for A and the algorithm stops exactly (with $h = 0$) for $k < n$, or v_1 is not invariant and the algorithm has to be run to completion ($k = n$).

In finite precision, the analysis is more delicate, since the mathematical analysis for convergence ($h \to 0$) is valid only when round-off can be ignored. And it is well known that round-off cannot be ignored when "convergence" takes place [13, 14, 16].

"Convergence" in finite precision means "near-reducibility", and this can happen with $|h|$ **large**, although this seems numerically counter-infinitive at first sight.

The algorithmic dynamics for "convergence" entails that there exist points in $\sigma(H^-)$, $\sigma(H)$ and $\sigma(H^+)$ which are very close, in spite of the fact that an exact coincidence is ruled out by the assumption of irreducibility for A.

The dynamics expressed in finite precision makes it possible that a value $z \in \sigma(H^+)$ which is close to $\sigma(\Omega)$ corresponds to a large h : z can be *nearly critical*. Therefore a complete explanation for the "convergence" of Krylov methods in finite precision requires to complement the classical point of view of exact *convergence* ($h \to 0$), valid when the arithmetic can be regarded as **exact**, by the novel notion of *criticality* ($|h| \to \infty$) which takes care of the effect of **finite precision** when they cannot be ignored.

The reader is refered to [5] to see precisely how this new notion clarifies the finite precision behaviour of such key aspects of Krylov methods as the Arnoldi residual, an algorithmic justification for restart and the extreme robustness to very large perturbations [15]. The notion of criticality offers therefore a theoretical justification for highly successful heuristics. It also shows why $|h_{k+1\ k}|$ small can be a misleading indicator for the nearness to exact reducibility.

References

[1] F. Chatelin. **Spectral approximation of linear operators**. Academic Press, New York, 1983.

[2] F. Chatelin. **Valeurs propres de matrices**. Masson, Paris, 1988.

[3] F. Chatelin. **Eigenvalues of matrices**. Wiley, Chichester, 1993. Enlarged Translation of the French Publication with Masson.

[4] F. Chaitin-Chatelin. About Singularities in Inexact Computing. Technical Report TR/PA/02/106, CERFACS, Toulouse, France, 2002.

[5] F. Chaitin-Chatelin. The Arnoldi method in the light of Homotopic Deviation theory. Technical Report TR/PA/03/15, CERFACS, Toulouse, France, 2003.

[6] F. Chaitin-Chatelin and V. Frayssé. **Lectures on Finite Precision Computation**. Publ., Philadelphia, 1996.

[7] F. Chaitin-Chatelin and E. Traviesas. Homotopic perturbation - Unfolding the field of singularities of a matrix by a complex parameter: a global geometric approach. Technical Report TR/PA/01/84, CERFACS, 2001.

[8] F. Chaitin-Chatelin and E. Traviesas. Qualitative Computing. Technical Report TR/PA/02/58, CERFACS, Toulouse, France, 2002. To appear in **Handbook of Computation**, B. Einarsson ed. , SIAM Philadelphia.

[9] F. Chaitin-Chatelin, V. Toumazou, E. Traviesas. *Accuracy assessment for eigencomputations: variety of backward errors and pseudospectra.* Lin. Alg. App. 309, 73-83, 2000. Also available as Cerfacs Rep. TR/PA/99/03.

[10] F. Chaitin-Chatelin and M.B. van Gijzen. Homotopic Deviation theory with an application to computational acoustics. Technical Report TR/PA/04/05, CERFACS, Toulouse, France, 2004.

[11] F. Chaitin-Chatelin. Computing beyond analyticity. Matrix Algorithms in Inexact and Uncertain Computing. Technical Report TR/PA/03/110, CERFACS, Toulouse, France, 2003.

[12] P. Lancaster, M. Tismenetsky. **Theory of Matrices**. Academic Press, New York, 1987.

[13] F. Chaitin-Chatelin, E. Traviesas and L. Plantié. *Understanding Krylov methods in finite precision* in Numerical Analysis and Applications, NAA 2000 (L. Vulkov, J. Wasviewski, P. Yalamov eds.), Springer Verlag Lectures Notes in CS, vol. 1988, pp. 187-197, 2000. Also available as Cerfacs Rep. TR/PA/00/40.

[14] F. Chaitin-Chatelin. Comprendre les méthodes de Krylov en précision finie : le programme du Groupe Qualitative Computing au CERFACS. Technical Report TR/PA/00/11, CERFACS, Toulouse, France, 2000.

[15] F. Chaitin-Chatelin and T. Meškauskas. Inner-outer iterations for mode solver in structural mechanics: application to the Code-Aster. Contract Report FR/PA/02/56, CERFACS, Toulouse, France, 2002.

[16] B.N. Parlett. **The Symmetric Eigenvalue Problem.** Prentice Hall, Englewood Cliffs, 1980.

[17] J. Moro, J.V. Burke and M.L. Overton, *On the Lidskii-Lyusternik-Vishik Perturbation Theory for Eigenvalues with Arbitrary Jordan Structure.* SIAM J. Matrix Anal. Appl. 18 (1997), pp. 793-817.

All Cerfacs Reports are available from:
`http://www.cerfacs.fr/algor/reports/index.html`