Matrix nearness problems and eigenvalue optimization

Nicola Guglielmi and Christian Lubich

January 30, 2023

Preface

Table of Contents

Chapter I.
Introduction by examples

Some from

- Complex, real and structured pseudospectra
- Structured distance to singularity
- Structured singular values
- Nearest singular matrix pencil
- H^∞ norm
- Complex, real and structured stability radii
- Nearest stable matrix
- Nearest passive or non-passive system
- Graphs: constrained partitioning, spectral clustering stability

Chapter II.
Basic eigenvalue optimization problems

chap:proto

We describe an algorithmic approach to a class of eigenvalue optimization problems that uses constrained gradient flows and the remarkable low-rank structure of the optimizers. This chapter is basic in the sense that we illustrate essential ideas and techniques on a particular problem class that will be vastly extended later in this book. The eigenvalue optimization problems considered in this chapter arise in computing complex, real or structured pseudoscectra or their extremal points such as those giving the pseudospectral abscissa and radius. These problems (and extensions thereof) will reappear as the principal building block in the two-level approach to various matrix nearness problems to be discussed in later chapters. The problems come in three variants: optimizing eigenvalues over unstructured complex perturbations of a given matrix, over real perturbations, or over structured perturbations, which are restricted to a given complex- or real-linear subspace of matrices, for example matrices with a given sparsity pattern or matrices with given range and co-range or Hamiltonian matrices. In all these cases there is a common underlying rank-1 property of optimizers that will be used to advantage in algorithms.

II.1 Unstructured complex case

c:proto-complex

II.1.1 Problem description

Let $A \in \mathbb{C}^{n,n}$ be a given matrix and let $\lambda(A) \in \mathbb{C}$ be a target eigenvalue of A, for example:

- the eigenvalue of minimal or maximal real part;
- the eigenvalue of minimal or maximal modulus;
- the closest eigenvalue to a given set in the complex plane.

Here the target eigenvalue need not depend continuously on the matrix A when several eigenvalues are simultaneously extremal, but it depends continuously on A when the extremal eigenvalue is unique.

The objective is to minimize a given function f of the target eigenvalue $\lambda(A+\Delta)$ over perturbation matrices Δ of a prescribed norm ε. We consider the following eigenvalue optimization problem: For a given $\varepsilon > 0$, find

$$\arg \min_{\mathit{\Delta} \in \mathbb{C}^{n,n},\, \|\mathit{\Delta}\|_F = \varepsilon} f\left(\lambda\left(A+\mathit{\Delta}\right), \overline{\lambda}\left(A+\mathit{\Delta}\right)\right), \tag{1.1} \qquad \boxed{\texttt{eq:optimiz0}}$$

where $\|\mathit{\Delta}\|_F$ is the Frobenius norm of the matrix $\mathit{\Delta}$, i.e. the Euclidean norm of the vector of matrix entries; where $\lambda(A+\mathit{\Delta})$ is the considered target eigenvalue of the perturbed matrix $A+\mathit{\Delta}$, and where

$$f : \mathbb{C}^2 \to \mathbb{C} \quad \text{with} \quad f\left(\lambda, \overline{\lambda}\right) = f\left(\overline{\lambda}, \lambda\right) \in \mathbb{R} \quad \text{for all } \lambda \in \mathbb{C} \tag{1.2} \qquad \boxed{\texttt{ass:f}}$$

is a given smooth function. While our theory applies to general functions f with (1.2), in our examples we often consider specific cases where f or $-f$ evaluated at $\left(\lambda, \overline{\lambda}\right)$ equals

$$\mathrm{Re}\,\lambda = \frac{\lambda + \overline{\lambda}}{2} \quad \text{or} \quad |\lambda|^2 = \lambda\overline{\lambda}.$$

We note that $\mathrm{Im}\,\lambda = \frac{1}{2\mathrm{i}}(\lambda - \overline{\lambda})$ does not satisfy (1.2), but this case can be included in the present setting by first rotating A to $-\mathrm{i}A$ and then considering the real part. The $\arg\max$ case is treated in the same way, replacing f by $-f$.

For example, as will be dicussed in detail in Chapter IV, the real part function is used in studying the distance to instability (or stability radius) of a matrix with all eigenvalues in the left complex half-plane. The interest is in computing the nearest matrix $A+\mathit{\Delta}$ to A for which the rightmost eigenvalue is on the imaginary axis. Here, “nearest” will refer to the Frobenius norm $\|\mathit{\Delta}\|_F$. Similarly, the squared modulus function is used when A is a matrix with all eigenvalues in the unit disk, to compute the nearest matrix $A+\mathit{\Delta}$ to A for which the eigenvalue of largest modulus is on the unit circle.

It is convenient to write

$$\mathit{\Delta} = \varepsilon E \quad \text{with } \|E\|_F = 1$$

and

$$F_\varepsilon(E) = f\left(\lambda\left(A+\varepsilon E\right), \overline{\lambda}\left(A+\varepsilon E\right)\right) \tag{1.3} \qquad \boxed{\texttt{Feps}}$$

so that Problem (1.1) is equivalent to finding

$$\arg \min_{E \in \mathbb{C}^{n,n},\, \|E\|_F = 1} F_\varepsilon(E). \tag{1.4} \qquad \boxed{\texttt{eq:optimiz}}$$

Problem (1.1) or (1.4) is a nonconvex, nonsmooth optimization problem.

In a variant to the above problem, the inequality constraints $\|\mathit{\Delta}\|_F \le \varepsilon$ and $\|E\|_F \le 1$ will also be considered in (1.1) and (1.4), respectively.

There are obvious generalizations to the above problem, which we will actually encounter in applications in later chapters:

- The objective function f might depend on several or all eigenvalues of $A+\mathit{\Delta}$ instead of only a single target eigenvalue.
- The objective function might depend also on eigenvectors of $A+\mathit{\Delta}$.

However, in this chapter we shall only consider the function f as in (1.1)–(1.2).

II.1.2 Free gradient

Let us begin with some notations and normalizations. Let x and y be left and right eigenvectors, respectively, associated with a simple eigenvalue λ of a matrix M: $x, y \in \mathbb{C}^n \setminus \{0\}$ with $x^*M = \lambda x^*$ and $My = \lambda y$, where $x^* = \overline{x}^\top$. Unless specified differently, we assume that the eigenvectors are normalized such that

$$\|x\| = \|y\| = 1 \quad \text{and} \quad x^*y \text{ is real and positive.} \tag{1.5}$$

eq:scaling

The norm $\|\cdot\|$ is chosen as the Euclidean norm. Any pair of left and right eigenvectors x and y can be scaled in this way.

We denote by

$$\langle X, Y \rangle = \sum_{i,j} \overline{x}_{ij} y_{ij} = \mathrm{tr}(X^*Y)$$

the inner product in $\mathbb{C}^{n,n}$ that induces the Frobenius norm $\|X\|_F = \langle X, X\rangle^{1/2}$.

The following lemma will allow us to compute the steepest descent direction of the functional F_ε.

lem:gradient **Lemma 1.1 (Free gradient).** *Let $E(t) \in \mathbb{C}^{n,n}$, for real t near t_0, be a continuously differentiable path of matrices, with the derivative denoted by $\dot{E}(t)$. Assume that $\lambda(t)$ is a simple eigenvalue of $A + \varepsilon E(t)$ depending continuously on t, with associated left and right eigenvectors $x(t)$ and $y(t)$ satisfying* (1.5)*, and let the eigenvalue condition number be*

$$\kappa(t) = \frac{1}{x(t)^* y(t)} > 0.$$

Then, $F_\varepsilon(E(t)) = f\big(\lambda(t), \overline{\lambda(t)}\big)$ is continuously differentiable w.r.t. t and we have

$$\frac{1}{\varepsilon\kappa(t)} \frac{d}{dt} F_\varepsilon(E(t)) = \mathrm{Re}\,\big\langle G_\varepsilon(E(t)), \dot{E}(t)\big\rangle, \tag{1.6}$$

eq:deriv

where the (rescaled) gradient of F_ε is the rank-1 matrix

$$G_\varepsilon(E) = 2 f_{\overline{\lambda}}\, x y^* \in \mathbb{C}^{n,n} \tag{1.7}$$

eq:freegrad

with $f_{\overline{\lambda}} = \dfrac{\partial f}{\partial \overline{\lambda}}(\lambda, \overline{\lambda})$ for the eigenvalue $\lambda = \lambda(A + \varepsilon E)$ and the corresponding left and right eigenvectors x and y normalized by (1.5).

Proof. We first observe that (1.2) implies $f_{\overline{\lambda}} = \overline{f_\lambda} = \overline{\dfrac{\partial f}{\partial \lambda}(\lambda, \overline{\lambda})}$. Using Theorem VIII.1.1, we obtain that $F_\varepsilon(E(t))$ is differentiable with

$$\begin{aligned} \frac{d}{dt} F_\varepsilon\left(E(t)\right) &= f_\lambda\, \dot{\lambda} + f_{\overline{\lambda}}\, \dot{\overline{\lambda}} \\ &= \frac{\varepsilon}{x^*y}\left(f_\lambda\, x^*\dot{E}y + f_{\overline{\lambda}}\, \overline{x^*\dot{E}y}\right) = \frac{\varepsilon}{x^*y}\, 2\,\mathrm{Re}\left(f_\lambda\, x^*\dot{E}y\right), \end{aligned} \tag{1.8}$$

where we omit the omnipresent dependence on t on the right-hand side. Noting that

$$\mathrm{Re}\big(f_\lambda\, x^* \dot{E} y\big) = \mathrm{Re}\,\big\langle \overline{f_\lambda}\, xy^*, \dot{E} \big\rangle,$$

we obtain (3.14)–(3.15). □

ex:G **Example 1.2.** For $f(\lambda,\overline{\lambda}) = -\frac{1}{2}(\lambda+\overline{\lambda}) = -\mathrm{Re}\,\lambda$ we have $2f_{\overline{\lambda}} = -1$ and hence $G_\varepsilon(E) = -xy^*$, which is nonzero for all λ. For $f(\lambda,\overline{\lambda}) = -\frac{1}{2}|\lambda|^2 = -\frac{1}{2}\lambda\overline{\lambda}$ we have $2f_{\overline{\lambda}} = -\lambda$. In this case we obtain $G_\varepsilon(E) = -\lambda\, xy^*$, which is nonzero whenever $\lambda \neq 0$.

II.1.3 Projected gradient

To comply with the constraint $\|E(t)\|_F^2 = 1$, we must have

$$0 = \frac{1}{2}\,\frac{d}{dt}\|E(t)\|_F^2 = \mathrm{Re}\,\langle E(t), \dot{E}(t)\rangle. \tag{1.9}$$

eq:normconstr

In view of Lemma 1.1 we are thus led to the following constrained optimization problem for the admissible direction of steepest descent.

lem:opt **Lemma 1.3 (Direction of steepest admissible descent).** *Let $E, G \in \mathbb{C}^{n,n}$ with $\|E\|_F = 1$. A solution of the optimization problem*

$$Z_\star \;=\; \arg \min_{\|Z\|_F=1,\, \mathrm{Re}\,\langle E,Z\rangle=0} \mathrm{Re}\,\langle G, Z\rangle, \tag{1.10}$$

is given by

$$\mu Z_\star = -G + \mathrm{Re}\langle G, E\rangle\, E, \tag{1.11}$$

eq:Eopt

where μ is the Frobenius norm of the matrix on the right-hand side. The solution is unique if G is not a multiple of E.

Proof. The result follows on noting that the real part of the complex inner product on $\mathbb{C}^{n,n}$ is a real inner product on $\mathbb{R}^{2n,2n}$, and the real inner product with a given vector (which here is a matrix) is maximized over a subspace by orthogonally projecting the vector onto that subspace. The expression in (1.11) is the orthogonal projection of $-G$ onto the orthogonal complement of the span of E, which is the tangent space at E of the manifold of matrices of unit Frobenius norm. □

II.1.4 Norm-constrained gradient flow

Lemmas 1.1 and 1.3 show that the admissible direction of steepest descent of the functional F_ε at a matrix E of unit Frobenius norm is given by the positive multiples of the matrix $-G_\varepsilon(E) + \mathrm{Re}\,\langle G_\varepsilon(E), E\rangle E$. This leads us to consider the (rescaled) *gradient flow on the manifold of $n \times n$ complex matrices of unit Frobenius norm*:

$$\dot{E} = -G_\varepsilon(E) + \mathrm{Re}\,\langle G_\varepsilon(E), E\rangle E, \tag{1.12}$$

ode-E

where we omitted the ubiquitous argument t.

By construction of this ordinary differential equation, we have $\mathrm{Re}\langle E, \dot{E}\rangle = 0$ along its solutions, and so the Frobenius norm 1 is conserved. Since we follow the admissible direction of steepest descent of the functional F_ε along solutions $E(t)$ of this differential equation, we obtain the following monotonicity property.

thm:monotone

Theorem 1.4 (Monotonicity). *Assume that $\lambda(t)$ is a simple eigenvalue of $A+\varepsilon E(t)$ depending continuously on t. Let $E(t)$ of unit Frobenius norm satisfy the differential equation* (4.13). *Then,*

$$\frac{d}{dt} F_\varepsilon(E(t)) \le 0. \tag{1.13}$$

eq:pos

Proof. Although the result follows directly from Lemmas 1.1 and 1.3, we compute the derivative explicitly. We write $G = G_\varepsilon(E)$ for short and take the inner product of (4.13) with $\dot{E}$. Using that $\mathrm{Re}\langle E, \dot{E}\rangle = 0$, we find

$$\|\dot{E}\|_F^2 = -\mathrm{Re}\langle G - \mathrm{Re}\langle G, E\rangle E, \dot{E}\rangle = -\mathrm{Re}\langle G, \dot{E}\rangle$$

and hence (4.13) and Lemma 1.1 yield

$$\frac{1}{\varepsilon\kappa}\,\frac{d}{dt} F_\varepsilon(E(t)) = -\|G - \mathrm{Re}\,\langle G, E\rangle E\|_F^2 \le 0, \tag{1.14}$$

c-s

which gives the precise rate of decay of F_ε along a trajectory $E(t)$ of (4.13). □

The stationary points of the differential equation (4.13) are characterized as follows.

thm:stat

Theorem 1.5 (Stationary points). *Let $E_\star \in \mathbb{C}^{n,n}$ with $\|E_\star\|_F = 1$ be such that*

(i) The target eigenvalue $\lambda(A+\varepsilon E)$ is simple at $E = E_\star$ and depends continuously on E in a neighborhood of $E_\star$.
(ii) The gradient $G_\varepsilon(E_\star)$ is nonzero.

Let $E(t) \in \mathbb{C}^{n,n}$ be the solution of (4.13) *passing through $E_\star$. Then the following are equivalent:*

1. $\dfrac{d}{dt} F_\varepsilon\,(E(t)) = 0.$

2. $\dot{E} = 0.$

3. $E_\star$ is a real multiple of $G_\varepsilon(E_\star)$.

Proof. Clearly, 3. implies 2., which implies 1. Finally, (1.14) shows that 1. implies 3. □

Remark 1.6 (Degeneracies). In degenerate situations where $G_\varepsilon(E_\star) = 0$, we cannot conclude from 2. to 3., i.e., that the stationary point is a multiple of $G_\varepsilon(E_\star)$. For the case $f(\lambda, \overline{\lambda}) = -\mathrm{Re}\,\lambda$ we have seen in Example 1.2 that $G_\varepsilon(E) = -xy^* \neq 0$, where x, y

are normalized eigenvectors to the target eigenvalue $\lambda(A+\varepsilon E)$. For $f(\lambda,\overline{\lambda}) = -\frac{1}{2}|\lambda|^2$ we have $G_\varepsilon(E) = -\lambda x y^* \neq 0$ for $\lambda \neq 0$. For other functions f we might encounter $G_\varepsilon(E_\star) = 0$, but such a degeneracy can be regarded as an exceptional situation, which will not be considered further.

rem:stat-min **Remark 1.7 (Stationary points and optimizers).** Every global minimum is a local minimum, and every local minimum is a stationary point. The converse is clearly not true. Stationary points of the gradient system that are not a local minimum, are unstable. It can thus be expected that generically a trajectory will end up in a local minimum. Running several different trajectories can reduce the risk of being caught in a local minimum instead of a global minimum.

rem:ineq **Remark 1.8 (Inequality constraints).** When we have the inequality constraint $\|\varDelta\|_F \leq \varepsilon$ in (1.1) or equivalently $\|E\|_F \leq 1$ in (1.4), the situation changes only slightly. If $\|E\|_F < 1$, every direction is admissible, and the direction of steepest descent is given by the negative gradient $-G_\varepsilon(E)$. So we choose the free gradient flow

$$\dot{E} = -G_\varepsilon(E) \qquad \text{as long as } \ \|E(t)\|_F < 1. \tag{1.15}$$

ode-E-free

When $\|E(t)\|_F = 1$, then there are two possible cases. If $\mathrm{Re}\,\langle G_\varepsilon(E), E\rangle \geq 0$, then the solution of (1.15) has (omitting the argument t)

$$\frac{d}{dt}\|E(t)\|_F^2 = 2\,\mathrm{Re}\,\langle \dot{E}, E\rangle = -2\,\mathrm{Re}\,\langle G_\varepsilon(E), E\rangle \leq 0,$$

and hence the solution of (1.15) remains of Frobenius norm at most 1.

Else, if $\mathrm{Re}\,\langle G_\varepsilon(E), E\rangle < 0$, the admissible direction of steepest descent is given by the right-hand side of (4.13), i.e. $-G_\varepsilon(E) + \mathrm{Re}\,\langle G_\varepsilon(E), E\rangle E$, and so we choose that differential equation to evolve E. The situation can be summarized as taking, if $\|E(t)\|_F = 1$,

$$\dot{E} = -G_\varepsilon(E) + \mu E \quad \text{ with } \ \mu = \min\bigl(0, \mathrm{Re}\,\langle G_\varepsilon(E), E\rangle\bigr). \tag{1.16}$$

ode-E-mu

Along solutions of (1.16), the functional F_ε decays monotonically, and stationary points of (1.16) with $G_\varepsilon(E) \neq 0$ are characterized, by the same argument as in Theorem 1.5, as

$$E \text{ is a } \textit{negative} \text{ real multiple of } G_\varepsilon(E). \tag{1.17}$$

stat-neg

If it can be excluded that the gradient G_ε vanishes at an optimizer (as in Example 1.2), it can thus be concluded that the optimizer of the problem with inequality constraints is a stationary point of the gradient flow (4.13) for the problem with equality constraints.

rem:mult-eig **Remark 1.9 (Multiple and discontinuous eigenvalues).** We mention some situations where the assumption of a smoothly evolving simple eigenvalue is violated. As such situations are either non-generic or can happen generically only at isolated times t, they do not affect the computation after discretization of the differential equation.

— Along a trajectory $E(t)$, the target eigenvalue $\lambda(t) = \lambda(A + \varepsilon E(t))$ may become discontinuous. For example, in the case of the eigenvalue of largest real part, a different branch of eigenvalues may get to have the largest real part. In such a case of discontinuity, the differential equation is further solved, with descent of the largest real part until finally a stationary point is approximately reached.

— A multiple eigenvalue $\lambda(t)$ may occur at some finite t because of a coalescence of eigenvalues. Even if some continuous trajectory runs into a coalescence, this is non-generic to happen after discretization of the differential equation, and so the computation will not be affected.

— A multiple eigenvalue may appear in a stationary point, in the limit $t \to \infty$. The computation will stop before, and items 1.-3. in Theorem 1.5 will then be satisfied approximately, in view of (1.14).

Although the situations above do not affect the time-stepping of the gradient system, close-to-multiple eigenvalues do impair the accuracy of the computed left and right eigenvectors that appear in the gradient.

II.1.5 Rank-1 property of optimizers

We call an optimizer $E_\star$ of (1.4) *non-degenerate* if conditions (i) and (ii) of Theorem 1.5 are satisfied. Since optimizers are necessarily stationary points of the norm-constrained gradient flow (3.6), Theorem 1.5 and Lemma 1.1 immediately imply the following remarkable property.

cor:rank-1

Corollary 1.10 (Rank of optimizers). *If $E_\star$ is a non-degenerate optimizer of problem* (1.4)*, then $E_\star$ is of rank* 1.

Let us summarize how this rank-1 property came about: An optimizer is a stationary point of the norm-constrained gradient flow (4.13). This implies that the optimizer E is a real multiple of the free gradient $G_\varepsilon(E)$, which is of rank 1 as a consequence of the derivative formula for simple eigenvalues.

This corollary motivates us to search for a differential equation on the manifold of rank-1 matrices of norm 1 with the property that the functional F_ε decreases along its solutions and has the same stationary points as the differential equation (4.13). Working with rank-1 matrices $E = uv^*$ given by two vectors $u, v \in \mathbb{C}^n$ instead of general complex $n \times n$ matrices is computationally favourable, especially for high dimensions n, for two independent reasons:

(i) Storage and computations are substantially reduced when the two n-vectors u, v are used instead of the full $n \times n$ matrix E.

(ii) The computation of the target eigenvalue $\lambda(t)$ of $A + \varepsilon E(t)$ using inverse iteration is largely simplified thanks to the Sherman-Morrison formula

$$(A + \varepsilon uv^* - \mu I)^{-1} = (A - \mu I)^{-1} - \frac{(A - \mu I)^{-1} \varepsilon uv^* (A - \mu I)^{-1}}{1 + v^*(A - \mu I)^{-1} \varepsilon u}.$$

Moreover, after transforming the given matrix $A \in \mathbb{C}^{n,n}$ to Hessenberg form by a unitary similarity transformation, linear systems with the shifted matrix $A - \mu I$ for varying shifts μ can be solved with $O(n^2)$ operations each.

For sparse matrices A, Krylov subspace methods for the perturbed matrix $A + \varepsilon E$ take advantage when E is of rank 1, since matrix-vector products with $E = uv^*$ just require computing an inner product with v.

II.1.6 Rank-1 matrices and their tangent matrices

subsec:rank-1

We denote by $\mathcal{R}_1 = \mathcal{R}_1(\mathbb{C}^{n,n})$ the manifold of complex rank-1 matrices of dimension $n \times n$ and write $E \in \mathcal{R}_1$ in a non-unique way as

$$E = \sigma u v^*,$$

where $\sigma \in \mathbb{C} \setminus \{0\}$ and $u, v \in \mathbb{C}^n$ have unit norm. The tangent space $T_E\mathcal{R}_1$ at $E \in \mathcal{R}_1$ consists of the derivatives of paths in $\mathcal{R}_1$ passing through E. Tangent matrices $\dot{E} \in T_E\mathcal{R}_1$ are then of the form

$$\dot{E} = \dot{\sigma} u v^* + \sigma \dot{u} v^* + \sigma u \dot{v}^*, \tag{1.18}$$

E-dot-1

where $\dot{\sigma} \in \mathbb{C}$ is arbitrary and $\dot{u}, \dot{v} \in \mathbb{C}^n$ are such that $\mathrm{Re}(u^*\dot{u}) = 0$ and $\mathrm{Re}(v^*\dot{v}) = 0$ (because of the norm constraint on u and v). They are uniquely determined by $\dot{E}$ and σ, u, v if we impose the orthogonality conditions $u^*\dot{u} = 0,\ v^*\dot{v} = 0$. Multiplying $\dot{E}$ with u^* from the left and with v from the right, we then obtain

$$\dot{\sigma} = u^*\dot{E}v, \quad \sigma\dot{u} = \dot{E}v - \dot{\sigma}u, \quad \sigma\dot{v}^* = u^*\dot{E} - \dot{\sigma}v^*. \tag{1.19}$$

sigma-u-v-Edot

Extending this construction, we arrive at a useful explicit formula for the projection onto the tangent space that is orthogonal with respect to the Frobenius inner product $\langle \cdot, \cdot \rangle$.

em:P-formula-1

Lemma 1.11 (Rank-1 tangent space projection). *The orthogonal projection from $\mathbb{C}^{n,n}$ onto the tangent space $T_E\mathcal{R}_1$ at $E = \sigma uv^* \in \mathcal{R}_1$ is given by*

$$P_E(Z) = Z - (I - uu^*)Z(I - vv^*) \quad \textit{for } Z \in \mathbb{C}^{n,n}. \tag{1.20}$$

P-formula-1

Proof. Let $P_E(Z)$ be defined by (1.20). To prove that $P_E(Z) \in T_E\mathcal{R}_1$, we show that $P_E(Z)$ can be written in the form (1.18). Let $\dot{\sigma}, \dot{u}, \dot{v}$ be defined like in (1.19), but now with $\dot{E} \in T_E\mathcal{R}_1$ replaced by arbitrary $Z \in \mathbb{C}^{n,n}$, i.e.,

$$\dot{\sigma} = u^*Zv, \quad \sigma\dot{u} = Zv - \dot{\sigma}u, \quad \sigma\dot{v}^* = u^*Z - \dot{\sigma}v^*. \tag{1.21}$$

sigma-u-v-Z

We obtain the corresponding matrix in the tangent space $T_E\mathcal{R}_1$, see (1.18), as

$$\begin{aligned}
&\dot{\sigma}uv^* + \sigma\dot{u}v^* + \sigma u\dot{v}^* \\
&= \dot{\sigma}uv^* + (Zv - \dot{\sigma}u)v^* + u(u^*Z - \dot{\sigma}v^*) \\
&= Zvv^* - uu^*Zvv^* + uu^*Z = P_E(Z).
\end{aligned}$$

This shows that

$$P_E(Z) = \dot\sigma u v^* + \sigma \dot u v^* + \sigma u \dot v^* \in T_E\mathcal{R}_1. \tag{1.22}$$

P-formula-dots

Furthermore,

$$\langle P_E(Z), \dot E\rangle = \langle Z, \dot E\rangle \qquad \text{for all } \dot E \in T_E\mathcal{R}_1,$$

because $\langle (I-uu^*)Z(I-vv^*), \dot E\rangle = \langle Z, (I-uu^*)\dot E(I-vv^*)\rangle = 0$ by (1.18). Hence, $P_E(Z)$ is indeed the orthogonal projection of Z onto $T_E\mathcal{R}_1$. □

We note that $P_E(E) = E$ for $E \in \mathcal{R}_1$, or equivalently, $E \in T_E\mathcal{R}_1$, which will be an often used property.

II.1.7 Rank-1 constrained gradient flow

-gradient-flow

In the differential equation (4.13) we project the right-hand side to the tangent space $T_E\mathcal{R}_1$:

$$\dot E = -P_E\Big(G_\varepsilon(E) - \mathrm{Re}\langle G_\varepsilon(E), E\rangle E\Big). \tag{1.23}$$

ode-E-1

This yields a differential equation on the rank-1 manifold $\mathcal{R}_1$. In view of Lemma 1.1, it is the (rescaled) gradient flow of the functional F_ε constrained to the manifold $\mathcal{R}_1$.

Assume now that for some t, the Frobenius norm of $E = E(t)$ is 1. Since $P_E(E) = E$, we have with $Z = -G_\varepsilon(E) + \mathrm{Re}\langle G_\varepsilon(E), E\rangle E$ that

$$\mathrm{Re}\,\langle E, \dot E\rangle = \mathrm{Re}\,\langle E, P_E(Z)\rangle = \mathrm{Re}\,\langle P_E(E), Z\rangle = \mathrm{Re}\,\langle E, Z\rangle = 0.$$

Hence, solutions $E(t)$ of (1.23) stay of Frobenius norm 1 for all t.

The proof of Lemma 1.11 also provides the following differential equations for the factors of $E(t) = \sigma(t)u(t)v(t)^*$, which can be discretized by standard numerical integrators.

lem:suv-1

Lemma 1.12 (Differential equations for the three factors). *For $E = \sigma u v^* \in \mathcal{R}_1$ with nonzero $\sigma \in \mathbb{C}$ and with $u \in \mathbb{C}^n$ and $v \in \mathbb{C}^n$ of unit norm, the equation $\dot E = P_E(Z)$ is equivalent to $\dot E = \dot\sigma u v^* + \sigma \dot u v^* + \sigma u \dot v^*$, where*

$$\begin{aligned} \dot\sigma &= u^* Z v \\ \dot u &= (I - uu^*) Z v \sigma^{-1} \\ \dot v &= (I - vv^*) Z^* u \overline{\sigma}^{-1}. \end{aligned} \tag{1.24}$$

Proof. The result follows immediately from (1.21) and (1.22). □

Since we are only interested in solutions of Frobenius norm 1 of (1.23), we can simplify the representation of E to $E = uv^*$ with u and v of unit norm (without the extra factor σ of unit modulus).

lem:uv-1 **Lemma 1.13 (Differential equations for the two vectors).** *For an initial value $E(0) = u(0)v(0)^*$ with $u(0)$ and $v(0)$ of unit norm, the solution of* (1.23) *is given as $E(t) = u(t)v(t)^*$, where u and v solve the system of differential equations (for $G = G_\varepsilon(E)$)*

$$\begin{aligned} \dot u &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(u^*Gv)u - (I-uu^*)Gv \\ \dot v &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(v^*G^*u)v - (I-vv^*)G^*u, \end{aligned} \tag{1.25}$$

ode-uv

which preserves $\|u(t)\| = \|v(t)\| = 1$ for all t.

We note that for $G = G_\varepsilon(E) = 2f_{\overline{\lambda}}\,xy^*$ (see Lemma 1.1) and with $\alpha = u^*x$, $\beta = v^*y$ and $\gamma = 2f_{\overline{\lambda}}$ we obtain the differential equations

$$\begin{aligned} \dot u &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\alpha\overline{\beta}\gamma)u + \alpha\overline{\beta}\gamma\, u - \overline{\beta}\gamma\, x \\ \dot v &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\overline{\alpha}\beta\overline{\gamma})v + \overline{\alpha}\beta\overline{\gamma}\, v - \overline{\alpha\gamma}\, y. \end{aligned} \tag{1.26}$$

ode-uv-short

Proof. We introduce the projection $\widetilde{P}_E$ onto the tangent space at $E = uv^*$ of the submanifold of rank-1 matrices of unit Frobenius norm,

$$\widetilde{P}_E(G) = P_E(G - \mathrm{Re}\langle G, E\rangle E) = P_E(G) - \mathrm{Re}\langle G, E\rangle E.$$

We find

$$\begin{aligned} \widetilde{P}_E(G) &= Gvv^* - uu^*Gvv^* + uu^*G - \mathrm{Re}\langle G, uv^*\rangle uv^* \\ &= (I-uu^*)Gvv^* + uu^*G(I-vv^*) + uu^*Gvv^* - \mathrm{Re}(u^*Gv)uv^* \\ &= (I-uu^*)Gvv^* + uu^*G(I-vv^*) + \mathrm{i}\,\mathrm{Im}(u^*Gv)uv^* \\ &= \Bigl(\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(u^*Gv)u + (I-uu^*)Gv\Bigr)v^* + u\Bigl(\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(u^*Gv)v^* + u^*G(I-vv^*)\Bigr). \end{aligned}$$

For $\dot E = \dot u v^* + u\dot v^*$ we thus have $\dot E = -\widetilde{P}_E(G)$ if u and v satisfy (1.25). Since then $\mathrm{Re}(u^*\dot u) = 0$ and $\mathrm{Re}(v^*\dot v) = 0$, the unit norm of u and v is preserved. □

The projected differential equation (1.23) has the same monotonicity property as the differential equation (4.13).

m:monotone-C-1 **Theorem 1.14 (Monotonicity).** *Let $E(t) \in \mathcal{R}_1$ of unit Frobenius norm be a solution to the differential equation* (1.23). *If $\lambda(t)$ is a simple eigenvalue of $A + \varepsilon E(t)$, then*

$$\frac{d}{dt}F_\varepsilon\bigl(E(t)\bigr) \le 0. \tag{1.27}$$

eq:pos-C-1

Proof. As in the proof of Theorem 1.4, we abbreviate $G = G_\varepsilon(E)$ and obtain from (1.23) and $\dot E \in T_E\mathcal{R}_1$ and $\mathrm{Re}\langle E, \dot E\rangle = 0$ that

$$\|\dot E\|_F^2 = -\mathrm{Re}\bigl\langle P_E\bigl(G - \mathrm{Re}\langle G, E\rangle E\bigr), \dot E\bigr\rangle = -\mathrm{Re}\bigl\langle G - \mathrm{Re}\langle G, E\rangle E, \dot E\bigr\rangle = -\langle G, \dot E\rangle$$

and hence Lemma 1.1 and (1.23) yield

$$\frac{1}{\varepsilon\kappa}\,\frac{d}{dt}F_\varepsilon(E(t)) = -\bigl\|P_E\bigl(G - \mathrm{Re}\langle G, E\rangle E\bigr)\bigr\|_F^2, \tag{1.28}$$

c-s-1

which yields the monotone decay. □

Comparing the differential equations (4.13) and (1.23) immediately shows that every stationary point of (4.13) is also a stationary point of the projected differential equation (1.23). Remarkably, the converse is also true for the stationary points E of unit Frobenius norm with $P_E(G_\varepsilon(E)) \neq 0$. Violation of this non-degeneracy condition is very exceptional, as we will explain below.

thm:stat-1

Theorem 1.15 (Stationary points). *Let the rank-1 matrix $E \in \mathcal{R}_1$ be of unit Frobenius norm and assume that $P_E(G_\varepsilon(E)) \neq 0$. If E is a stationary point of the rank-1 projected differential equation* (1.23)*, then E is already a stationary point of the differential equation* (4.13)*.*

Proof. We show that E is a nonzero real multiple of $G_\varepsilon(E)$. By Theorem 1.5, E is then a stationary point of the differential equation (4.13).

For a stationary point E of (1.23), we must have equality in (1.28), which shows that $P_E(G)$ (again with $G = G_\varepsilon(E)$) is a nonzero real multiple of E. Hence, in view of $P_E(E) = E$, we can write G as

$$G = \mu E + W, \qquad \text{where } \mu \neq 0 \text{ is real and } P_E(W) = 0.$$

Since E is of rank 1 and of unit Frobenius norm, E can be written as $E = uv^*$ with $\|u\| = \|v\| = 1$. We then have

$$W = W - P_E(W) = (I - uu^*)W(I - vv^*).$$

On the other hand, $G = 2\overline{f}_\lambda xy^*$ is also of rank 1. So we have

$$2\overline{f}_\lambda xy^* = \mu uv^* + (I - uu^*)W(I - vv^*).$$

Multiplying from the right with v yields that x is a complex multiple of u, and multiplying from the left with u^* yields that y is a complex multiple of v. Hence, G is a complex multiple of E. Since we already know that $P_E(G)$ is a nonzero real multiple of $P_E(E) = E$, it follows that G is the same real multiple of E. By Theorem 1.5, E is therefore a stationary point of the differential equation (4.13). □

em:exceptional

Remark 1.16 (Non-degeneracy condition). Let us discuss the condition $P_E(G_\varepsilon(E)) \neq 0$. We recall that $G = G_\varepsilon(E)$ is a multiple of xy^*, where x and y are left and right eigenvectors, respectively, to the eigenvalue λ of $A + \varepsilon E$. In which situation can we have $P_E(G) = 0$ whereas $G \neq 0$?

For $E = uv^*$, $P_E(G) = 0$ implies $G = (I - uu^*)G(I - vv^*)$, which yields $Gv = 0$ and $u^*G = 0$ and therefore $y^*v = 0$ and $u^*x = 0$. So we have $Ey = 0$ and $x^*E = 0$. This implies that λ is already an eigenvalue of A with the same left and right eigenvectors x, y as for $A + \varepsilon E$, which is a very exceptional situation.

II.1.8 Numerical integration by a splitting method

ec:proto-numer

We need to integrate numerically the differential equations (1.26), viz.

$$\begin{aligned}\dot{u} &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\alpha\overline{\beta}\gamma)u + \alpha\overline{\beta}\gamma\, u - \overline{\beta}\gamma\, x\\ \dot{v} &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\overline{\alpha}\beta\overline{\gamma})v + \overline{\alpha}\beta\overline{\gamma}\, v - \overline{\alpha\gamma}\, y,\end{aligned}$$

where $\alpha = u^*x \in \mathbb{C}$, $\beta = v^*y \in \mathbb{C}$ and $\gamma = 2f_{\overline{\lambda}} \in \mathbb{C}$.

The objective here is not to follow a particular trajectory accurately, but to arrive quickly at a stationary point. The simplest method is the normalized Euler method, where the result after an Euler step (i.e., a steepest descent step) is normalized to unit norm for both the u- and v-component. This can be combined with an Armijo-type line-search strategy to determine the step size adaptively.

We found, however, that a more efficient method is obtained with a *splitting method* instead of the Euler method. The splitting method consists of a first step applied to the differential equations

$$\begin{aligned}\dot{u} &= \alpha\overline{\beta}\gamma\, u - \overline{\beta}\gamma\, x\\ \dot{v} &= \overline{\alpha}\beta\overline{\gamma}\, v - \overline{\alpha\gamma}\, y\end{aligned} \qquad (1.29)$$ ode-uv-horiz

followed by a step for the differential equations

$$\begin{aligned}\dot{u} &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\alpha\overline{\beta}\gamma)u\\ \dot{v} &= -\tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\overline{\alpha}\beta\overline{\gamma})v.\end{aligned} \qquad (1.30)$$ ode-uv-rot

As the next lemma shows, the first differential equation moves λ in the direction of $-\gamma = -2f_{\overline{\lambda}}$. In particular, the motion is horizontal if $f_{\overline{\lambda}}$ is always real. The second differential equation is a mere rotation of u and v.

em:gamma-motion **Lemma 1.17 (Eigenvalue motion in the direction of $-f_{\overline{\lambda}}$).** *Along a path of simple eigenvalues $\lambda(t)$ of $A + \varepsilon u(t)v(t)^*$, where u, v of unit norm solve* (1.29)*, we have that*

$$\dot{\lambda}(t) \textit{ is a nonnegative real multiple of } -\frac{\partial f}{\partial \overline{\lambda}}(\lambda(t), \overline{\lambda}(t)).$$

Proof. The standard perturbation theory of eigenvalues shows that

$$\dot{\lambda} = \frac{1}{x^*y}\left(x^*\frac{d}{dt}(A + \varepsilon uv^*)\, y\right) = \varepsilon\,\frac{x^*\,(\dot{u}v^* + u\dot{v}^*)\,y}{x^*y}.$$

With $\alpha = u^*x$ and $\beta = v^*y$ and with x, y normalized by (1.5), we obtain from (1.29)

$$\frac{\dot{\lambda}}{\gamma} = -\frac{\varepsilon}{x^*y}\Big(|\alpha|^2\cdot\big(1 - |\beta|^2\big) + |\beta|^2\cdot\big(1 - |\alpha|^2\big)\Big) \in \mathbb{R},\ \le 0,$$

which proves the statement, since $\gamma = 2f_{\overline{\lambda}}$. □

In general, splitting methods do not preserve stationary points. Here, it does.

lem:stat-split **Lemma 1.18 (Stationary points).** *If (u, v) is a stationary point of the differential equations* (1.26)*, then it is also a stationary point of the differential equations* (1.29) *and* (1.30).

Proof. If (u, v) is a stationary point of (1.26), then u is proportional to x and v is proportional to y. Hence, $x = \alpha u$ and $y = \beta v$. This implies that (u, v) is a stationary point of (1.29), and hence also of (1.30). □

Fully discrete splitting algorithm. Starting from initial values u_k, v_k, we denote by x_k and y_k the left and right eigenvectors to the target eigenvalue λ_k of $A + \varepsilon u_k v_k^*$, and set

$$\alpha_k = u_k^* x_k, \qquad \beta_k = v_k^* y_k, \qquad \gamma_k = 2 f_{\overline{\lambda}_k}. \tag{1.31}$$

alpha-beta-gamma-n

We apply the Euler method with step size h to (1.29) to obtain

$$\begin{aligned} \widehat{u}(h) &= u_k + h\left(\alpha_k \overline{\beta}_k \gamma_k \, u_k - \overline{\beta}_k \gamma_k \, x_k\right) \\ \widehat{v}(h) &= v_k + h\left(\overline{\alpha}_k \beta_k \overline{\gamma}_k \, v_k - \overline{\alpha_k \gamma_k} \, y_k\right), \end{aligned} \tag{1.32}$$

eul-horiz

followed by a normalization to unit norm

$$\widetilde{u}(h) = \frac{\widehat{u}(h)}{\|\widehat{u}(h)\|}, \quad \widetilde{v}(h) = \frac{\widehat{v}(h)}{\|\widehat{v}(h)\|}. \tag{1.33}$$

eq:normal

Then, as a second step, we integrate the rotating differential equations (1.30) by setting, with $\vartheta = -\frac{1}{2}\,\mathrm{Im}\,\left(\alpha_k \overline{\beta_k} \gamma_k\right)$,

$$u(h) = \mathrm{e}^{\mathrm{i}\vartheta h}\, \widetilde{u}(h), \qquad v(h) = \mathrm{e}^{-\mathrm{i}\vartheta h}\, \widetilde{v}(h), \tag{1.34}$$

eq:rotate

and compute the target eigenvalue $\lambda(h)$ of $A + \varepsilon u(h) v(h)^*$. We note that this fully discrete algorithm still preserves stationary points.

One motivation for choosing this method is that near a stationary point, the motion is almost rotational since $x \approx \alpha u$ and $y \approx \beta v$. The dominating term determining the motion is then the rotational term on the right-hand side of (1.26), which is integrated by a rotation in the above scheme (the integration would be exact if α, β, γ were constant).

This algorithm requires in each step one computation of rightmost eigenvalues and associated eigenvectors of rank-1 perturbations to the matrix A, which can be computed at relatively small computational cost for large sparse matrices A, either combining the Cayley transformation approach with the Sherman-Morrison formula or by using an implicitly restarted Arnoldi method (as implemented in ARPACK and used in the MATLAB function *eigs*).

(We also tried a variant where α, β, γ in the rotation step are updated from $(\widetilde{u}(h), \widetilde{v}(h))$ and the left and right eigenvectors to the target eigenvalue $\widetilde{\lambda}(h)$ of $A + \varepsilon \widetilde{u}(h) \widetilde{v}(h)^*$. In our numerical experiments we found, however, that the slight improvement in the speed of convergence to the stationary state does not justify the nearly doubled computational cost per step.)

Step size selection. We use an Armijo-type line search strategy to determine a step size that reduces the functional $f(\lambda, \overline{\lambda})$. For the non-discretized differential equation (1.23), we know from (1.28) that the decay rate is given by

$$\frac{d}{dt} F_\varepsilon(E(t)) = -\varepsilon\kappa \Big(\|P_E(G)\|_F^2 - \big(\mathrm{Re}\,\langle G, E\rangle\big)^2 \Big) \leq 0.$$

Here we note that for $E = uv^*$ and again with $\alpha = u^* x$, $\beta = v^* x$, $\gamma = 2 f_{\overline{\lambda}}$,

$$\mathrm{Re}\langle G,E\rangle = \mathrm{Re}(\alpha\overline{\beta}\gamma)$$

and

$$P_E(G) = \gamma(\alpha u y^* + \overline{\beta} x v^* - \alpha\overline{\beta} u v^*) = \gamma(u,x)\begin{pmatrix} -\alpha\overline{\beta} & \alpha \\ \overline{\beta} & 0 \end{pmatrix}(v,y)^*.$$

A calculation shows that the squared Frobenius norm of this rank-2 matrix equals

$$\|P_E(G)\|_F^2 = |\gamma|^2\big(|\alpha|^2 + |\beta|^2 - |\alpha|^2|\beta|^2\big).$$

We set

$$g_k = \varepsilon\kappa\Big(\|P_E(G)\|_F^2 - \big(\mathrm{Re}\,\langle G,E\rangle\big)^2\Big) \ge 0$$

for the choice $E = E_k = u_k v_k^*$, $G = G_\varepsilon(E_k) = 2f_{\overline{\lambda}}(\lambda_k,\overline{\lambda_k})x_k y_k^*$, and $\kappa = \kappa_k = 1/(x_k^* y_k)$. In view of the above formulas g_k is computed simply as

$$g_k = \varepsilon\kappa_k\Big(|\gamma_k|^2\big(|\alpha_k|^2 + |\beta_k|^2 - |\alpha_k|^2|\beta_k|^2\big) - \mathrm{Re}(\alpha_k\overline{\beta}_k\gamma_k)^2\Big). \qquad (1.35) \qquad \boxed{\texttt{g-n-formula}}$$

Let

$$f_k = f(\lambda_k,\overline{\lambda_k}), \qquad f(h) = f(\lambda(h),\overline{\lambda(h)}).$$

We accept the result of the step with step size h if

$$f(h) < f_k.$$

If for some fixed $\theta > 1$,

$$f(h) \ge f_k - (h/\theta)g_k,$$

then we reduce the step size for the next step to h/θ. If the step size has not been reduced in the previous step, we try for a larger step size. Algorithm 1 describes the step from t_k to $t_{k+1} = t_k + h_k$.

alg_prEul

Algorithm 1: Integration step for the rank-1 constrained gradient system

```
Data: $A, \varepsilon, \theta > 1, u_k \approx u(t_k), v_k \approx v(t_k)$, $h_k$ (proposed step size)
Result: $u_{k+1}, v_{k+1}, h_{k+1}$
begin
    Initialize the step size by the proposed step size, $h = h_k$
    Compute the value $f_k = f(\lambda_k, \overline{\lambda_k})$
    Compute left/right eigenvectors $x_k, y_k$ of $A + \varepsilon u_k v_k^*$ to $\lambda_k$ such that
      $\|x_k\| = \|y_k\| = 1, x_k^* y_k > 0$
    Compute $\alpha_k, \beta_k, \gamma_k$ by (1.31) and $g_k$ by (1.35)
    Initialize $f(h) = f_k$
    while $f(h) \geq f_k$ do
        Compute $u(h), v(h)$ according to (1.32)-(1.34)
        Compute $\lambda(h)$ target eigenvalue of $A + \varepsilon u(h) v(h)^*$
        Compute the value $f(h) = f\big(\lambda(h), \overline{\lambda(h)}\big)$
        if $f(h) \geq f_k$ then
            Reduce the step size, $h := h/\theta$
    if $f(h) \geq f_k - (h/\theta) g_k$ then
        Reduce the step size for the next step, $h_{\text{next}} := h/\theta$
    else if $h = h_k$ then
        Set $h_{\text{next}} := \theta h_k$ (augment the stepsize if no rejection has occurred)
    else
        Set $h_{\text{next}} := h_k$
    Set $h_{k+1} = h_{\text{next}}$, $\lambda_{k+1} = \lambda(h)$, and the starting values for the next step as
      $u_{k+1} = u(h), v_{k+1} = v(h)$
    return
```

Numerical example. An illustration is given in Fig. 9 for $f(\lambda, \overline{\lambda}) = -\frac{1}{2}(\lambda + \overline{\lambda})$, which yields $\gamma = 1$, and A the randomly chosen 6×6 matrix

$$A = \begin{pmatrix} 0.1019 & -0.8350 & 0.2966 & -0.0756 & -2.2079 & 1.2682 \\ 1.1813 & -1.4224 & -0.8664 & 0.8003 & -1.3413 & 1.3547 \\ -1.2457 & -0.1737 & -1.1910 & -0.3194 & -0.2909 & 0.8230 \\ -0.7830 & -0.5115 & -0.0109 & 0.8860 & 0.4878 & 0.3246 \\ -0.5740 & 0.0268 & 1.1950 & -0.1729 & 0.9966 & -0.8003 \\ -0.3815 & -0.4476 & -0.9740 & 1.4030 & 1.0361 & 0.7399 \end{pmatrix}. \quad (1.36)$$

eq:example

The initial step size is set to $h = 0.1$. The orange curve is the boundary of the ε-pseudospectrum of A, that is

$$\{z \in \mathbb{C} : \ z \text{ is an eigenvalue of } A + \varepsilon E \text{ for some } E \in \mathbb{C}^{n,n} \text{ with } \|E\| = \varepsilon\},$$

and with our choice of f we aim to find a rightmost point of the ε-pseudospectrum (this problem will be discussed in more detail in Chapter III).

A comparison of the 12 iterates (i.e. accepted steps) generated by the splitting integrator (in blue) and an exponential Euler method (in red) shows that the splitting integrator is faster and more accurate after the same number of steps.

II.2 Real case

sec:proto-real

II.2.1 Problem description.

We now consider problem (1.1) for a *real* matrix $A \in \mathbb{R}^{n,n}$ and *real* perturbations $\Delta \in \mathbb{R}^{n,n}$: find

$$\arg \min_{\Delta \in \mathbb{R}^{n,n}, \|\Delta\|_F = \varepsilon} f\left(\lambda\left(A+\Delta\right), \overline{\lambda}\left(A+\Delta\right)\right) \tag{2.1}$$

eq:optimiz0rF

where again $\lambda(A+\Delta)$ is the target eigenvalue of the perturbed matrix $A+\Delta$, and f satisfies (1.2). As in the complex case, it is convenient to write

$$\Delta = \varepsilon E \quad \text{with } \; \|E\|_F = 1$$

and to use the notation

$$F_\varepsilon(E) = f\left(\lambda\left(A+\varepsilon E\right), \overline{\lambda}\left(A+\varepsilon E\right)\right) \tag{2.2}$$

so that (2.1) can be rewritten as

$$\arg \min_{E \in \mathbb{R}^{n,n}, \|E\|_F = 1} F_\varepsilon(E). \tag{2.3}$$

eq:optimizrF

II.2.2 Norm-constrained gradient flow and rank of optimizers

dient-flow-real

The programme of the previous section extends to the real case with minor but important modifications.

Free gradient. Consider a smooth path of *real* matrices $E(t) \in \mathbb{R}^{n,n}$. Since $\dot{E}(t)$ is then also a real matrix, we have by Lemma 1.1

$$\frac{1}{\varepsilon \kappa(t)} \frac{d}{dt} F_\varepsilon(E(t)) = \left\langle G_\varepsilon^{\mathbb{R}}(E(t)), \dot{E}(t) \right\rangle \tag{2.4}$$

eq:deriv-real

with the rescaled real gradient

$$G_\varepsilon^{\mathbb{R}}(E) := \operatorname{Re} G_\varepsilon(E) = \operatorname{Re}(2 f_{\overline{\lambda}} x y^*) \in \mathbb{R}^{n,n}, \tag{2.5}$$

gradient-real

which has rank at most 2 (as a sum of two rank-1 matrices). As ε is fixed and only the real case is considered in this section, we often write for short

$$G(E) := G_\varepsilon^{\mathbb{R}}(E).$$

fig:splitt

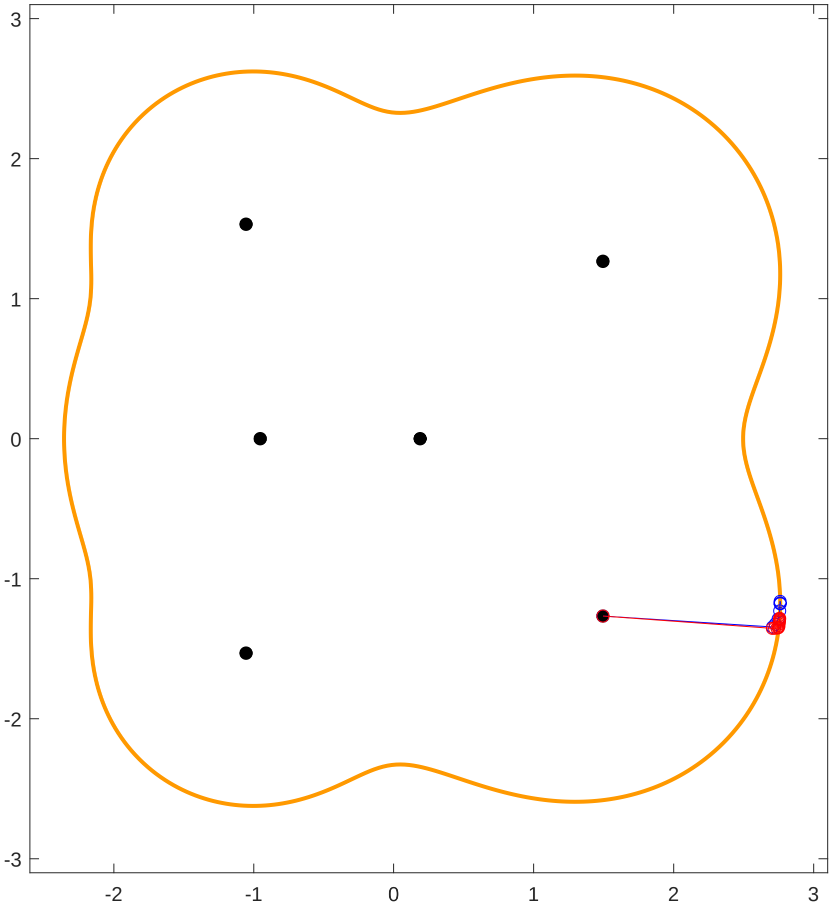

Fig. 1.1. In blue: iterates λ_k of the splitting integrator (Algorithm 1) applied to the matrix A of (III.1.36) with $f(\lambda, \overline{\lambda}) = \mathrm{Re}\lambda$ and $\varepsilon = 1$. In red: iterates of the exponential Euler method.

fig:expeul

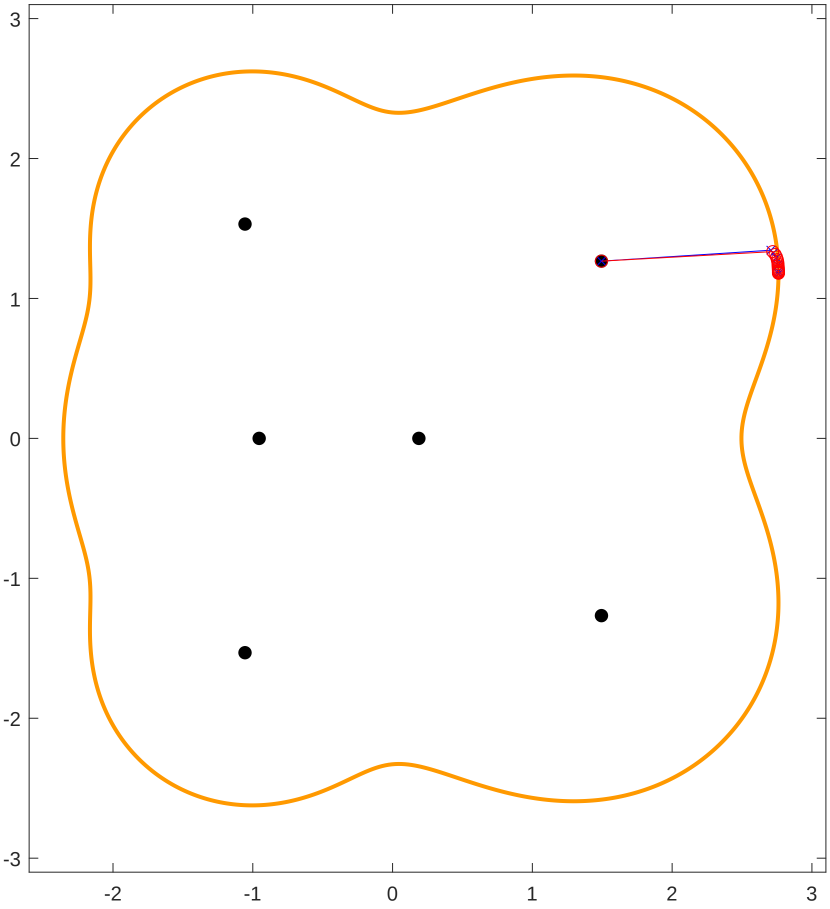

Fig. 1.2. Zoom of the iterates produced by the two methods. The splitting method clearly reaches the equilibrium faster.

Lemma 1.3 on the direction of steepest admissible descent extends without ado to the real case: Consider real matrices and take everywhere the real inner product instead of the real part of the complex inner product.

Constrained gradient flow. We consider the gradient flow on the manifold of *real* $n \times n$ matrices of unit Frobenius norm,

$$\dot{E} = -G(E) + \langle G(E), E \rangle E. \tag{2.6}$$

ode-ErF

Monotonicity. Assuming simple eigenvalues along the trajectory, we then still have the monotonicity property of Theorem 1.4,

$$\frac{d}{dt} F_\varepsilon(E(t)) = -\| G(E) - \langle G(E), E \rangle E \|_F^2 \le 0, \tag{2.7}$$

eq:pos-real

with essentially the same proof (the real inner product replaces the real part of the complex inner product).

Stationary points. Also the characterization of stationary points as given in Theorem 1.5 extends with the same proof: Let $E \in \mathbb{R}^{n,n}$ with $\|E\|_F = 1$ be such that the eigenvalue $\lambda(A + \varepsilon E)$ is simple and $G_\varepsilon^{\mathbb{R}}(E) \neq 0$. Then,

$$\begin{array}{l} E \text{ is a stationary point of the differential equation (2.6)} \\ \text{if and only if } E \text{ is a real multiple of } G_\varepsilon^{\mathbb{R}}(E). \end{array} \tag{2.8}$$

stat-real

stat-real

As a consequence, optimizers of (2.3) have rank at most 2. We can determine the precise rank as follows.

thm:rank

Theorem 2.1 (Rank of optimizers). *For $A \in \mathbb{R}^{n,n}$ and $\varepsilon > 0$, let $E \in \mathbb{R}^{n,n}$ with $\|E\|_F = 1$ be a stationary point of the differential equation* (2.6) *such that the eigenvalue $\lambda = \lambda(A + \varepsilon E)$ is simple and $G_\varepsilon^{\mathbb{R}}(E) \neq 0$. Then, the rank of E is as follows.*

(a) If λ is real, then E has rank 1.
(b) If $\operatorname{Im} \lambda \neq 0$, *then E has rank* 2.

Proof. (a) If λ is real, then f_λ is real and x and y can be chosen real, hence $G(E) = \operatorname{Re}(\overline{f}_\lambda x y^*) = f_\lambda x y^\top$ is of rank 1, and so is every nonzero real multiple, in particular E.

(b) We set $w = f_\lambda y$ and separate the real and imaginary parts in $x = x_R + \mathrm{i} x_I$ and $w = w_R + \mathrm{i} w_I$. If $\operatorname{Re}(x w^*) = x_R w_R^\top + x_I w_I^\top$ is of rank 1, then x_R and x_I are linearly dependent or w_R and w_I are linearly dependent. Let us first assume the former. In this case there is a real α such that $x = \cos(\alpha) v + \mathrm{i} \sin(\alpha) v$ for some nonzero real vector v. Rotating both x and y by $\mathrm{e}^{-\mathrm{i}\alpha}$ does not change the required property $x^* y > 0$, so we can assume without loss of generality that x is a real left eigenvector of the real matrix $A + \varepsilon E$, which implies that the corresponding eigenvalue λ is real. The argument is analogous when w_R and w_I are linearly dependent. □

II.2.3 Low-rank matrices and their tangent matrices

subsec:low-rank

Theorem 2.1 motivates us to search for a differential equation on the manifold of real rank-2 matrices that has the same stationary points with non-real target eigenvalue as (2.6). For the stationary points with real target eigenvalues we use a differential equation for rank-1 matrices as in Section II.1.7.

In the following we consider differential equations of the manifold of real $n \times n$ matrices of rank r, denoted

$$\mathcal{M}_r = \mathcal{M}_r(\mathbb{R}^{n,n}) = \{E \in \mathbb{R}^{n,n} \,:\, \mathrm{rank}(E) = r\}.$$

While only ranks 1 and 2 are of interest for the optimization problem studied in this section, we now consider the case of a general fixed rank r, since it is not more complicated than rank 2 and will be useful later. We proceed similarly to the rank-1 case considered in Section II.1.6.

Every real rank-r matrix E of dimension $n \times n$ can be written in the form

$$E = USV^\top \tag{2.9}$$

USV

where $U \in \mathbb{R}^{n,r}$ and $V \in \mathbb{R}^{n,r}$ have orthonormal columns, i.e.,

$$U^\top U = I_r, \qquad V^\top V = I_r, \tag{2.10}$$

UV-orth

(with the identity matrix I_r of dimension r), and $S \in \mathbb{R}^{r,r}$ is nonsingular. The singular value decomposition yields S diagonal, but here we will not assume a special form of S. The representation (2.9) is not unique: replacing U by $\widetilde{U} = UP$ and V by $\widetilde{V} = VQ$ with orthogonal matrices $P, Q \in \mathbb{R}^{r,r}$, and correspondingly S by $\widetilde{S} = P^\top SQ$, yields the same matrix $E = USV^\top = \widetilde{U}\widetilde{S}\widetilde{V}^\top$.

Every tangent matrix $\dot{E} \in T_E\mathcal{M}_r$ is of the form

$$\dot{E} = \dot{U}SV^\top + U\dot{S}V^\top + US\dot{V}^\top, \tag{2.11}$$

Edot-r

where $\dot{S} \in \mathbb{R}^{r,r}$, and $U^\top \dot{U}$ and $V^\top \dot{V}$ are skew-symmetric (as results from differentiating $U^\top U = I_r$ and $V^\top V = I_r$). The matrices $\dot{S}, \dot{U}, \dot{V}$ are uniquely determined by $\dot{E}$ and U, S, V if we impose the orthogonality conditions

$$U^\top \dot{U} = 0, \qquad V^\top \dot{V} = 0. \tag{2.12}$$

orth

Multiplying $\dot{E}$ with $U^\top$ from the left and with V from the right, we then obtain

$$\dot{S} = U^\top \dot{E} V, \quad \dot{U}S = \dot{E}V - U\dot{S}, \quad S\dot{V}^\top = U^\top \dot{E} - \dot{S}V^\top,$$

which yields $\dot{S}, \dot{U}, \dot{V}$ in terms of $\dot{E}$. Extending this construction, we arrive at an explicit formula for the orthogonal projection onto the tangent space. Here, orthogonality refers to the real Frobenius inner product.

em:P-formula-r

Lemma 2.2 (Rank-r tangent space projection). *The orthogonal projection from $\mathbb{R}^{n,n}$ onto the tangent space $T_E\mathcal{M}_r$ at $E = USV^\top \in \mathcal{M}_r$ is given by*

$$P_E(Z) = Z - (I - UU^\top)Z(I - VV^\top) \quad \textit{for } Z \in \mathbb{C}^{n,n}. \tag{2.13}$$

P-formula-r

Proof. The proof is a direct extension of the proof of Lemma 1.11. Let $P_E(Z)$ be defined by (2.13). Determining (similarly to above) $\dot{S}, \dot{U}, \dot{V}$ by

$$\dot{S} = U^\top ZV, \quad \dot{U}S = ZV - U\dot{S}, \quad S\dot{V}^\top = U^\top Z - \dot{S}V^\top,$$

we obtain $P_E(Z) = UU^\top ZVV^\top - ZVV^\top - UU^\top Z$ in the form (2.11) with (2.12) and hence

$$P_E(Z) \in T_E\mathcal{M}_r.$$

Furthermore,

$$\langle P_E(Z), \dot{E}\rangle = \langle Z, \dot{E}\rangle \qquad \text{for all } \dot{E} \in T_E\mathcal{M}_r,$$

because $\langle (I - UU^\top)Z(I - VV^\top), \dot{E}\rangle = \langle Z, (I - UU^\top)\dot{E}(I - VV^\top)\rangle = 0$ by (2.11). Hence, $P_E(Z)$ is the orthogonal projection of Z onto $T_E\mathcal{M}_r$. □

We note that $P_E(E) = E$ for $E \in \mathcal{M}_r$, or equivalently, $E \in T_E\mathcal{M}_r$.

II.2.4 Rank-r constrained gradient flow

-gradient-flow

In the differential equation (2.6) we replace the right-hand side by its orthogonal projection onto $T_E\mathcal{M}_r$, so that solutions starting with rank r will retain rank r for all times:

$$\dot{E} = P_E\Big(-G(E) + \langle G(E), E\rangle E\Big). \tag{2.14}$$

ode-ErF-2

Since $E \in T_E\mathcal{M}_r$, we have $P_E(E) = E$ and $\langle E, Z\rangle = \langle E, P_E(Z)\rangle$, and hence the differential equation can be rewritten as

$$\dot{E} = -P_E(G(E)) + \langle E, P_E(G(E))\rangle E, \tag{2.15}$$

ode-ErF-2-v2

which differs from (2.6) only in that $G(E)$ is replaced by its orthogonal projection to $T_E\mathcal{M}_r$. This shows that $\langle E, \dot{E}\rangle = 0$, so that the unit norm of E is conserved along solutions of (2.14).

To obtain the differential equation in a form that uses the factors in $E = USV^\top$ rather than the full $n \times n$ matrix E, we use the following result, which follows directly from the proof of Lemma 2.2.

lem:USV

Lemma 2.3 (Differential equations for the factors). *For $E = USV^\top \in \mathcal{M}_r$ with nonsingular $S \in \mathbb{R}^{r,r}$ and with $U \in \mathbb{R}^{n,r}$ and $V \in \mathbb{R}^{n,r}$ having orthonormal columns, the equation $\dot{E} = P_E(Z)$ is equivalent to $\dot{E} = \dot{U}SV^\top + U\dot{S}V^\top + US\dot{V}^\top$, where*

$$\begin{aligned} \dot{S} &= U^\top ZV \\ \dot{U} &= (I - UU^\top)ZVS^{-1} \\ \dot{V} &= (I - VV^\top)Z^\top US^{-\top}. \end{aligned} \tag{2.16}$$

With $Z = -G(E) + \langle G(E), E\rangle E$ and $r = 2$, this yields that the differential equation (2.14) for $E = USV^\top$ is equivalent to a system of differential equations for S, U, V. On the right-hand side appears the inverse of the matrix S, which may be ill-conditioned. In the present context, this appears when the target eigenvalue gets close to the real axis (see Theorem 2.1) so that E, and hence S, becomes almost of rank 1. In such a situation of a small singular value in S, standard numerical integrators become unstable or yield plainly wrong numerical solutions unless used with a tiny stepsize proportional to the smallest nonzero singular value. Later in this section we will describe a numerical integrator that is robust to small singular values.

Monotonicity. Assuming simple eigenvalues almost everywhere along the trajectory, we still have the monotonicity property of Theorem 1.14 along solutions $E(t)$ of (2.14),

$$\frac{d}{dt}F_\varepsilon(E(t)) = -\|P_E(G(E)) - \langle E, P_E(G(E))\rangle E\|_F^2 \le 0, \tag{2.17}$$

eq:pos-real-r

with essentially the same proof (the real inner product replaces the real part of the complex inner product).

Stationary points. Comparing the differential equations (2.6) and (2.14) immediately shows that every stationary point of (4.13) is also a stationary point of the projected differential equation (2.14). As in Theorem 1.15, the converse is also true for the stationary points E of unit Frobenius norm with $P_E(G(E)) \neq 0$. This shows that the low-rank projection does not create spurious stationary points.

thm:stat-r

Theorem 2.4 (Stationary points). *Let $E \in \mathcal{M}_2$ be of unit Frobenius norm and assume that $P_E(G(E)) \neq 0$. If E is a stationary point of the projected differential equation* (2.14)*, then E is already a stationary point of the differential equation* (2.6)*.*

Proof. The proof extends the proof of Theorem 1.15. We show that E is a real multiple of $G_\varepsilon(E)$. By (2.8), E is then a stationary point of the differential equation (4.13).

For a stationary point E of (2.14), we must have equality in (1.28), which shows that $P_E(G)$ (with $G = G(E)$ for short) is a nonzero real multiple of E. Hence, in view of $P_E(E) = E$, we can write G as

$$G = \mu E + W, \qquad \text{where } \mu \neq 0 \text{ is real and } P_E(W) = 0.$$

With $E = USV^\top$ as above, we then have

$$W = W - P_E(W) = (I - UU^\top)W(I - VV^\top).$$

Since G is of rank at most 2, it can be written in the form $G = XRY^\top$, where $X, Y \in \mathbb{R}^{n,2}$ have orthonormal columns and $R \in \mathbb{R}^{2,2}$. So we have

$$XRY^\top = \mu USV^\top + (I - UU^\top)W(I - VV^\top).$$

Multiplying from the right with V yields $X(RY^\top V) = \mu US$, which shows that X has the same range as U, and multiplying from the left with $U^\top$ yields that Y has the same range as V. Hence, G has the same range and corange as E, which implies that $P_E(G) = G$. Since we already know that $P_E(G)$ is a nonzero real multiple of $P_E(E) = E$, it follows that G is the same real multiple of E. By (2.8), E is therefore a stationary point of (2.6). □

As in Remark 1.16, it is shown that if $G(E)$ is of rank 2 and $P_E(G(E)) = 0$, then it follows that $Ey = 0$ and $x^*E = 0$, which implies that λ is already an eigenvalue of the unperturbed matrix A with the same eigenvectors x and y, which is a very exceptional situation.

II.2.5 Time-stepping for the low-rank differential equation

rank-integrator

A robust integrator. The following method adapts the low-rank integrator of Ceruti & Lubich (2021) to the norm-constrained situation considered here. It first updates the basis matrices U and V with orthonormal columns in parallel and then uses a Galerkin approximation to the differential equation (2.14) in the updated basis. This integrator has been shown to be robust to the presence of small singular values, which would here appear in the case of a target eigenvalue near the real axis.

One time step of integration from time t_k to $t_{k+1} = t_k + h$ starting from a factored rank-r matrix $E_k = U_k S_k V_k^\top$ of unit Frobenius norm computes an updated rank-r factorization $E_{k+1} = U_{k+1} S_{k+1} V_{k+1}^\top$ of unit Frobenius norm as follows.

1. Update the basis matrices $U_k \to U_{k+1}$ and $V_k \to V_{k+1}$:

   Integrate from $t = t_k$ to $t_{k+1} = t_k + h$ the $n \times r$ matrix differential equation

   $$\dot{K}(t) = -G(K(t)V_k^\top)V_k, \qquad K(t_k) = U_k S_k.$$

   Perform a QR factorization $K(t_{k+1}) = U_{k+1} R_{k+1}$ and compute the $r \times r$ matrix $M = U_{k+1}^\top U_k$.

   Integrate from $t = t_k$ to t_{k+1} the $n \times r$ matrix differential equation

   $$\dot{L}(t) = -G(U_k L(t)^\top)^\top U_k, \qquad L(t_k) = V_k S_k^\top.$$

   Perform a QR factorization $L(t_{k+1}) = V_{k+1} \widetilde{R}_{k+1}$ and compute the $r \times r$ matrix $N = V_{k+1}^\top V_k$.

2. Update $S_k \to S_{k+1}$:

   Integrate from $t = t_k$ to t_{k+1} the $r \times r$ matrix differential equation

   $$\dot{S}(t) = -U_{k+1}^\top G(U_{k+1} S(t) V_{k+1}^\top) V_{k+1}, \qquad S(t_k) = \frac{M S_k N^\top}{\| M S_k N^\top \|_F},$$

   and set $S_{k+1} = S(t_{k+1}) / \| S(t_{k+1}) \|_F$.

The differential equations in the substeps are solved approximately by a step of some standard numerical integrator, e.g. the explicit Euler method or a low-order explicit Runge–Kutta method such as the second-order Heun method. We denote the result of the fully discrete method with stepsize h as $U(h)$, $V(h)$, $S(h)$.

Stepsize selection. As in Section II.1.8, the stepsize selection is done by an Armijo-type strategy. Along solutions $E(t) \in \mathcal{M}_2$ of (2.14), we have by (2.4)–(2.5)

$$\frac{d}{dt}F_\varepsilon(E(t)) = -g(E(t)) \quad \text{with} \quad g(E) = \varepsilon\kappa\big(\|P_E(G)\|_F^2 - \langle G, E\rangle^2\big) \geq 0$$

where $\kappa = 1/(x^*y) > 0$, $G = G_\varepsilon^{\mathbb{R}}(E) = \mathrm{Re}(2f_{\overline{\lambda}}xy^*)$ with the normalized left and right eigenvectors x and y to the eigenvalue $\lambda(A + \varepsilon E)$.

We note that on separating real and imaginary parts in $2f_{\overline{\lambda}}x = w_R + \mathrm{i}w_I$ and $y = y_R + \mathrm{i}y_I$ and defining the $n \times 2$ real matrices $W = (w_R, w_I)$ and $Y = (y_R, y_I)$, we have the real factorization

$$G = WY^\top.$$

With the rank-2 matrix $E = USV^\top$ in factorized from as above, we can then compute $g(E)$ without actually forming the $n \times n$ matrices E and G: noting that with the 2×2 matrices $P = U^\top W$ and $Q = V^\top Y$ we have

$$\langle G, E\rangle = \langle WY^\top, USV^\top\rangle = \langle PQ^\top, S\rangle_{\mathbb{R}^{2\times 2}}$$

and

$$P_E(G) = UPY^\top - UPQ^\top V^\top + WQ^\top V^\top,$$

which yields after a straightforward computation

$$\|P_E(G)\|_F^2 = \|PY^\top\|_F^2 + \|WQ^\top\|_F^2 - \|PQ^\top\|_F^2,$$

so that finally $g = g(E)$ is given by

$$g = \varepsilon\kappa\big(\|PY^\top\|_F^2 + \|WQ^\top\|_F^2 - \|PQ^\top\|_F^2 - \langle PQ^\top, S\rangle^2\big). \qquad (2.18) \qquad \boxed{\texttt{g-n-formula-r}}$$

With this quantity g, the Armijo-type stepsize selection is then done as in Section II.1.8. A time step of the method is summarized in Algorithm 2.

alg:real

Algorithm 2: Integration step for the rank-2 constrained real gradient system

```
Data: $A, \varepsilon, \theta > 1, U_k \approx U(t_k), V_k \approx V(t_k) \in \mathbb{R}^{n \times 2}$ with orthonormal columns,
      $S_k \approx S(t_k) \in \mathbb{R}^{2 \times 2}$ of unit Frobenius norm, target eigenvalue
      $\lambda_k = \lambda(A + \varepsilon U_k S_k V_k^\top)$, $h_k$ (proposed step size)
Result: $U_{k+1}, V_{k+1}, S_{k+1}, \lambda_{k+1}, h_{k+1}$
begin
    Initialize the step size by the proposed step size, $h = h_k$
    Compute $f_k = f(\lambda_k, \overline{\lambda_k})$
    Compute the left/right eigenvectors $x_k, y_k$ to $\lambda_k$ such that
      $\|x_k\| = \|y_k\| = 1, x_k^* y_k > 0$
    Compute $g_k$ by (2.18)
    Initialize $f(h) = f_k$
    while $f(h) \geq f_k$ do
        Compute $U(h), V(h), S(h)$ by the above rank-2 integrator
        Compute $\lambda(h)$ target eigenvalue of $A + \varepsilon U(h) S(h) V(h)^\top$
        Compute $f(h) = f\big(\lambda(h), \overline{\lambda(h)}\big)$
        if $f(h) \geq f_k$ then
            Reduce the step size, $h := h/\theta$
    Initialize $h_{\text{next}} = h$
    if $f(h) \geq f_k - (h/\theta) g_k$ then
        Reduce the step size for the next step, $h_{\text{next}} := h/\theta$
    if $h_{\text{next}} = h_k$ then
        Compute $U(\theta h), V(\theta h), S(\theta h)$ by the above rank-2 integrator
        Compute $\lambda(\theta h)$ target eigenvalue of $A + \varepsilon U(\theta h) S(\theta h) V(\theta h)^\top$
        Compute $f(\theta h) = f\big(\lambda(\theta h), \overline{\lambda(\theta h)}\big)$
        if $f(h) > f(\theta h)$ then
            Enlarge the step size for the next step, $h := \theta h$ and then $h_{\text{next}} := h$
    Set $h_{k+1} = h_{\text{next}}$, $\lambda_{k+1} = \lambda(h)$, and the starting values for the next step as
      $U_{k+1} = U(h)$, $V_{k+1} = V(h)$, $S_{k+1} = S(h)$
    return
```

II.3 Structured cases

toto-structured

II.3.1 Problem description. Linear structures

Let $\mathcal{S}$ be a subspace of the vector space of complex or real $n \times n$ matrices, e.g. a space of matrices with a prescribed sparsity pattern, or Toeplitz matrices, or Hamiltonian matrices, etc.

As before, we set

$$F_\varepsilon(E) = f\left(\lambda\left(A + \varepsilon E\right), \overline{\lambda}\left(A + \varepsilon E\right)\right). \tag{3.1}$$

eq:optimiz0S

We now restrict the admissible perturbations εE to be in $\mathcal{S}$ and consider the *structured* eigenvalue optimization problem to find

$$\arg \min_{E \in \mathcal{S}, \|E\|_F = 1} F_\varepsilon(E). \tag{3.2}$$

eq:optimizS

II.3.2 Projection onto the structure

proj-structure

Let $\Pi^{\mathcal{S}}$ be the orthogonal projection (w.r.t. the Frobenius inner product) onto $\mathcal{S}$: for every $Z \in \mathbb{C}^{n,n}$,

$$\Pi^{\mathcal{S}} Z \in \mathcal{S} \quad \text{and} \quad \mathrm{Re}\langle \Pi^{\mathcal{S}} Z, W \rangle = \mathrm{Re}\langle Z, W \rangle \quad \forall W \in \mathcal{S}. \tag{3.3}$$

Pi-S

For a complex-linear subspace $\mathcal{S}$, taking the real part of the complex inner product can be omitted (because with $W \in \mathcal{S}$, then also $\mathrm{i}W \in \mathcal{S}$), but taking the real part is needed for real-linear subspaces. Note that for $\mathcal{S} = \mathbb{R}^{n,n}$, we then have $\Pi^{\mathcal{S}} Z = \mathrm{Re}\, Z$ for all $Z \in \mathbb{C}^{n,n}$. In the following examples, the stated action of $\Pi^{\mathcal{S}}$ is readily verified.

Example 3.1 (Sparse matrices). If $\mathcal{S}$ is the space of complex matrices with a prescribed sparsity pattern, then $\Pi^{\mathcal{S}} Z$ leaves the entries of Z on the sparsity pattern unchanged and annihilates those outside the sparsity pattern.

Example 3.2 (Matrices with prescribed range and co-range). An example of particular interest in control theory is the perturbation space

$$\mathcal{S} = \{ B \Delta C \, : \, \Delta \in \mathbb{R}^{k,l} \},$$

where $B \in \mathbb{R}^{n,k}$ and $C \in \mathbb{R}^{l,n}$ with $k, l < n$ are given matrices of full rank. Here, $\Pi^{\mathcal{S}} Z = B B^\dagger Z C^\dagger C$, where $B^\dagger$ and $C^\dagger$ are the Moore–Penrose inverses of B and C, respectively.

Example 3.3 (Toeplitz matrices). If $\mathcal{S}$ is the space of complex $n \times n$ Toeplitz matrices, then $\Pi^{\mathcal{S}} Z$ is obtained by replacing in each diagonal all the entries of Z by their arithmetic mean. For real Toeplitz matrices, the same action is done on ReZ.

Example 3.4 (Hamiltonian matrices). If $\mathcal{S}$ is the space of $2d \times 2d$ real Hamiltonian matrices, then $\Pi^{\mathcal{S}} Z = J^{-1} \mathrm{Sym}(\mathrm{Re}(JZ))$, where $\mathrm{Sym}(\cdot)$ takes the symmetric part of a matrix and

$$J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix},$$

for which $J^{-1} = J^\top = -J$. We recall that a real matrix A is Hamiltonian if JA is symmetric.

II.3.3 Structure- and norm-constrained gradient flow

gradient-flow-S

The programme of Section II.1 extends to structured cases as follows.

Structured gradient. Consider a smooth path of *structured* matrices $E(t) \in \mathcal{S}$. Since then also $\dot{E}(t) \in \mathcal{S}$, we have by Lemma 1.1

$$\frac{1}{\varepsilon \kappa(t)} \frac{d}{dt} F_\varepsilon(E(t)) = \mathrm{Re}\big\langle G_\varepsilon^{\mathcal{S}}(E(t)), \dot{E}(t) \big\rangle \tag{3.4}$$

eq:deriv-S

with the rescaled structured gradient

$$G_\varepsilon^{\mathcal{S}}(E) := \mathit{\Pi}^{\mathcal{S}}\, G_\varepsilon(E) = \mathit{\Pi}^{\mathcal{S}}(2 f_{\overline{\lambda}}\, xy^*) \in \mathcal{S}, \tag{3.5}$$

gradient-S

where x, y are the left and right eigenvectors, normalized to unit norm and with positive inner product, associated with a simple eigenvalue λ of $A+\varepsilon E$, and $f_{\overline{\lambda}} = (\partial f/\partial \overline{\lambda})(\lambda, \overline{\lambda})$.

We note that $G_\varepsilon^{\mathcal{S}}(E)$ is the orthogonal projection onto $\mathcal{S}$ of a rank-1 matrix.

Lemma 1.3 on the direction of steepest admissible descent extends immediately to the structured case: If $E, G \in \mathcal{S}$ in Lemma 1.3, then also $Z_\star$ of (1.11) is in $\mathcal{S}$.

zero-gradient-S

Lemma 3.5 (Non-vanishing structured gradient). *Let $A, E \in \mathcal{S}$ and $\varepsilon > 0$, and let λ be a simple target eigenvalue of $A + \varepsilon E$.*

(i) Complex case: $\mathcal{S}$ is a complex-linear subspace of $\mathbb{C}^{n,n}$. Then,

$$G_\varepsilon^{\mathcal{S}}(E) \neq 0 \quad \textit{if} \quad \overline{\lambda} f_{\overline{\lambda}} \neq 0.$$

(ii) Real case: $\mathcal{S}$ is a real-linear subspace of $\mathbb{R}^{n,n}$. Then,

$$G_\varepsilon^{\mathcal{S}}(E) \neq 0 \quad \textit{if} \quad \mathrm{Re}(\overline{\lambda} f_{\overline{\lambda}}) \neq 0.$$

We emphasize that also A needs to be in $\mathcal{S}$. The result does not hold true when $A \notin \mathcal{S}$.

Proof. We give the proof for the real case. The complex case is analogous but slightly simpler. We take the real inner product of $G_\varepsilon^{\mathcal{S}}(E)$ with $A + \varepsilon E \in \mathcal{S}$ and use the definition (3.5) of $G_\varepsilon^{\mathcal{S}}(E)$:

$$\begin{aligned} &\langle G_\varepsilon^{\mathcal{S}}(E), A + \varepsilon E \rangle = \mathrm{Re}\langle \mathit{\Pi}^{\mathcal{S}}(2 f_{\overline{\lambda}}\, xy^*), A + \varepsilon E \rangle = \mathrm{Re}\langle 2 f_{\overline{\lambda}}\, xy^*, A + \varepsilon E \rangle \\ &\quad = \mathrm{Re}\big(2 f_\lambda\, x^*(A + \varepsilon E) y\big) = \mathrm{Re}\big(2 f_\lambda \lambda\, x^* y\big) = 2\,\mathrm{Re}\big(f_{\overline{\lambda}} \overline{\lambda}\big)\,(x^* y), \end{aligned}$$

where $x^* y > 0$. This yields the result. □

If the identity matrix I is in $\mathcal{S}$, then the condition for $G_\varepsilon^{\mathcal{S}}(E) \neq 0$ can be weakened:

– In the complex case, it then suffices to have $f_{\overline{\lambda}} \neq 0$. This is seen by taking the inner product with $A + \varepsilon E - \mu I \in \mathcal{S}$ for an arbitrary $\mu \in \mathbb{C}$.

– In the real case, if λ is real, then it suffices to have Re $f_{\overline{\lambda}} \neq 0$. If λ is non-real, then it even suffices to have $f_{\overline{\lambda}} \neq 0$. In both cases this is seen by taking the inner product with $A + \varepsilon E - \mu I \in \mathcal{S}$ for an arbitrary $\mu \in \mathbb{R}$.

Constrained gradient flow. We consider the gradient flow on the manifold of *structured* $n \times n$ matrices in $\mathcal{S}$ of unit Frobenius norm,

$$\dot{E} = -G_\varepsilon^{\mathcal{S}}(E) + \mathrm{Re}\langle G_\varepsilon^{\mathcal{S}}(E), E \rangle E. \tag{3.6}$$

ode-E-S

Monotonicity. Assuming simple eigenvalues along the trajectory, we then still have the monotonicity property of Theorem 1.4,

$$\frac{d}{dt} F_\varepsilon(E(t)) \le 0, \tag{3.7}$$

eq:pos-S

with essentially the same proof.

Stationary points. Also the characterization of stationary points as given in Theorem 1.5 extends with the same proof: Let $E \in \mathcal{S}$ with $\|E\|_F = 1$ be such that the eigenvalue $\lambda(A + \varepsilon E)$ is simple and $G_\varepsilon^{\mathcal{S}}(E) \neq 0$. Then,

$$\begin{array}{l} E \text{ is a stationary point of the differential equation (3.6)} \quad \boxed{\texttt{stat-S}} \\ \text{if and only if } E \text{ is a real multiple of } G_\varepsilon^{\mathcal{S}}(E). \end{array} \tag{3.8}$$

stat-S

As a consequence, optimizers of (3.2) are projections onto $\mathcal{S}$ of rank-1 matrices. This motivates us to search for a differential equation on the manifold of rank-1 matrices that leads to the same stationary points.

II.3.4 Rank-1 matrix differential equation

Solutions of (3.6) can be written as $E(t) = \Pi^{\mathcal{S}} Z(t)$ where $Z(t)$ solves

$$\dot Z = -G_\varepsilon(E) + \mathrm{Re}\langle G_\varepsilon(E), E\rangle Z \quad \text{with } \; E = \Pi^{\mathcal{S}} Z. \tag{3.9}$$

ode-E-S-Z

We note that $\mathrm{Re}\langle E, \dot E\rangle = 0$ if $\|E\|_F = 1$, so that the unit Frobenius norm of $E(t)$ is conserved for all t. As every solution tends to a stationary point of rank 1, we project the right-hand side onto the tangent space $T_Y\mathcal{M}_1$ at Y of the manifold of complex rank-1 matrices $\mathcal{M}_1 = \mathcal{M}_1(\mathbb{C}^{n,n})$ and consider instead the projected differential equation with solutions of rank 1:

$$\dot Y = -P_Y G_\varepsilon(E) + \mathrm{Re}\langle P_Y G_\varepsilon(E), E\rangle Y \quad \text{with } \; E = \Pi^{\mathcal{S}} Y. \tag{3.10}$$

ode-E-S-1

Note that then

$$\dot E = -\Pi^{\mathcal{S}} P_Y G_\varepsilon(E) + \mathrm{Re}\langle \Pi^{\mathcal{S}} P_Y G_\varepsilon(E), E\rangle E \quad \text{with } \; E = \Pi^{\mathcal{S}} Y, \tag{3.11}$$

ode-E-S-1-Pi

which differs from the gradient flow (3.6) only in that the gradient $G_\varepsilon(E)$ is replaced by the rank-1 projected gradient $P_Y G_\varepsilon(E)$.

For $E = \Pi^{\mathcal{S}} Y$ of unit Frobenius norm,

$$\mathrm{Re}\langle E, \dot E\rangle = \mathrm{Re}\langle E, \dot Y\rangle = -\mathrm{Re}\langle E, P_Y G_\varepsilon(E)\rangle + \mathrm{Re}\langle P_Y G_\varepsilon(E), E\rangle \, \mathrm{Re}\langle E, Y\rangle = 0,$$

where we used that $\mathrm{Re}\langle E, Y\rangle = \mathrm{Re}\langle E, \Pi^{\mathcal{S}} Y\rangle = \mathrm{Re}\langle E, E\rangle = \|E\|_F^2 = 1$. So we have

$$\|E(t)\|_F = 1 \qquad \text{for all } t.$$

We write a rank-1 matrix $Y \in \mathcal{M}_1$ in a non-unique way as

$$Y = \rho u v^*,$$

where $\rho \in \mathbb{R}$, $\rho > 0$ and $u, v \in \mathbb{C}^n$ have unit norm. The following lemma extends Lemma 1.13 to the structured situation. It shows how the rank-1 differential equation (3.10) can be restated in terms of differential equations for the factors u, v and an explicit formula for ρ.

lem:uv-1-S **Lemma 3.6 (Differential equations for the factors).** *Every solution $Y(t) \in \mathcal{M}_1$ of the rank-1 differential equation* (3.10) *with $\|\Pi^{\mathcal{S}} Y(t)\|_F = 1$ can be written as $Y(t) = \rho(t)u(t)v(t)^*$ where $u(t)$ and $v(t)$ of unit norm satisfy the differential equations*

$$\begin{aligned}
\rho\dot{u} &= -\tfrac{\mathrm{i}}{2}\mathrm{Im}(u^*Gv)u - (I - uu^*)Gv,\\
\rho\dot{v} &= -\tfrac{\mathrm{i}}{2}\mathrm{Im}(v^*Gu)v - (I - vv^*)G^*u,
\end{aligned}$$

where $G = G_\varepsilon(E)$ for $E = \Pi^{\mathcal{S}} Y = \rho\, \Pi^{\mathcal{S}}(uv^*)$ and $\rho = 1/\|\Pi^{\mathcal{S}}(uv^*)\|_F$.

We find that with the exception of the additional positive factor ρ on the left-hand side, these differential equations are of the same form as in Lemma 1.13. Note that ρ is only related to the speed with which a trajectory is percursed, but does not affect the trajectory itself. However, here $G = G_\varepsilon(E)$ for a different matrix $E = \Pi^{\mathcal{S}}(\rho uv^*)$ instead of $E = uv^*$ in (1.25).

Proof. The equation for ρ is obvious because $1 = \|E\|_F = \|\Pi^{\mathcal{S}}(\rho uv^*)\|_F = \rho\|\Pi^{\mathcal{S}}(uv^*)\|_F$. We write the right-hand side of (3.10) and use (1.20) to obtain for $Y = \rho uv^*$

$$\begin{aligned}
\dot{Y} &= -P_Y G + \mathrm{Re}\langle P_Y G, E\rangle Y\\
&= -\,(I - uu^*)Gvv^* - uu^*G(I - vv^*) - uu^*Gvv^* + \mathrm{Re}\Big\langle P_Y G, E\Big\rangle Y\\
&= -\,\Big((I - uu^*)Gvv^* + \tfrac{\mathrm{i}}{2}\mathrm{Im}(u^*Gv)u\Big)v^* - u\Big(u^*G(I - vv^*) + \tfrac{\mathrm{i}}{2}\mathrm{Im}(u^*Gv)v^*\Big)\\
&\quad -\Big(\mathrm{Re}(u^*Gv) + \mathrm{Re}\langle P_Y G, E\rangle\rho\Big)uv^*.
\end{aligned}$$

Since this is equal to $\dot{Y} = (\rho\dot{u})v^* + u(\rho\dot{v}^*) + \dot{\rho}uv^*$, we can equate $\rho\dot{u}$, $\rho\dot{v}^*$ and $\dot{\rho}$ with the three terms in big brackets. So we obtain the stated differential equations for u and v (and another one for ρ, which will not be needed). Further we have $(d/dt)\|u\|^2 = 2\,\mathrm{Re}(u^*\dot{u}) = 0$ and analogously for v, which yields that u and v stay of unit norm. □

We note that for $G = G_\varepsilon(E) = 2f_{\overline{\lambda}}\, xy^*$ (see Lemma 1.1) and with $\alpha = u^*x$, $\beta = v^*y$ and $\gamma = 2f_{\overline{\lambda}}$, we obtain differential equations that differ from (1.26) only in the additional factor ρ on the left-hand side:

$$\begin{aligned}
\rho\dot{u} &= \alpha\overline{\beta}\gamma\, u - \overline{\beta}\gamma\, x - \tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\alpha\overline{\beta}\gamma)u\\
\rho\dot{v} &= \overline{\alpha}\beta\overline{\gamma}\, v - \overline{\alpha\gamma}\, y - \tfrac{\mathrm{i}}{2}\,\mathrm{Im}(\overline{\alpha}\beta\overline{\gamma})v.
\end{aligned} \qquad (3.12)$$

ode-uv-short-S

Stationary points. The following theorem states that under some non-degeneracy conditions (see Remark 1.16), the differential equations (3.6) and (3.10) yield the same stationary points.

thm:stat-S

Theorem 3.7 (Relating stationary points). *(a) Let $E \in \mathcal{S}$ of unit Frobenius norm be a stationary point of the gradient system* (3.6) *that satisfies $\Pi^{\mathcal{S}} G_\varepsilon(E) \neq 0$. Then, $E = \Pi^{\mathcal{S}} Y$ for an $Y \in \mathbb{C}^{n,n}$ of rank 1 that is a stationary point of the differential equation* (3.10).

(b) Conversely, let $Y \in \mathbb{C}^{n,n}$ of rank 1 be a stationary point of the differential equation (3.10) *such that $E = \Pi^{\mathcal{S}} Y$ has unit Frobenius norm and $P_Y G_\varepsilon(E) \neq 0$. Then, E is a stationary point of the gradient system* (3.6).

Proof. Let $G = G_\varepsilon(E)$ in this proof for short.

(a) By (4.14), $E = \mu^{-1} \Pi^{\mathcal{S}} G$ for some nonzero real μ. Then, $Y := \mu^{-1} G$ is of rank 1 and we have $E = \Pi^{\mathcal{S}} Y$. We further note that $P_Y G = \mu P_Y Y = \mu Y = G$. We thus have

$$-P_Y G + \mathrm{Re}\langle P_Y G, E\rangle Y = -G + \mathrm{Re}\langle G, E\rangle Y.$$

Here we find that

$$\mathrm{Re}\langle G, E\rangle = \mathrm{Re}\langle \Pi^{\mathcal{S}} G, E\rangle = \mathrm{Re}\langle \mu E, E\rangle = \mu \|E\|_F^2 = \mu.$$

So we have

$$-G + \mathrm{Re}\langle G, E\rangle Y = -G + \mu Y = 0$$

by the definition of Y. This shows that Y is a stationary point of (3.10).

(b) By the argument used in the proof of Theorem 1.15, stationary points $Y \in \mathcal{M}_1$ of the differential equation (3.10) are characterized as real multiples of G. Hence, $E = \Pi^{\mathcal{S}} Y$ is a real multiple of $\Pi^{\mathcal{S}} G$, and by (4.14), $E = \Pi^{\mathcal{S}} Y$ is a stationary point of (3.6). □

Possible loss of global monotonicity and preservation of local monotonicity near stationary points. Since the projections $\Pi^{\mathcal{S}}$ and P_Y do not commute, we cannot guarantee the monotonicity (3.7) along solutions of (3.10). However, in all our numerical experiments we observed a monotonic decrease of the functional in all steps except possibly (and rarely) in the first step. In the following we will explain this monotonic behaviour locally near a stationary point, but we have no theoretical explanation for the numerically observed monotonic behavior far from stationary points.

The first observation, already made in the proof of Theorem 3.7, is that at a stationary point Y of (3.10), we have $P_Y G_\varepsilon(E) = G_\varepsilon(E)$ for $E = \Pi^{\mathcal{S}} Y$. Therefore, close to a stationary point, $P_Y G_\varepsilon(E)$ will be close to $G_\varepsilon(E)$. It turns out that it is even *quadratically* close. This is made more precise in the following lemma.

lem:loc-S

Lemma 3.8 (Projected gradient near a stationary point). *Let $Y_\star \in \mathcal{M}_1$ with $E_\star = \Pi^{\mathcal{S}} Y_\star \in \mathcal{S}$ of unit Frobenius norm. Let $Y_\star$ be a stationary point of the rank-1 projected differential equation* (3.10)*, with an associated target eigenvalue λ of $A + \varepsilon E_\star$ that is simple. Then, there exist $\overline{\delta} > 0$ and a real C such that for all positive $\delta \leq \overline{\delta}$ and all $Y \in \mathcal{M}_1$ with $\|Y - Y_\star\| \leq \delta$ and associated $E = \Pi^{\mathcal{S}} Y$ of unit norm, we have*

$$\|P_Y G_\varepsilon\,(E) - G_\varepsilon\,(E)\,\| \leq C\delta^2. \tag{3.13}$$

Proof. We consider a smooth path $Y(\tau) = u(\tau)v(\tau)^* \in \mathcal{M}_1$ (with $u(\tau), v(\tau) \in \mathbb{C}^n$) and associated $E(\tau) = \varPi^{\mathcal{S}} Y(\tau)$ of unit Frobenius norm with initial value

$$Y(0) = Y_\star = \mu^{-1} G_\star \quad \text{for some nonzero real } \mu \quad \text{and}$$
$$G_\star = G_\varepsilon\,(E_\star)) = 2\overline{f}_\lambda xy^*,$$

where $E_\star = \varPi^{\mathcal{S}} Y_\star$ is of unit Frobenius norm and (λ, x, y) is the target eigentriplet of $A + \varepsilon E_\star$ associated with the target eigenvalue λ.

A direct calculation of the first-order terms in the Taylor expansions of $P_{Y(\tau)} G_\varepsilon(E(\tau))$ and $G_\varepsilon(E(\tau))$, which uses the formula (1.20) for the projection $P_{Y(\tau)}$ and the formula (3.15) of the rescaled gradient $G_\varepsilon(E(\tau))$, surprisingly yields that the Taylor expansions of $P_{Y(\tau)} G_\varepsilon(E(\tau))$ and $G_\varepsilon(E(\tau))$ at $\tau = 0$ coincide up to $O(\tau^2)$. This gives the stated result. □

As a direct consequence of this lemma, a comparison of the differential equations (3.11) and (3.6) yields that δ-close to a stationary point, the functional decreases monotonically along solutions of (3.10) up to $O(\delta^2)$, and even with the same negative derivative as for the gradient flow (3.6) up to $O(\delta^2)$. Note that the derivative of the functional is proportional to $-\delta$ in a δ-neighbourhood of a strong local minimum. Guglielmi, Lubich & Sicilia (**?**) use Lemma 3.8 to prove a result on local convergence as $t \to \infty$ to strong local minima of the functional F_ε of (1.4) for $E(t) = \varPi^{\mathcal{S}} Y(t)$ of unit Frobenius norm associated with solutions $Y(t)$ of the rank-1 differential equation (3.10).

II.3.5 Discrete algorithm and numerical example

Since the differential equations for u and v have essentially the same form as in the unstructured case, the splitting algorithm of Subsection II.1.8 extends in a straightforward way, and also the stepsize selection is readily extended. We refer to Guglielmi, Lubich & Sicilia (**?**) for details.

******** Numerical experiment with a sparse matrix ? **************

II.4 Notes

The review article by Lewis and Overton (1996) remains a basic reference on eigenvalue optimization, including a fascinating account of the history of the subject. There is, however, only a slight overlap of problems and techniques considered here and there.

The book by Absil, Mahony & Sepulchre (2008) on optimization on matrix manifolds discusses alternative gradient-based methods to those considered here, though not specifically for eigenvalue optimization nor low-rank matrix manifolds.

Low-rank property of optimizers. The low-rank structure of optimizers in an eigenvalue optimization problem was first used by Guglielmi & Overton (2011) who devised a rank-1 matrix iteration to compute the complex ε-pseudospectral abscissa and radius; see Section III.2 below.

The approach to eigenvalue optimization via a norm-constrained gradient system and the associated low-rank dynamics was first proposed and studied by Guglielmi & Lubich (2011, 2012, 2013), where it was used to compute the complex and real ε-pseudospectral abscissa and radius as well as sections of the boundary of the ε-pseudospectrum (see Chapter III).

Our discussion of low-rank dynamics in Sections II.1.7 and II.2.4 is based on Koch & Lubich (2007). Numerical integrators for low-rank matrix differential equations that are robust to small singular values are given by the projector-splitting integrator of Lubich & Oseledets (2014) and the basis-update & Galerkin integrator of Ceruti & Lubich (2021), of which a norm-preserving variant is presented in Section II.2.5.

Our presentation of the structured eigenvalue optimization problem and the underlying rank-1 property in Section II.3 follows Guglielmi, Lubich & Sicilia **?**.

Frobenius norm vs. matrix 2-norm. In the approach described in this chapter (and further on in this work), perturbations are measured and constrained in the Frobenius norm. This choice is made because the Frobenius norm, other than the matrix 2-norm, is induced by an inner product, which simplifies many arguments. Not least, it allows us to work with gradient systems. However, the approach taken here with functional-reducing differential equations and their associated low-rank dynamics is relevant also for the matrix 2-norm, in two different ways:

(a) In the general complex case, the optimizers with respect to the Frobenius norm are of rank 1, and so their Frobenius norm equals their 2-norm. Since generally, the 2-norm of a matrix does not exceed its Frobenius norm, it follows that the rank-1 Frobenius-norm optimizers constrained by $\|\varDelta\|_F \le \varepsilon$ are simultaneously the 2-norm optimizers constrained by $\|\varDelta\|_2 \le \varepsilon$.

(b) In the real case and in structured cases, where the Frobenius-norm and 2-norm optimizers are generally different, functional-reducing differential equations for the 2-norm-constrained problem can be given that have very similar properties to the gradient systems for the Frobenius-norm-constrained problem considered here; see Guglielmi & Lubich (2013) for the computation of the 2-norm real pseudospectral abscissa and Guglielmi, Kressner & Lubich (2015) for 2-norm-constrained eigenvalue optimization problems for Hamiltonian matrices.

Chapter III. Pseudospectra

chap:pseudo

III.1 Complex, real and structured pseudospectra

III.1.1 Motivation and definitions

c:ps-motivation

As a motivating example for the (complex, real or structured) ε-pseudospectrum of a matrix A, we consider the linear dynamical system $\dot{x}(t) = Ax(t)$. The system is asymptotically stable, i.e., solutions $x(t)$ converge to zero as $t \to \infty$ for all initial data, if and only if all eigenvalues of A have negative real part. We now ask for the sensitivity of the asymptotic stability under (complex, real or structured) perturbations Δ of norm bounded by a given $\varepsilon > 0$. This clearly depends on the choice of norm, and here we consider the Frobenius norm:

$$\|\cdot\| = \|\cdot\|_F.$$

For a *normal* matrix, the spectral decomposition yields that the perturbed system remains asymptotically stable for an arbitrary complex perturbation of norm at most ε if for each eigenvalue λ of A, the real part is bounded by $\mathrm{Re}\,\lambda + \varepsilon < 0$. This condition is, however, not sufficient for *non-normal* matrices A and it may not be necessary if the class of admissible perturbations is restricted to some subspace $\mathcal{S}$ of structured perturbations.

The question posed is thus: Is the following real number negative?

$$\alpha_\varepsilon^{\mathcal{S}}(A) = \max\{\mathrm{Re}\,\lambda : \text{ There exists } \Delta \in \mathcal{S} \text{ with } \|\Delta\| \le \varepsilon \text{ such that } \lambda \text{ is an eigenvalue of } A + \Delta\}.$$

This question is answered by solving a problem (II.1.1) for $\mathcal{S} = \mathbb{C}^{n,n}$, (II.2.1) for $\mathcal{S} = \mathbb{R}^{n,n}$, or (II.3.1)–(II.3.2) for an arbitrary complex-linear or real-linear subspace $\mathcal{S}$, in each case with the function to be minimized given by $f(\lambda, \overline{\lambda}) = -\frac{1}{2}(\lambda + \overline{\lambda}) = -\mathrm{Re}\,\lambda$ (i.e., we maximize $\mathrm{Re}\,\lambda$).

It is useful to rephrase the question in terms of the ε-pseudospectrum, which is defined as follows.

def:ps

Definition 1.1. The ($\mathcal{S}$-structured) ε*-pseudospectrum* of the matrix A is the set

$$\Lambda_\varepsilon^{\mathcal{S}}(A) = \{\lambda \in \mathbb{C} : \ \lambda \in \Lambda(A + \Delta) \text{ for some } \Delta \in \mathcal{S} \text{ with } \|\Delta\| \le \varepsilon\}, \tag{1.1}$$

eq:epsps

where $\Lambda(M) \subset \mathbb{C}$ denotes the spectrum (i.e., set of eigenvalues) of a square matrix M.

For $\mathcal{S} = \mathbb{C}^{n,n}$, $\Lambda_\varepsilon(A) = \Lambda_\varepsilon^{\mathbb{C}}(A) = \Lambda_\varepsilon^{\mathcal{S}}(A)$ is known as the *complex* ε-pseudospectrum, and for $\mathcal{S} = \mathbb{R}^{n,n}$, $\Lambda_\varepsilon^{\mathbb{R}}(A) = \Lambda_\varepsilon^{\mathcal{S}}(A)$ is known as the *real* ε-pseudospectrum.

The above quantity $\alpha_\varepsilon^{\mathcal{S}}(A)$, which is known as the ($\mathcal{S}$-structured) *ε-pseudospectral abscissa* of the matrix A, can then be rewritten more compactly as

$$\alpha_\varepsilon^{\mathcal{S}}(A) = \max\{\mathrm{Re}\,\lambda : \ \lambda \in \Lambda_\varepsilon^{\mathcal{S}}(A)\}. \tag{1.2}$$

stability-abscissa

An analogous quantity, of interest for discrete-time linear dynamical systems $x_{k+1} = Ax_k$, is the ($\mathcal{S}$-structured) *ε-pseudospectral radius* of the matrix A,

$$\rho_\varepsilon^{\mathcal{S}}(A) = \max\{|\lambda| : \ \lambda \in \Lambda_\varepsilon^{\mathcal{S}}(A)\}. \tag{1.3}$$

stability-radius

In these notions, we drop the superscript $\mathcal{S}$ when we consider the unstructured complex case $\mathcal{S} = \mathbb{C}^{n,n}$. We then write $\Lambda_\varepsilon(A)$, $\alpha_\varepsilon(A)$ and $\rho_\varepsilon(A)$.

III.1.2 Complex pseudospectrum and singular values

The complex ε-pseudospectrum can be characterized in terms of singular values. The singular value decomposition of a matrix $M \in \mathbb{C}^{n,n}$ is $M = U\Sigma V^*$, with unitary matrices $U = (u_1, \dots, u_n)$ and $V = (v_1, \dots, v_n)$ formed by the left and right singular vectors u_k and v_k, respectively, and with the real diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$ of the singular values $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$. We use the notation $\sigma_k(M)$ for the kth singular value of M when we wish to indicate the dependence on M, and we write $\sigma_{\min}(M) = \sigma_n(M)$ for the smallest singular value.

thm:ps-sv

Theorem 1.2 (Singular values and eigenvalues). *The complex ε-pseudospectrum of $A \in \mathbb{C}^{n,n}$ is characterized as*

$$\begin{aligned} \Lambda_\varepsilon(A) &= \{\lambda \in \mathbb{C} : \ \sigma_{\min}(A - \lambda I) \leq \varepsilon\} \qquad (1.4)\\ &= \{\lambda \in \mathbb{C} : \ \lambda \in \Lambda(A + \Delta) \textit{ for some } \Delta \in \mathbb{C}^{n,n} \textit{ with } \|\Delta\| \leq \varepsilon\}. \end{aligned}$$

ps-sv

Moreover, the perturbation matrix Δ can be restricted to be of rank 1.

Proof. The result relies on the fact that the distance to singularity of a matrix M equals its smallest singular value:

$$\sigma_{\min}(M) = \min\{\|\Delta\| : \ \Delta \in \mathbb{C}^{n,n} \text{ is such that } M + \Delta \text{ is singular}\}. \tag{1.5}$$

dist-sing

The perturbation of minimal norm is then the rank-1 matrix

$$\Delta_\star = -\sigma_n u_n v_n^* \tag{1.6}$$

Delta-star

(unique if $\sigma_n < \sigma_{n-1}$), where $\sigma_n = \sigma_{\min}(M)$ and u_n, v_n are the nth singular vectors, which yields that $M + \Delta_\star$ has the same singular value decomposition as M except that the smallest singular value is replaced by zero.

Choosing $M = A - \lambda I$ for $\lambda \in \mathbb{C}$ thus shows that $\sigma_{\min}(A - \lambda I) \leq \varepsilon$ if and only if there exists a matrix $\Delta \in \mathbb{C}^{n,n}$ of norm at most ε such that $A - \lambda I + \Delta$ is singular, or in other words, that λ is an eigenvalue of $A + \Delta$. □

Since $\sigma_{\min}(A - \lambda I)$ depends continuously on λ, an immediate consequence of Theorem 1.2 is that the boundary of the ε-pseudospectrum of A is given as

$$\partial \Lambda_\varepsilon(A) = \{\lambda \in \mathbb{C} : \ \sigma_{\min}(A - \lambda I) = \varepsilon\}. \tag{1.7}$$

ps-sv-bdy

rem:ps-norms

Remark 1.3 (Frobenius norm and matrix 2-norm). Since for rank-1 matrices, the Frobenius norm and the matrix 2-norm are the same, Theorem 1.2 and its proof show that the complex ε-pseudospectra defined with respect to these two norms are identical. This no longer holds true for real and structured pseudospectra.

Since $1/\sigma_{\min}(A - \lambda I) = \sigma_{\max}\big((A - \lambda I)^{-1}\big) = \|(A - \lambda I)^{-1}\|_2$, we can reformulate (1.4) in terms of resolvents $(A - \lambda I)^{-1}$ as

$$\Lambda_\varepsilon(A) = \{\lambda \in \mathbb{C} : \ \|(A - \lambda I)^{-1}\|_2 \geq 1/\varepsilon\}. \tag{1.8}$$

ps-res

This allows us to characterize the ε-pseudospectral abscissa (1.2) as

$$\alpha_\varepsilon(A) = \max\{\mathrm{Re}\,\lambda \ : \ \|(A - \lambda I)^{-1}\|_2 \geq 1/\varepsilon\},$$

which implies

$$\frac{1}{\varepsilon} = \max_{\mathrm{Re}\,\lambda \geq \alpha_\varepsilon(A)} \|(A - \lambda I)^{-1}\|_2. \tag{1.9}$$

eps-res-bound

If all eigenvalues of A have negative real part, we define the *stability radius* (or *distance to instability*) as $\varepsilon_\star > 0$ such that $\alpha_{\varepsilon_\star}(A) = 0$, i.e., there exists a perturbation $\Delta \in \mathbb{C}^{n,n}$ of Frobenius norm $\varepsilon_\star$ such that $A + \Delta$ has an eigenvalue on the imaginary axis, as opposed to all perturbations of smaller norm. The above formula then yields that the inverse stability radius $1/\varepsilon_\star$ is the smallest upper bound of $\|(A - \lambda I)^{-1}\|_2$ for λ in the right half-plane:

$$\frac{1}{\varepsilon_\star} = \max_{\mathrm{Re}\,\lambda \geq 0} \|(A - \lambda I)^{-1}\|_2. \tag{1.10}$$

oeps-res-bound

As we discuss next, the ε-pseudospectral abscissa $\alpha_\varepsilon(A)$ and the stability radius $\varepsilon_\star$ are important quantities in bounding solutions of linear differential equations.

rem:ps-exp

Estimating transient behaviour of linear differential equations via resolvent bounds. We describe two approaches to bounding solutions to linear differential equations, one for the matrix exponential e^{tA}, which corresponds to the homogeneous initial value problem $\dot{x}(t) = Ax(t)$ with an arbitrary initial value $x(0) = x_0$, and the other approach for the inhomogeneous problem $\dot{x}(t) = Ax(t) + f(t)$ with zero initial value.

(i) Via Theorem 1.2, the transient behaviour of $\|\mathrm{e}^{tA}\|_2$ can be bounded in terms of the complex pseudospectrum. Here we illustrate this with a simple robust bound: Let $\Gamma_\varepsilon \subset \mathbb{C}$ be the boundary curve of a piecewise smooth domain (or several non-overlapping domains) whose closure covers $\Lambda_\varepsilon(A)$, and assume further that the real part of the rightmost point of Γ_ε equals the pseudospectral abscissa $\alpha_\varepsilon(A)$. In particular, we may take $\Gamma_\varepsilon = \partial\Lambda_\varepsilon(A)$ when this is a piecewise regular curve. Using the Cauchy integral representation

$$\mathrm{e}^{tA} = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_\varepsilon} \mathrm{e}^{t\lambda}\,(\lambda I - A)^{-1}\,d\lambda$$

and noting that by (1.8), $\|(\lambda I - A)^{-1}\|_2 \le 1/\varepsilon$ on Γ_ε, we find by taking norms that

$$\|\mathrm{e}^{tA}\|_2 \le \frac{\gamma_\varepsilon(t)}{\varepsilon} \quad \text{with} \quad \gamma_\varepsilon(t) = \frac{1}{2\pi}\int_{\Gamma_\varepsilon} |\mathrm{e}^{t\lambda}|\,|d\lambda| \le \frac{|\Gamma_\varepsilon|}{2\pi}\,\mathrm{e}^{t\alpha_\varepsilon(A)}, \qquad (1.11)$$

transient-bound

where $|\Gamma_\varepsilon|$ is the length of Γ_ε. This bound holds for every $\varepsilon > 0$.

The same argument can be applied to a perturbed matrix $A + \Delta$ with $\Delta \in \mathbb{C}^{n,n}$ bounded by $\|\Delta\|_2 \le \delta$. Using the bound $\sigma_{\min}(A + \Delta - \lambda I) \ge \sigma_{\min}(A - \lambda I) - \delta$, we then obtain the robust bound

$$\|\mathrm{e}^{t(A+\Delta)}\|_2 \le \frac{\gamma_\varepsilon(t)}{\varepsilon - \delta} \qquad \text{for } \|\Delta\|_2 \le \delta \text{ and for all } \varepsilon > \delta. \qquad (1.12)$$

transient-perturbed

This bound can be optimized over $\varepsilon > \delta$, provided that a bound for $|\Gamma_\varepsilon|$ and an algorithm for computing the pseudospectral abscissa $\alpha_\varepsilon(A)$ are available. An improved related bound for structured perturbations $\Delta \in \mathcal{S}$ will later be given in (V.5.3).

(ii) We consider the linear differential equation $\dot{x}(t) = Ax(t) + f(t)$ with zero initial value for inhomogeneities $f \in L^2(0,\infty;\mathbb{C}^n)$, where we assume that all eigenvalues of A have negative real part. We extend $x(t)$ and $f(t)$ to $t < 0$ by zero. Their Fourier transforms $\widehat{x}$ and $\widehat{f}$ are then related by $\mathrm{i}\omega\,\widehat{x}(\omega) = A\widehat{x}(\omega) + \widehat{f}(\omega)$ for all $\omega \in \mathbb{R}$, i.e.,

$$\widehat{x}(\omega) = (\mathrm{i}\omega I - A)^{-1}\widehat{f}(\omega), \qquad \omega \in \mathbb{R},$$

and hence the Plancherel formula yields

$$\begin{aligned} &\int_{\mathbb{R}} \|x(t)\|^2\,dt = \int_{\mathbb{R}} \|\widehat{x}(\omega)\|^2\,d\omega = \int_{\mathbb{R}} \|(\mathrm{i}\omega I - A)^{-1}\widehat{f}(\omega)\|^2\,d\omega \\ &\le \max_{\omega\in\mathbb{R}} \|(\mathrm{i}\omega I - A)^{-1}\|_2^2 \int_{\mathbb{R}} \|\widehat{f}(\omega)\|^2\,d\omega = \max_{\omega\in\mathbb{R}} \|(\mathrm{i}\omega I - A)^{-1}\|_2^2 \int_{\mathbb{R}} \|f(t)\|^2\,dt. \end{aligned}$$

Using (1.10) and the causality property that $x(t)$, for $0 \le t \le T$, only depends on $f(\tau)$ with $0 \le \tau \le t \le T$ (which allows us to extend $f(t)$ by 0 for $t > T$), we thus obtain the bound

$$\left(\int_0^T \|x(t)\|^2\,dt\right)^{1/2} \le \frac{1}{\varepsilon_\star}\left(\int_0^T \|f(t)\|^2\,dt\right)^{1/2}, \qquad 0 \le T \le \infty, \qquad (1.13)$$

oeps-L2-bound

where $\varepsilon_\star$ is the stability radius of A, i.e. $\alpha_{\varepsilon_\star}(A) = 0$. A related robust bound for $A + \Delta$ with structured perturbations $\Delta \in \mathcal{S}$ will later be given in (V.5.2).

Extremal perturbations. The proof of Theorem 1.2 also yields the following result on the perturbations Δ of minimal norm ε such that $A + \Delta$ has a prescribed eigenvalue λ on the boundary $\partial\Lambda_\varepsilon(A)$ of the complex ε-pseudospectrum of A.

lem:Delta-C

Lemma 1.4 (Extremal complex perturbations). *Let $\lambda \in \partial\Lambda_\varepsilon(A)$, and let $\Delta \in \mathbb{C}^{n,n}$ of norm ε be such that $A+\Delta$ has the eigenvalue λ. Then, Δ is of rank* 1.

Assume now that ε is a simple *singular value of $A-\lambda I$ and that the corresponding left and right singular vectors are not orthogonal to each other. Then,*

$$\Delta = \varepsilon \mathrm{e}^{\mathrm{i}\theta} x y^*,$$

where $\mathrm{e}^{\mathrm{i}\theta}$ is the outer normal to $\partial\Lambda_\varepsilon(A)$ at λ, which is uniquely determined, and x and y are left and right eigenvectors of $A+\Delta$ to the eigenvalue λ, of unit norm and with $x^*y > 0$.

Proof. By (1.7), we have $\sigma_{\min}(A-\lambda I) = \varepsilon$. The proof of Theorem 1.2 then shows that $\Delta = -\varepsilon u v^*$, where u and v are left and right singular vectors of $A-\lambda I$, with

$$(A-\lambda I+\Delta)v = 0 \quad \text{and} \quad u^*(A-\lambda I+\Delta) = 0,$$

or equivalently, $(A+\Delta)v = \lambda v$ and $u^*(A+\Delta) = \lambda u^*$. This shows that u and v are left and right eigenvectors of $A+\Delta$.

Assume now that ε is a simple singular value of $A-\lambda I$ and that $u^*v \neq 0$. We show that $\partial\Lambda_\varepsilon(A)$ has the outer normal $\mathrm{e}^{\mathrm{i}\theta}$ at λ, where the angle θ is determined by

$$u^*v = -\rho \mathrm{e}^{-\mathrm{i}\theta}, \quad \rho > 0.$$

Let $\gamma(t)$, for t near 0, be a path in the complex plane with $\gamma(0) = \lambda \in \partial\Lambda_\varepsilon(A)$. With $\nu = \dot\gamma(0)$ we have, by the standard derivative formula of simple eigenvalues,

$$\frac{d}{dt}\Big|_{t=0} \sigma_{\min}(A-\gamma(t)I) = \frac{\begin{pmatrix} u \\ v \end{pmatrix}^* \dfrac{d}{dt}\Big|_{t=0} \begin{pmatrix} 0 & A-\gamma(t)I \\ (A-\gamma(t)I)^* & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}}{\begin{pmatrix} u \\ v \end{pmatrix}^* \begin{pmatrix} u \\ v \end{pmatrix}}$$

$$= -\tfrac{1}{2}\,\mathrm{Re}(\nu\, u^*v) = \tfrac{1}{2}\rho\,\mathrm{Re}(\nu \mathrm{e}^{-\mathrm{i}\theta}).$$

This shows that $\nu = \mathrm{e}^{\mathrm{i}\theta}$ is the unique direction of steepest ascent, which is orthogonal to the level set $\partial\Lambda_\varepsilon(A)$ and points out of $\Lambda_\varepsilon(A)$. Hence, $\mathrm{e}^{\mathrm{i}\theta}$ is the outer normal to $\partial\Lambda_\varepsilon(A)$ at λ.

We set $x = -\mathrm{e}^{-\mathrm{i}\theta}u$ and $y = v$, which gives us a pair of left and right eigenvectors of $A+\Delta$ with $x^*y = \rho > 0$. We then have

$$\Delta = -\varepsilon u v^* = \varepsilon \mathrm{e}^{\mathrm{i}\theta} x y^*,$$

which proves the result. □

As we will see, Lemmas 1.2 and 1.4 motivate different approaches to computing the boundary of the complex ε-pseudospectrum (or just extremal points such as a rightmost point): methods that steer the smallest singular value of $A-\lambda I$ to ε, and methods that iterate on rank-1 matrices.

III.1.3 Real and structured pseudospectra

For real and structured pseudospectra, there is apparently no characterization in terms of singular value decompositions. Nevertheless, there are analogues of Lemma 1.4 at least at boundary points $\lambda \in \partial \varLambda_\varepsilon^{\mathcal{S}}(A)$ where the boundary is differentiable and thus admits an outer normal $\mathrm{e}^{\mathrm{i}\theta}$.

lem:Delta-R

Lemma 1.5 (Extremal real perturbations). *Let $\lambda \in \partial \varLambda_\varepsilon^{\mathbb{R}}(A)$ be on a smooth section of the boundary, with outer normal $\mathrm{e}^{\mathrm{i}\theta}$ at λ. Let $\varDelta \in \mathbb{R}^{n,n}$ of Frobenius norm ε be such that $A+\varDelta$ has λ as a simple eigenvalue. Let x and y be left and right eigenvectors of $A+\varDelta$ to the eigenvalue λ, of unit norm and with $x^*y>0$. If the matrix $\mathrm{Re}(\mathrm{e}^{\mathrm{i}\theta}xy^*)$ is non-zero, then*

$$\varDelta = \varepsilon\alpha\,\mathrm{Re}(\mathrm{e}^{\mathrm{i}\theta}xy^*),$$

where $\alpha = 1/\|\mathrm{Re}(\mathrm{e}^{\mathrm{i}\theta}xy^*)\|_F > 0$. In particular, $\varDelta$ is of rank at most 2.

We note that the condition $\mathrm{Re}(\mathrm{e}^{\mathrm{i}\theta}xy^*) \neq 0$ excludes sections of $\partial \varLambda_\varepsilon^{\mathbb{R}}(A)$ that consist of intervals on the real line.

Proof. The proof is based on Section II.2.2, in particular (II.2.8). In the proof we denote the given $\lambda \in \partial \varLambda_\varepsilon^{\mathbb{R}}(A)$ as λ_0 and use λ to denote a complex variable. Similarly, we denote the matrix $\varDelta$ in the statement of the lemma by $\varDelta_0$ and use $\varDelta$ for a generic matrix in $\mathbb{R}^{n,n}$.

We set $\mu = \lambda_0 + \delta\mathrm{e}^{\mathrm{i}\theta}$ with a small $\delta > 0$. For the function

$$f(\lambda,\overline{\lambda}) = |\mu-\lambda|^2 = (\mu-\lambda)(\overline{\mu}-\overline{\lambda})$$

we consider the optimization problem (II.2.1) with $\lambda(A+\varDelta)$ denoting the eigenvalue of $A+\varDelta$ closest to λ_0, which has $\varDelta_0 = \varepsilon E_0$ with $\|E_0\| = 1$ as a solution, and $\lambda(A+\varDelta_0) = \lambda_0$. In particular, E_0 is then a stationary point of the gradient system (II.2.6), and hence satisfies (II.2.8) with a negative factor as in (II.1.17), which yields that $\varDelta_0$ has the stated form, with α a negative multiple of $\partial f/\partial\overline{\lambda}(\lambda_0,\overline{\lambda}_0) = -(\mu-\lambda_0) = -\delta\mathrm{e}^{\mathrm{i}\theta} \neq 0$. This holds under the condition that the gradient $G = \mathrm{Re}(-\delta\mathrm{e}^{\mathrm{i}\theta}xy^*)$ is non-zero. □

lem:Delta-S

Lemma 1.6 (Extremal structured perturbations). *Let $\lambda \in \partial \varLambda_\varepsilon^{\mathcal{S}}(A)$ be on a smooth section of the boundary, with outer normal $\mathrm{e}^{\mathrm{i}\theta}$ at λ. Let $\varDelta \in \mathcal{S}$ of Frobenius norm ε be such that $A+\varDelta$ has λ as a simple eigenvalue. Let x and y be left and right eigenvectors of $A+\varDelta$ to the eigenvalue λ, of unit norm and with $x^*y>0$. If the matrix $\varPi^{\mathcal{S}}(\mathrm{e}^{\mathrm{i}\theta}xy^*)$ is non-zero, where $\varPi^{\mathcal{S}}$ is the orthogonal projection (II.3.3) onto $\mathcal{S}$, then*

$$\varDelta = \varepsilon\alpha\,\varPi^{\mathcal{S}}(\mathrm{e}^{\mathrm{i}\theta}xy^*),$$

where $\alpha = 1/\|\varPi^{\mathcal{S}}(\mathrm{e}^{\mathrm{i}\theta}xy^*)\|_F > 0$.

Proof. The proof is analogous to that of Lemma 1.5, using now the structure-constrained gradient flow and the characterization of its stationary points of Section II.3.3. □

III.2 Computing the pseudospectral abscissa

sec:psa

As we have seen in the previous section, knowing the ε-pseudospectral abscissa $\alpha_\varepsilon(A)$ is important for ensuring robust stability of a linear dynamical system. Various intriguing algorithms based on different ideas have been proposed to compute the complex pseudospectral abscissa:

- the criss-cross algorithm of Burke, Lewis & Overton (2003), which is based on Theorem 1.2 and on Byers' Lemma given below;
- the rank-1 iteration of Guglielmi & Overton (2011), which is based on Lemma 1.4;
- the rank-1 projected gradient flow algorithm of Guglielmi & Lubich (2011); see Section II.1 with $f(\lambda,\overline{\lambda}) = -\frac{1}{2}(\lambda+\overline{\lambda}) = -\mathrm{Re}\,\lambda$;
- the subspace method of Kressner & Vandereycken (2014).

With the adaptations of Sections II.2 and II.3, the low-rank projected gradient flow approach can also be used for computing real and structured pseudospectral abscissas.

III.2.1 Criss-cross algorithm

ec:criss-cross

This remarkable algorithm was proposed and analysed by Burke, Lewis & Overton (2003). It uses a sequence of vertical and horizontal searches in the complex plane to identify the intersection of a given line with $\partial\Lambda_\varepsilon(A)$. Horizontal searches yield updates to the approximation of $\alpha_\varepsilon(A)$ while vertical searches find favourable locations for the horizontal searches.

The criss-cross algorithm computes a monotonically growing sequence (α^k) that converges to the complex ε-pseudospectral abscissa $\alpha_\varepsilon(A)$. In its basic form, it can be written as follows:

```
0. Initialize $\alpha^0 = \max\{\mathrm{Re}\,\lambda \,:\, \lambda \in \Lambda(A)\}$.
1. For $k = 0,1,2,\dots$ iterate
1.1 (Vertical search)

        Find all real numbers $\beta_j$, in increasing order for $j$ from 0 to $m$,
        such that $\alpha^k + \mathrm{i}\beta_j \in \partial\Lambda_\varepsilon(A)$.          (2.1)   cc-vertical

1.2 (Horizontal search) For $j = 0,\dots,m-1$,

        let the midpoint $\beta_{j+1/2} = \frac{1}{2}(\beta_j + \beta_{j+1})$;
        if $\alpha^k + \mathrm{i}\beta_{j+1/2} \in \Lambda_\varepsilon(A)$, find the largest real number $\widehat{\alpha}_{j+1/2}$
        such that $\widehat{\alpha}_{j+1/2} + \mathrm{i}\beta_{j+1/2} \in \partial\Lambda_\varepsilon(A)$ .          (2.2)   cc-horizontal

1.3 Take $\alpha^{k+1}$ as the maximum of the $\widehat{\alpha}_{j+1/2}$.
```

To turn this into a viable algorithm, (2.1) and (2.2) need to be computed efficiently. This becomes possible thanks to the following basic lemma, applied to A for (2.1) and to the rotated matrix $\mathrm{i}A$ for (2.2).

lem:byers **Lemma 2.1 (Byers' Lemma).** *Let $A \in \mathbb{C}^{n,n}$. For given real numbers α and β, the number $\varepsilon > 0$ is a singular value of the matrix*

$$A - (\alpha + \mathrm{i}\beta)I$$

if and only if $\mathrm{i}\beta$ *is an eigenvalue of the Hamiltonian matrix*

$$H(A,\alpha) = \begin{pmatrix} -(A-\alpha I)^* & \varepsilon I \\ -\varepsilon I & A - \alpha I \end{pmatrix}. \tag{2.3}$$ eq:H

Proof. After taking $A - \alpha I$ in the role of A, we can assume $\alpha = 0$ in the following. The imaginary number $\mathrm{i}\beta$ is an eigenvalue of the Hamiltonian matrix (2.3) if and only if there exist nonzero vectors u and v such that

$$\begin{pmatrix} -A^* & \varepsilon I \\ -\varepsilon I & A \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \mathrm{i}\beta \begin{pmatrix} u \\ v \end{pmatrix}. \tag{2.4}$$ eq:eigH

This is equivalent to

$$(A - \mathrm{i}\beta I)^*\, u = \varepsilon v, \qquad (A - \mathrm{i}\beta I)\, v = \varepsilon u, \tag{2.5}$$ eq:uv

which expresses that ε is a singular value of $A - \mathrm{i}\beta I$. □

Using Lemma 2.1, the *vertical search* (2.1) is done by computing all purely imaginary eigenvalues $\mathrm{i}\beta$ of the Hamiltonian matrix and discarding those eigenvalues among them for which ε is not the smallest singular value of $A - (\alpha_k + \mathrm{i}\beta)I$. The *horizontal search* (2.2) is done by computing the purely imaginary eigenvalue $\mathrm{i}\widehat{\alpha}_{j+1/2}$ of largest imaginary part of the Hamiltonian matrix that corresponds to the matrix $\mathrm{i}A$ in the role of A and $-\beta_{j+1/2}$ in the role of α. The method is written in pseudocode in Algorithm 3.

The computational cost of an iteration step of the criss-cross algorithm is thus determined by computing imaginary eigenvalues of complex Hamiltonian $2n \times 2n$ matrices. All imaginary eigenvalues are needed in the vertical search and the ones of largest imaginary part in the horizontal search. In addition, the smallest singular values of complex $n \times n$ matrices need to be computed to decide if a given complex number is in $\Lambda_\varepsilon(A)$.

An illustration is given in Fig. 7 for the 6×6 matrix given in (II.1.36).

Unconditional convergence. As the following theorem by Burke, Lewis & Overton (2003) shows, the sequence generated by the criss-cross algorithm always converges to the ε-pseudospectral abscissa.

thm:cc-conv **Theorem 2.2 (Convergence of the criss-cross algorithm).** *For every matrix $A \in \mathbb{C}^{n,n}$, the sequence (α^k) of the criss-cross algorithm converges to the pseudospectral abscissa $\alpha_\varepsilon(A)$.*

Proof. By construction, the sequence (α_k) is a monotonically increasing sequence of real parts of points on $\partial\Lambda_\varepsilon(A)$, and it is bounded as $\Lambda_\varepsilon(A)$ is bounded. Therefore, (α_k) converges to a limit $\alpha^\star$, which is the real part of some point on $\partial\Lambda_\varepsilon(A)$. Hence, $\alpha^\star \le \alpha_\varepsilon(A)$. It remains to show that actually $\alpha^\star = \alpha_\varepsilon(A)$.

fig:cc

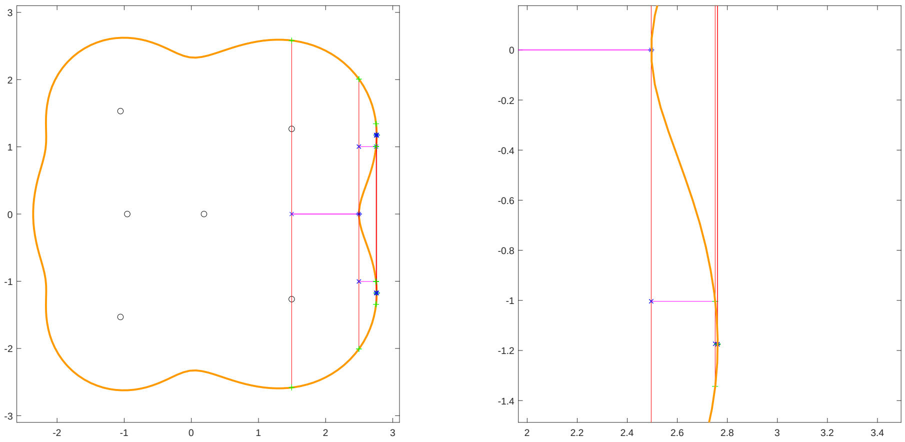

Fig. 2.1. Iterates of the criss cross algorithm applied to the matrix A of (1.36) and $\varepsilon = 1$. Right picture: zoom close to a rightmost point

Algorithm 3: Criss-cross algorithm

```
Data: Matrix $A$, $\varepsilon > 0$, tol a given positive tolerance
Result: $\alpha_\varepsilon(A)$
begin
    Set $k = 1$, $\alpha^0 = \alpha(A)$
    while $\alpha^k - \alpha^{k-1} > \mathrm{tol}$ do
        Vertical iteration.
        Find all imaginary eigenvalues $\{\widehat{\beta}_j\}$ of $H(A, \alpha^k)$ (see (2.3)) and collect those for
          which $\varepsilon$ is the smallest singular value of $A - (\alpha^k + \mathrm{i}\widehat{\beta}_j)$ as $\{\beta_j\}_{j=0}^m$,
              $\beta_0 \le \beta_1 \le \beta_2 \le \ldots \le \beta_{m-1} \le \beta_m.$
          Horizontal iteration.
        for $j = 0, \ldots, m-1$ do
            Compute the midpoints $\beta_{j+1/2} = \frac{\beta_j + \beta_{j+1}}{2}$
            Find the highest imaginary eigenvalue (i.e. with largest imaginary part)
              $\mathrm{i}\alpha_{j+1/2}$ of $H(\mathrm{i}A, \beta_{j+1/2})$ (see (2.3))
        Set $k = k + 1$
        Set $\alpha^{k+1} = \max_{j=0,\ldots,m-1} \alpha_{j+1/2}$
```

alg_cc

To this end, we use the fact that every path-connected component of $\Lambda_\varepsilon(A)$ contains an eigenvalue of A. This is readily seen as follows: For any $\lambda_1 \in \Lambda_\varepsilon(A)$, there exists a matrix $\Delta \in \mathbb{C}^{n,n}$ of norm at most ε such that λ_1 is an eigenvalue of $A + \Delta$. Consider now the path $A + \theta\Delta$, $0 \le \theta \le 1$. By the continuity of eigenvalues, to this path corresponds a path of eigenvalues $\lambda(\theta)$ of $A + \theta\Delta$ with $\lambda(1) = \lambda_1$, which connects λ_1 with the eigenvalue $\lambda(0)$ of A.

Suppose $\alpha^\star < \alpha_\varepsilon(A)$. We lead this to a contradiction. Let $\lambda_1 \in \Lambda_\varepsilon(A)$ be such that $\mathrm{Re}\,\lambda_1 = \alpha_\varepsilon(A)$. Then there is a path $\lambda(\theta)$ to an eigenvalue λ_0 of A, which by construction has a real part that does not exceed α^0 and hence is smaller than $\alpha^\star$. So there exist $0 < \theta^\star < 1$ such that $\lambda(\theta^\star) \in \Lambda_{\theta^\star\varepsilon}(A)$ has real part $\alpha^\star$, so that $\lambda(\theta^\star) = \alpha^\star + \mathrm{i}\beta^\star$ for some real $\beta^\star$. There exists a smallest interval $[\beta_0, \beta_1]$ that contains $\beta^\star$ and has boundary points such that $\alpha^\star + \mathrm{i}\beta_0, \alpha^\star + \mathrm{i}\beta_1 \in \partial\Lambda_\varepsilon(A)$. Then, the points $\alpha^\star + \mathrm{i}\beta$ with $\beta_0 < \beta < \beta_1$ are in the interior of $\Lambda_\varepsilon(A)$, and in particular this holds true for the midpoint $\beta = \frac{1}{2}(\beta_0 + \beta_1)$. Hence there exists an $\alpha > \alpha^\star$ such that $\alpha + \mathrm{i}\beta \in \partial\Lambda_\varepsilon(A)$. But then, also the criss-cross algorithm would have found an $\alpha^k > \alpha^\star$, in contradiction to the maximality of $\alpha^\star$. So we must have $\alpha^\star = \alpha_\varepsilon(A)$. □

Locally quadratic convergence. We here show that the criss-cross algorithm converges locally quadratically under the following regularity assumption:

At every right-most point of the ε-pseudospectrum of A, the boundary curve $\partial\Lambda_\varepsilon(A)$ is smooth with nonzero curvature. (2.6) right-reg

This condition is stronger than the condition of a simple smallest singular value of $A-\lambda I$ at right-most points $\lambda \in \Lambda_\varepsilon(A)$ imposed by Burke, Lewis & Overton (2003), but it allows for a short proof of locally quadratic convergence.

m:cc-conv-quad

Theorem 2.3 (Locally quadratic convergence of the criss-cross algorithm). *Under condition* (2.6)*, the sequence* (α^k) *of the criss-cross algorithm converges locally quadratically to* $\alpha^\star = \alpha_\varepsilon(A)$ *:*

$$0 \le \alpha^\star - \alpha^{k+1} \le C(\alpha^\star - \alpha^k)^2,$$

where C *is independent of k, provided that* α^k *is sufficiently close to* $\alpha^\star$*.*

Proof. Near a right-most boundary point $\alpha^\star + \mathrm{i}\beta^\star$ of $\Lambda_\varepsilon(A)$, boundary points $\alpha + \mathrm{i}\beta$ are related by

$$\alpha^\star - \alpha = f(\beta) = c^2(\beta^\star - \beta)^2 + O((\beta^\star - \beta)^3),$$

where $c \neq 0$ by condition (2.6). For the variables $\eta = \alpha^\star - \alpha$ and $\xi = c(\beta^\star - \beta)$ this relation becomes

$$\eta = \varphi(\xi) = \xi^2 + O(\xi^3).$$

For a small $\delta > 0$, let now $\xi_+ = \delta$ and choose $\xi_- = -\delta + O(\delta^2)$ such that $\varphi(\xi_-) = \varphi(\xi_+) = \delta^2(1 + O(\delta))$. Then,

$$\tfrac{1}{2}(\xi_+ + \xi_-) = O(\delta^2) \quad \text{and hence} \quad \varphi\big(\tfrac{1}{2}(\xi_+ + \xi_-)\big) = O(\delta^4),$$

which yields

$$\varphi\big(\tfrac{1}{2}(\xi_+ + \xi_-)\big) \le C\varphi(\xi_+)^2.$$

Translated back to the original variables, this yields the stated result. □

ec:r1-iteration

III.2.2 Iteration on rank-1 matrices

Guglielmi and Overton (2011) proposed a strikingly simple iterative algorithm for computing the pseudospectral abscissa that uses a sequence of rank-1 perturbations of the matrix. Working with rank-1 perturbations appears natural in view of Lemma 1.4. Moreover, this lemma (with $\theta = 0$) shows that at a point $\lambda \in \partial\Lambda_\varepsilon(A)$ such that $\mathrm{Re}\lambda = \alpha_\varepsilon(A)$, where the outer normal is horizontal to the right, the corresponding matrix perturbation Δ of norm ε is such that $\Delta = \varepsilon x y^*$, where x and y are left and right eigenvectors, of unit norm and with $xy^* > 0$, to the eigenvalue λ of $A + \Delta$. This motivates the following fixed-point iteration.

Basic rank-1 iteration. The basic iteration starts from two vectors u_0 and v_0 of unit norm and runs as follows for $k = 0, 1, 2, \dots$:

Given a rank-1 matrix $E_k = u_k v_k^*$ of unit norm, compute the rightmost eigenvalue λ_k of $A + \varepsilon E_k$ and left and right eigenvectors x_k and y_k, of unit norm and with $x_k^* y_k > 0$, and set $E_{k+1} = u_{k+1} v_{k+1}^* := x_k y_k^*$.

Algorithm 4 gives a formal description. This algorithm requires in each step one computation of rightmost eigenvalues and associated eigenvectors of rank-1 perturbations to the matrix A, which can be computed at relatively small computational cost for large sparse matrices A, either combining the Cayley transformation approach with the Sherman-Morrison formula or by using an implicitly restarted Arnoldi method (as implemented in ARPACK and used in the MATLAB function *eigs*).

Algorithm 4: Rank-1 iteration

```
Data: Matrix $A$, $\varepsilon > 0$, tol a given positive tolerance
Result: $r \le \alpha_\varepsilon(A)$, $x, y$
begin
    Compute $x_0$ and $y_0$ left and right eigenvectors to the rightmost eigenvalue of $A$, both
      normalized to unit norm and with $x_0^* y_0 > 0$.
    Let $r_{-1} = -\infty$
    Compute $x_0, y_0$, left and right eigenvectors to the rightmost eigenvalue $\lambda_0$ of $A$, with
      $x_0, y_0$ of unit norm such that $x_0^* y_0 > 0$
    Set $r_0 = \mathrm{Re}(\lambda_0)$
    Set $k = 0$
    while $r_k - r_{k-1} > \mathrm{tol}$ do
        Compute $x_{k+1}, y_{k+1}$, left and right eigenvectors to the rightmost eigenvalue $\lambda_{k+1}$
          of $A + \varepsilon E_k$, $E_k = x_k y_k^*$, with $x_k, y_k$ of unit norm such that $x_k^* y_k > 0$
        Set $k = k + 1$
        Set $r_k = \mathrm{Re}(\lambda_k)$
    Set $r = r_k$
    Set $x = x_k, y = y_k$
```

alg_GO

An illustration is given in Fig. 10 for the 6 × 6 matrix (1.36).

The expectation is that $\mathrm{Re}\,\lambda_k$ converges to the ε-pseudospecral abscissa $\alpha_\varepsilon(A)$, as is observed in numerical experiments. Indeed, *if* the iteration converges, $\lambda_k \to \lambda$ and $x_k \to x$, $y_k \to y$, then x, y are of unit norm with $x^* y > 0$ and

$$x, y \text{ are left and right eigenvectors to the rightmost eigenvalue } \lambda \text{ of } A + \varepsilon x y^*. \qquad (2.7)$$

stat-lim

This implies $(A - \lambda I)y = -\varepsilon x$ and $x^*(A - \lambda I) = -\varepsilon y^*$, which shows that $A - \lambda I$ has the singular value ε (as is required for having $\lambda \in \partial \Lambda_\varepsilon(A)$ by (1.7)) — though ε is here not known to be the smallest singular value. Furthermore the gradient of the associated singular value is $x^* y > 0$, that is, the gradient is horizontal to the right in the complex plane. By Lemma 1.4, this implies that $\lambda \in \partial \Lambda_\varepsilon(A)$ with outer normal 1 if ε is indeed the *smallest* singular value of $A - \lambda I$.

Moreover, in the interpretation of Theorem II.1.5 and Remark II.1.8, the property (2.7) implies that $E = xy^*$ is a *stationary point* (though not necessarily a maximum) of the eigenvalue optimization problem to find

$$\arg \max_{\|E\|_F \le 1} \mathrm{Re}\,\lambda(A + \varepsilon E), \qquad (2.8)$$

eigopt-psa

i.e. Problem (II.1.1) with $f(\lambda, \overline{\lambda}) = -\mathrm{Re}\,\lambda$.

There exist no results about global convergence of the rank-1 iteration. Local linear convergence can be shown for a sufficiently small ratio of the two smallest singular values, fig:goit

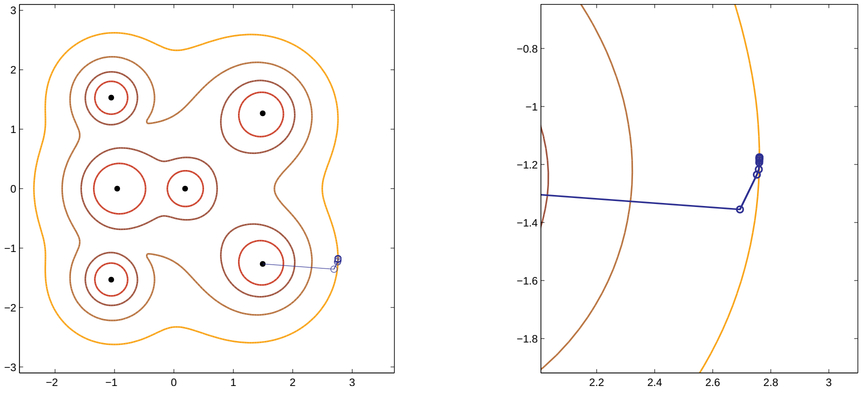

Fig. 2.2. Iterates λ_k of Guglielmi and Overton's algorithm applied to the matrix A of (1.36) and $\varepsilon = 1$. Right: zoom.

$\varepsilon/\sigma_{n-1}(A-\lambda I)$, by studying the derivative of the iteration map at a stationary point. This requires bounds of derivatives of eigenvectors using appropriate representations of the group inverse, as laid out in the Appendix (and used later in this book).

Monotone rank-1 iteration. The simple rank-1 iteration described above is not guaranteed to yield a monotonically increasing sequence $(\mathrm{Re}\,\lambda_k)$. Guglielmi and Overton (2011) also proposed a monotone variant that is described in the following.

For given vectors u, v of unit norm, we start from the rank-1 perturbation $A+\varepsilon uv^*$ with rightmost eigenvalue λ_0, assumed to be simple. Let x, y be left and right eigenvectors associated with λ_0, of unit norm and with $x^*y > 0$. We still have a further degree of freedom in scaling x and y, i.e. choosing the argument of the complex numbers $\alpha = u^*x$ and $\beta = v^*y$ of fixed modulus.

— If $|\alpha| \geq |\beta|$, then we scale x such that α is real and positive. Since we require $x^*y > 0$, this also determines y uniquely.

— If $|\alpha| < |\beta|$, then we scale y such that β is real and positive. Since we require $x^*y > 0$, this also determines x uniquely.

With this particular scaling, we consider, for $0 \leq t \leq 1$, a family of matrices

$$B(t) = A + \varepsilon p(t)q(t)^*, \qquad 0 \leq t \leq 1, \tag{2.9}$$ eq:B

that interpolates between $A+\varepsilon uv^*$ at $t=0$ and $A+\varepsilon xy^*$ at $t=1$:

$$p(t) = \frac{tx+(1-t)u}{\|tx+(1-t)u\|}, \quad q(t) = \frac{ty+(1-t)v}{\|ty+(1-t)v\|}. \tag{2.10}$$ eq:xyt

The following lemma will allow us to formulate a rank-1 iteration with monotonically increasing $\mathrm{Re}\,\lambda_k$.

lem:loc-mon **Lemma 2.4 (Monotonicity near $t=0$).** *Let $B(t)$, $0 \leq t \leq 1$, be defined as above with the stated scaling of the eigenvectors. Let $\lambda(t)$, $0 \leq t \leq 1$, be the continuous path of eigenvalues of $B(t)$ with $\lambda(0) = \lambda_0$. If λ_0 is a simple eigenvalue of $A+\varepsilon uv^*$, then $\lambda(t)$ is differentiable at 0 and*

$$\mathrm{Re}\,\dot\lambda(0) \geq 0.$$

The inequality is strict except in the following two cases:

1. *$\alpha = \beta = 1$;*
2. *α and β are both real, of equal modulus and opposite sign.*

Proof. By the standard perturbation theory for simple eigenvalues we have, with $\kappa = 1/(x^*y) > 0$,

$$\dot\lambda(0) = \frac{x^*\dot B(t)y}{x^*y} = \varepsilon\kappa\Big(x^*\big(\dot p(0)q(0)^* + p(0)\dot q(0)^*\big)y\Big).$$

We find $p(0) = u$, $q(0) = v$ and

$$\dot p(0) = (x-u) - u\mathrm{Re}(u^*(x-u)) = x - u\mathrm{Re}(u^*x) = x - u\mathrm{Re}\,\alpha, \quad \dot q(0) = y - v\mathrm{Re}\,\beta$$

This yields

$$\begin{aligned}\mathrm{Re}\,\dot\lambda(0) &= \varepsilon\kappa \mathrm{Re}\Big(x^*(x-u\mathrm{Re}\,\alpha)v^*y + x^*u(y-v\mathrm{Re}\,\beta)^*y\Big)\\ &= \varepsilon\kappa\big(\mathrm{Re}\,\beta - \mathrm{Re}\,\alpha\,\mathrm{Re}(\overline{\alpha}\beta) + \mathrm{Re}\,\alpha - \mathrm{Re},\beta\,\mathrm{Re}(\overline{\alpha}\beta)\big),\end{aligned}$$

that is,

$$\mathrm{Re}\,\dot\lambda(0) = \varepsilon\kappa\left(1-\mathrm{Re}(\overline{\alpha}\beta)\right)\left(\mathrm{Re}\,\alpha + \mathrm{Re}\,\beta\right). \tag{2.11}$$

dot-lambda-ab

With our scaling, the right-hand side is positive except in Case 1. oder 2., where it vanishes. □

```
Algorithm 5: Rank-1 iteration: monotone version
  Data: Matrix $A$, $\varepsilon > 0$, tol a given positive tolerance
  Result: $r \le \alpha_\varepsilon(A), x, y$
  begin
      Compute $x_0$ and $y_0$ left and right eigenvectors to the rightmost eigenvalue of $A$, both
        normalized to unit norm and with $x_0^* y_0 > 0$.
      Let $r_{-1} = -\infty$
      Compute $x_0, y_0$, left and right eigenvectors to the rightmost eigenvalue $\lambda_0$ of $A$, with
        $x_0, y_0$ of unit norm such that $x_0^* y_0 > 0$
      Set $r_0 = \mathrm{Re}(\lambda_0)$
      Set $k = 0$
      while $r_k - r_{k-1} >$ tol do
          Compute $x_{k+1}, y_{k+1}$, left and right eigenvectors to the rightmost eigenvalue $\lambda_{k+1}$
            of $A + \varepsilon E_k$, $E_k = x_k y_k^*$, with $x_k, y_k$ of unit norm such that $x_k^* y_k > 0$
          Set $\lambda = \lambda_{k+1}$
          Set $t = 1$
          repeat
              Set $t = t/2$
              Compute $x(t), y(t)$ according to (2.10)
              Compute $\lambda(t)$
          until $\mathrm{Re}\lambda > \mathrm{Re}\lambda_k$
          Set $x_{k+1} = x(t)$, $y_{k+1} = y(t)$
          Set $k = k + 1$
          Set $r_k = \mathrm{Re}(\lambda_k)$
      Set $r = r_k, x = x_k, y = y_k$
      Halt
```

alg_GOM

Lemma 2.4 guarantees that Re $\lambda(t) >$ Re $\lambda(0)$ for sufficiently small t. Hence the idea is that to perform zero or more times the computation of the rightmost eigenvalue $\lambda(t)$ of (2.9)-(2.10) until

$$\mathrm{Re}\,\lambda(t) > \mathrm{Re}\,\lambda(0) + \frac{t}{2}\,\mathrm{Re}\,\dot\lambda(0)$$

with Re $\dot\lambda(0)$ given by (2.11), replacing t by $t/2$ until the inequality is fulfilled.

In the kth iteration step, starting from $(u, v) = (u_k, v_k)$ and $\lambda(0) = \lambda_k$, we determine in this way $t > 0$ such that Re $\lambda(t)$ satisfies the above condition and then set $\lambda_{k+1} = \lambda(t)$ and $(u_{k+1}, v_{k+1}) = (p(t), q(t))$. The sequence of the real parts of the eigenvalues Re λ_k is then monotonically increasing. In more detail, the variant is formulated in Algorithm 5.

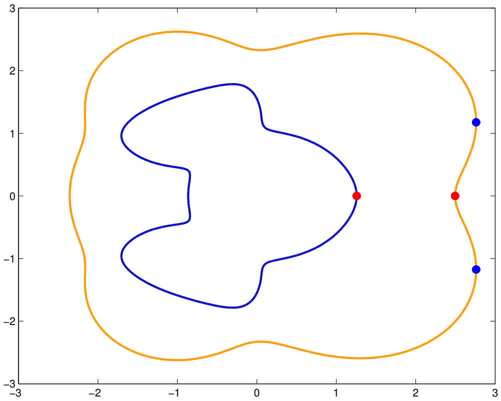

fig:extsv

Fig. 2.3. The outer curve is the boundary of the pseudospectrum $\Lambda_\varepsilon(A)$ and the inner curve is the ε-level set of $\sigma_{n-1}(A - zI)$.

th:convmon

Theorem 2.5 (Convergence of the monotone rank-1 iteration). *If the iteration sequence stays away from Case 2 in Lemma 2.4, then the monotone rank-1 iteration converges to a stationary point of the eigenvalue optimization problem* (2.8)*, i.e., the limits*

$$\lambda = \lim_{k\to\infty} \lambda_k, \quad u = \lim_{k\to\infty} u_k, \quad v = \lim_{k\to\infty} v_k$$

exist, and the stationarity condition (2.7) *is satisfied. In particular,* ε *is a singular value of* $A - \lambda I$ *with left and right singular vectors* u *and* v*. If* ε *is the smallest singular value, then* $\lambda \in \partial \Lambda_\varepsilon(A)$ *with horizontal rightward outer normal.*

We note that near a stationary point (2.7), where $\alpha = \beta = 1$, Case 2 in Lemma 2.4 cannot occur.

Proof. Since the sequence $(\mathrm{Re}\,\lambda_k)$ is monotonically increasing and bounded, it converges. This implies that in the limit, both sides of (2.11) are zero, and hence one of the two cases 1 or 2 in Lemma 2.4 must avail in the limit. By assumption, we have excluded the exceptional Case 2. In the remaining Case 1, $\alpha = \beta = 1$ in the limit, i.e. $u = x$ and $v = y$ in the limit, and hence the iteration converges and the stationarity condition (2.7) is fulfilled. As noted before, this implies the further statements. □

Figure 2.3 shows the ε-pseudospectrum of the matrix A in (1.36) for $\varepsilon = 1$ as well as the ε-level set of $\sigma_{n-1}(A - zI)$ (the inner curve).

The locally rightmost points to which the algorithms may converge are given by the 4 points emphasized. The blue ??? ones are locally attractive while the red ??? ones are locally unstable.

III.2.3 Discretized rank-1 matrix differential equation

A different iteration on rank-1 matrices results from the rank-1 projected gradient system of Section II.1.7 for the minimization function $f(\lambda, \overline{\lambda}) = -\frac{1}{2}(\lambda + \overline{\lambda}) = -\mathrm{Re}\,\lambda$ after discretization as in Section II.1.8 (see Algorithm II.1), as was first proposed similarly by Guglielmi & Lubich (2011), though with a different time stepping method. The so obtained rank-1 iteration yields a sequence of rank-1 matrices $E_k = u_k v_k^*$ of unit norm and a sequence of eigenvalues $\lambda_k \in \Lambda_\varepsilon(A)$ of $A + \varepsilon E_k$ with monotonically growing real part, which converges to a stationary point (2.7); see Theorems II.1.15 and Remark II.1.16, and also Lemma II.1.18 for the splitting method. The computational cost per step is essentially the same as in the rank-1 iterations of the previous subsection. A numerical example was already presented in Section II.1.8.

Conceptually, the approach of first deriving a suitable differential equation and then using an adaptive time-stepping to arrive at a stationary point is different from directly devising an iteration. Different tools are available and made use of in the two approaches. For example, the tangent space of the manifold of rank-1 matrices is a natural concept in the time-continuous setting though not so in the time-discrete setting. This enhanced toolbox results in efficient algorithms that would not be obtained from a purely discrete viewpoint.

III.2.4 Acceleration by a subspace method

Kressner and Vandereycken (2014) proposed a subspace method to accelerate the basic rank-1 iteration described in Subsection III.2.2. For the subspace expansion, they essentially do a step of Algorithm 4 and add the obtained eigenvector to the subspace. Along the way they compute orthonormal bases $V_k \in \mathbb{C}^{n\times k}$ of nested subspaces. A key element is the computation of the rightmost point of the ε-pseudospectrum of the rectangular matrix pencil $AV_k - \lambda V_k$ in place of $A - \lambda I$,

$$\Lambda_\varepsilon(AV_k, V_k) = \{\lambda \in \mathbb{C} \,:\, \sigma_{\min}(AV_k - \lambda V_k) = \varepsilon\}.$$

These pseudospectra are nested: $\Lambda_\varepsilon(AV_k, V_k) \subset \Lambda_\varepsilon(AV_{k+1}, V_{k+1}) \subset \Lambda_\varepsilon(A)$. The rightmost point of $\Lambda_\varepsilon(AV_k, V_k)$ is computed by a variant of the criss-cross algorithm. The basic algorithm is given in Algorithm 6.

Algorithm 6: Subspace method.

```
Data: Matrix $A$, $\varepsilon > 0$
Result: Approximation $\mu_\alpha$ to a locally rightmost point of $\Lambda_\varepsilon(A)$.
begin
    Compute the rightmost eigenvalue $\lambda_0$ and normalized right eigenvector $y_0$ of $A$.
    Set $\widehat{V}_1 = y_0$.
    for $k = 1, 2, \dots$ until converged do
        Compute the rightmost point $\mu_k$ of $\Lambda_\varepsilon(AV_k, V_k)$.
        Compute left/right singular vectors $u_k$ and $v_k$ to $\sigma_{\min}(A - \mu_k I)$. Set
          $E_k = -u_k v_k^*$.
        Compute the rightmost eigenvalue $\lambda_k$ and right eigenvector $y_k$ of $A + \varepsilon E_k$.
          Compute $V_{k+1} =$ orth$([V_k, y_k])$.
    Set $\mu_\alpha = \mu_k$.
```

alg:KV-subspace

Kressner and Vandereycken (2014) show that the sequence (μ_k) grows monotonically, as a consequence of the growth of the nested subspaces. A simplified version of the algorithm, where the right singular vector v_k instead of the right eigenvector y_k is added to the subspace, is shown to converge locally superlinearly to the pseudospectral abscissa.

III.3 Tracing the boundary of the pseudospectrum

sec:ps-tracing

In this section we describe two algorithms for boundary tracing. While there exist path-following methods to obtain pseudospectral contours (e.g. those implemented in Eigtool), the computation becomes expensive for large matrices. Here we use instead the low-rank structure of the extremal perturbations, which allows us to treat also large sparse matrices efficiently; see Section II.1.6.

We present two algorithms. The first algorithm, to which we refer as the *tangential/transversal algorithm*, makes use of a combination of the differential equation (II.1.25) and a similar differential equation that moves eigenvalues horizontally to the boundary. The second algorithm, which we call the *ladder algorithm*, aims to compute, for an iteratively constructed sequence of points outside the ε-pseudospectrum, the corresponding nearest points in the ε-pseudospectrum. Both algorithms require repeatedly the computation of the eigenvalue of a rank-1 perturbation to A that is nearest to a given complex number. This can be done efficiently also for large matrices, using inverse power iteration combined with the Sherman-Morrison formula.

The ladder algorithm extends readily to real and structured pseudospectra, for which the few algorithms proposed in the literature (as implemented in Seigtool) are restricted to just a few structures and turn out to be extremely demanding from a computational point of view.

III.3.1 Tangential/transversal algorithm

The algorithm alternates between a time step for the system of differential equations (II.1.25) and the following system of differential equations for vectors $u(t)$ and $v(t)$ of unit norm. This second system is a simplified variant of (II.1.25) for $G = -xy^*$, where x and y are left and right eigenvectors, respectively, both of unit norm and with $x^*y > 0$, of an eigenvalue λ of the rank-1 perturbed matrix $A + \varepsilon uv^*$:

$$\begin{aligned} \dot{u} &= (I - uu^*)xy^*v \\ \dot{v} &= (I - vv^*)yx^*u. \end{aligned} \tag{3.1}$$

ode-hor

The system preserves the unit norm of u and v, since $u^*\dot{u} = 0$ and $v^*\dot{v} = 0$. As we show in the next lemma, the system has the property that for a path of simple eigenvalues $\lambda(t)$ of $A + \varepsilon u(t)v(t)^*$, the derivative $\dot{\lambda}(t)$ is real and positive, continuing to a stationary point where $A - \lambda I$ has the singular value ε. By Theorem 1.2 it therefore stops at the boundary $\partial\Lambda_\varepsilon(A)$ when ε is the smallest singular value. While in theory, a trajectory might stop at an interior point where the singular value ε is not the smallest one, this appears to be an unstable case that is not observed in computations.

lem:hor-motion **Lemma 3.1 (Horizontal motion of an eigenvalue).** *Along a path of simple eigenvalues $\lambda(t)$ of $A + \varepsilon u(t)v(t)^*$, where u, v of unit norm solve* (3.1)*, we have that*

$$\dot{\lambda}(t) \textit{ is real and positive for all } t.$$

In the limit $\lambda_\star = \lim_{t\to\infty} \lambda(t)$, the matrix $A - \lambda_\star I$ has the singular value ε.

Proof. The standard perturbation theory of eigenvalues shows that

$$\dot{\lambda} = \frac{x^* \frac{d}{dt}(A + \varepsilon uv^*)\, y}{x^*y} = \varepsilon\, \frac{x^*\,(\dot{u}v^* + u\dot{v}^*)\, y}{x^*y}.$$

With $\alpha = u^*x$ and $\beta = v^*y$, we obtain

$$\dot{\lambda} = \frac{\varepsilon}{x^*y}\Big(|\alpha|^2 \cdot \|y - \beta v\|^2 + |\beta|^2 \cdot \|x - \alpha u\|^2 \Big) \in \mathbb{R},\ \geq 0.$$

In a stationary point of (3.1), u and x are collinear, and so are v and y. It follows that $uv^* = \mathrm{e}^{\mathrm{i}\theta}xy^*$ for some real θ. We thus have

$$(A + \varepsilon \mathrm{e}^{\mathrm{i}\theta} xy^*)y = \lambda y, \qquad x^*(A + \varepsilon \mathrm{e}^{\mathrm{i}\theta} xy^*) = \lambda x^*,$$

or equivalently

$$(A - \lambda I)y = \varepsilon \mathrm{e}^{\mathrm{i}\theta} x, \qquad (A - \lambda I)^* \mathrm{e}^{\mathrm{i}\theta} x = \varepsilon y,$$

which states that $\mathrm{e}^{\mathrm{i}\theta} x$ and y are left and right singular vectors to the singular value ε of $A - \lambda I$. □

While (3.1) moves eigenvalues horizontally to the right, the differential equation (II.1.25) moves an eigenvalue on the boundary along a path that starts tangentially to the boundary, as is shown by the following lemma.

lem:tang-motion

Lemma 3.2 (Tangential motion of an eigenvalue from the boundary). *Let $\lambda_0 \in \partial \Lambda_\varepsilon(A)$ be on a smooth section of the boundary, with outer normal $\mathrm{e}^{\mathrm{i}\theta}$ at λ_0 for $0 < |\theta| \le \pi/2$. Let u_0 and v_0 of unit norm be such that $A + \varepsilon u_0 v_0^*$ has λ_0 as a simple eigenvalue. Let $u(t)$ and $v(t)$ be solutions of the system of differential equations (II.1.25) with $G = -xy^*$ with initial values u_0 and v_0. Then, the path of eigenvalues $\lambda(t)$ of $A + \varepsilon u(t)v(t)^*$ with $\lambda(0) = \lambda_0$ has $\dot\lambda(0) \ne 0$ and*

$$\dot\lambda(0) \text{ is tangential to } \partial\Lambda_\varepsilon(A) \text{ at } \lambda_0.$$

Proof. By Lemma 1.4, we have $u_0 v_0^* = \mathrm{e}^{\mathrm{i}\theta} x_0 y_0^*$. We then find, inserting (II.1.25) for $\dot u$ and $\dot v$,

$$\begin{aligned}
\dot\lambda(0) &= \frac{x_0^* \, \frac{d}{dt}\big|_{t=0} (A + \varepsilon u(t) v(t)^*) \, y_0}{x_0^* y_0} \\
&= \frac{\varepsilon}{x_0^* y_0} \, x_0^* \big(\dot u(0) v(0)^* + u(0) \dot v(0)^* \big) y_0 \\
&= \frac{\varepsilon}{x_0^* y_0} \, \mathrm{i} \, \mathrm{Im}(u_0^* x_0 y_0^* v_0) x_0^* u_0 v_0^* y_0 \\
&= \frac{\varepsilon}{x_0^* y_0} \, \mathrm{i} \, \mathrm{Im}(\mathrm{e}^{-\mathrm{i}\theta}) \, \mathrm{e}^{\mathrm{i}\theta}.
\end{aligned}$$

Since $x_0^* y_0 > 0$, we find that $\dot\lambda(0)$ points into the tangential direction $-\mathrm{i}\,\mathrm{e}^{\mathrm{i}\theta} \operatorname{sign}(\theta)$. □

Description of the algorithm. Lemmas 3.1 and 3.2 motivate the following algorithm for tracing the boundary of the ε-pseudospectrum $\Lambda_\varepsilon(A)$. Suppose that λ_0 is a simple eigenvalue of $A + \varepsilon u_0 v_0^*$ lying on a smooth section of the boundary $\partial\Lambda_\varepsilon(A)$, with *a priori* unknown outer normal $\mathrm{e}^{\mathrm{i}\theta_0}$. By Lemma 1.4, $u_0 v_0^* = \mathrm{e}^{\mathrm{i}\theta_0} x_0 y_0^*$, where x_0 and y_0 are left and right eigenvectors of $A + \varepsilon u_0 v_0^*$, both of unit norm and with $x_0^* y_0 > 0$. This fact allows us to determine the outer normal as

$$\mathrm{e}^{\mathrm{i}\theta_0} = \frac{|u_0^* v_0|}{u_0^* v_0}. \tag{3.2}$$

normal-theta

We first consider the tangential differential equation (II.1.25) with $G = -xy^*$ for the rotated matrix $\mathrm{i}\mathrm{e}^{-\mathrm{i}\theta_0} A$, which leads us to the case $\theta = \pi/2$ in Lemma 3.2. We make

Algorithm 7: Tangential/transversal algorithm for tracing the boundary of the ε-pseudospectrum

```
Data: Matrix $A$, initial vectors $u, v$ of unit norm such that the eigenvalue $\lambda$ of $A + \varepsilon uv^*$
      lies on $\partial\Lambda_\varepsilon(A)$, stepsize $h$, number $N$ of desired boundary points, tol a given
      positive tolerance
Result: vector $\Gamma$ of $N$ consecutive boundary points
begin
    for i=1,...,N do
        Set $\zeta = u^* v/|u^* v|$
        Compute the approximate solution $\widetilde{u}_1, \widetilde{v}_1$ of (II.1.25) for the rotated matrix $\mathrm{i}\zeta A$
          with initial data $\mathrm{i}\zeta u, v$ doing a single normalized Euler step of size $h$
        Let $u_0 = -\mathrm{i}\widetilde{u}_1$, $v_0 = \widetilde{v}_1$
        for $k = 1, 2, \dots$ until convergence do
            Compute the approximate solution $u_k, v_k$ of (3.1) for the rotated matrix $\zeta A$
              with adaptive stepsize $h_k$ (as in Section II.1)
            Compute the rightmost eigenvalue $\lambda_k$ of $\zeta A + \varepsilon u_k\, v_k^*$
        Set $u = u_k/\zeta$, $v = v_k$
        Store $\lambda_k/\zeta$ into $\Gamma$
```

alg_1

a time step of stepsize h with the splitting method presented in Chapter II, followed by normalization of u and v; at $t_1 = t_0 + h$ this yields a rank-1 matrix $\widetilde{E}_1 = \tilde{u}_1 \tilde{v}_1^*$ of unit norm. (Note that this step does not require to actually compute the rotated matrix.) Since $\dot{\lambda}(t_0)$ is tangential to the boundary $\partial\Lambda_\varepsilon(A)$ at λ_0, the eigenvalue $\tilde{\lambda}_1$ of $A + \varepsilon\widetilde{E}_1$ lies in $\Lambda_\varepsilon(A)$ and is $O(h^2)$ close to the boundary.

With initial values $\tilde{u}_1, \tilde{v}_1$ we then consider the differential equation (3.1) for the rotated matrix $\mathrm{e}^{-\mathrm{i}\theta_0} A$ in order to reach the boundary with a horizontal trajectory. We integrate (3.1) until we stop at a stationary point $E_1 = u_1 v_1^*$. There, a singular value of $A - \lambda I$ equals ε (by Lemma 3.1), which takes us to the boundary $\partial\Lambda_\varepsilon(A)$ provided we start sufficiently close to it.

We then continue the above alternating integration from $\lambda_1 \in \partial\Lambda_\varepsilon(A)$ and the associated vectors u_1, v_1.

The algorithm as described computes a part of $\partial\Lambda_\varepsilon(A)$ to the right of λ_0. To go to the left, we change the direction of time in the tangential differential equation (II.1.25), i.e., we take a negative stepsize h.

Numerical experiment.

To be discussed.

We consider again the 6×6 random matrix (1.36) with $\varepsilon = 1$.

We obtain the result in Figure 3.1.

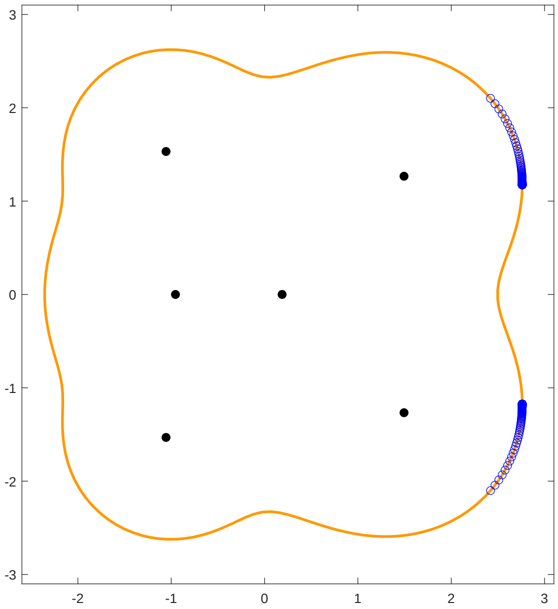

fig:tt

Fig. 3.1. In the left picture the ε-pseudospectrum and a section of its boundary (in blue) determined by the tangential transversal Algorithm 7 for the random matrix (1.36) and the value $\varepsilon = 1$.

III.3.2 Ladder algorithm

As in the previous algorithm, let $\lambda_0 \in \partial\Lambda_\varepsilon(A)$ be a simple eigenvalue of $A + \varepsilon u_0 v_0^*$ (with u_0 and v_0 of unit norm) that lies on a smooth section of $\partial\Lambda_\varepsilon(A)$, with outer normal $\mathrm{e}^{\mathrm{i}\theta_0}$ at λ_0 obtained from (3.2).

With a small distance $\delta > 0$, we define the nearby point μ_0 on the straight line normal to $\partial\Lambda_\varepsilon(A)$ at λ_0,

$$\mu_0 = \lambda_0 + \delta \mathrm{e}^{\mathrm{i}\theta_0}.$$

We add a tangential component, either to the left $(+)$ or to the right $(-)$,

$$\mu_1 = \mu_0 \pm \mathrm{i}\delta \mathrm{e}^{\mathrm{i}\theta_0}.$$

We then apply the eigenvalue optimization algorithm of Section II.1 for the function

Algorithm 8: Ladder algorithm for tracing the boundary of the ε-pseudospectrum

```
Data: Matrix $A$, initial vectors $u, v$ of unit norm such that the eigenvalue $\lambda$ of $A + \varepsilon uv^*$
      lies on $\partial\Lambda_\varepsilon(A)$, step size $\delta$, number $N$ of desired boundary points, tolerance
      $\mathrm{tol} > 0$
Result: vector $\Gamma$ of $N$ consecutive boundary points
begin
    for i=1,...,N do
        Set $\zeta = u^* v/|u^* v|$
        Set $\mu = \lambda + (1 - \mathrm{i})\delta/\zeta$
        Compute the nearest point to $\mu$ on $\partial\Lambda_\varepsilon(A)$ by the rank-1 eigenvalue optimization
          algorithm of Section II.1 for minimizing the function $f(\lambda, \overline{\lambda}) = |\lambda - \mu|^2$, with $u$
          and $v$ as the starting iterate and with the tolerance parameter tol. This yields an
          update of $u$, $v$ and $\lambda$.
        Store $\lambda$ into $\Gamma$.
```

alg_2

$$f(\lambda, \overline{\lambda}) = (\lambda - \mu_1)(\overline{\lambda} - \overline{\mu_1}) = |\lambda - \mu_1|^2,$$

choosing $E_0 = u_0 v_0^*$ as the starting iterate. That algorithm aims to compute u_1 and v_1 of unit norm such that $A + \varepsilon u_1 v_1^*$ has the boundary point $\lambda_1 \in \partial\Lambda_\varepsilon(A)$ nearest to μ_1 as an eigenvalue. At the point λ_1 we have the outer normal $(\mu_1 - \lambda_1)/|\mu_1 - \lambda_1|$.

We continue from λ_1 and u_1, v_1 in the same way as above, constructing a sequence λ_k ($k \geq 1$) of points on the boundary of the pseudospectrum $\Lambda_\varepsilon(A)$ with approximate spacing δ.

Remark 3.3. The points λ_k and μ_k and the straight lines between them form the "rope ladder" along the boundary of the pseudospectrum to which the name of the algorithm refers. We climb up or down on this ladder to construct the sequence of boundary points.

Remark 3.4. If the curvature of the boundary is larger than $1/\delta$, then it may happen that μ_k gets to lie inside $\Lambda_\varepsilon(A)$, and the algorithm will find $\lambda_k = \mu_k$. In such a situation, the step size δ needs to be reduced.

Numerical experiment. We apply the ladder algorithm to the matrix A in (II.1.36), for which the curvature of the ε-pseudospectrum is moderate. Making use of a circular arc external to the ε-pseudospectral contour allows for an accurate computation of a section of the boundary close to the point of maximum modulus.

Algorithm 8 produces the blue curve in Figure **??** which is superimposed to the boundary of the ε-pseudospectrum (computed to high accuracy by Eigtool by Wright, 2002). The average number of steps for each horizontal computation is 5.5 for an accuracy tolerance $\mathrm{tol} = 10^{-7}$.

III.3.3 Tracing the boundary of structured pseudospectra

The ladder algorithm is readily extended to compute sections of the boundary of the structured ε-pseudospectrum of a given matrix. The only available tool appears to be Seigtool

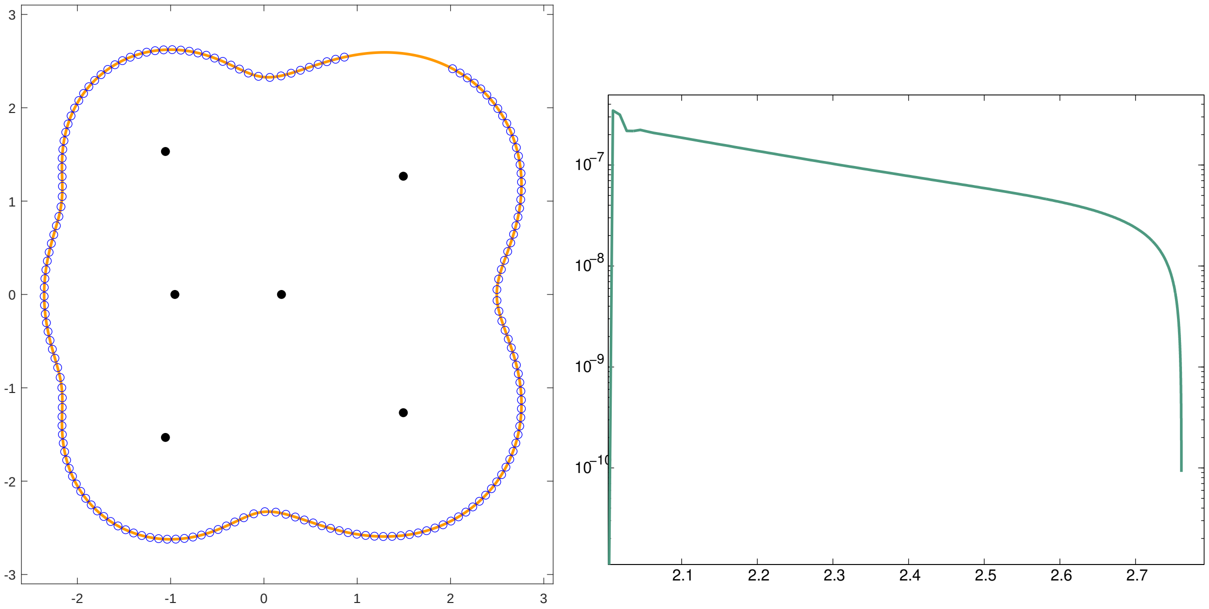

Fig. 3.2. In the left picture the ε-pseudospectrum (in orange), computed by Eigtool with $\varepsilon = 1$ of the matrix A in (II.1.36) and its boundary points (in blue) determined by the ladder algorithm (Algorithm 8). Right picture: $\varepsilon - \sigma_n(A - \lambda I)$, in logarithmic scale, as a function of Re(λ) for few points computed by the algorithm. (represented in blue in the left picture) **??**)

fig:1

by Kressner et al. **?**, which, however, treats only a limited number of structures: the real , the Hamiltonian and the skew Hamiltonian structure. Instead, pattern structures like Toeplitz, Hankel, banded and in general sparse matrices are not included and there do not exist specific algorithms to draw the associated structured pseudospectra.

Let us consider here two illustrative examples, in which we apply a structured version of the ladder algorithm.

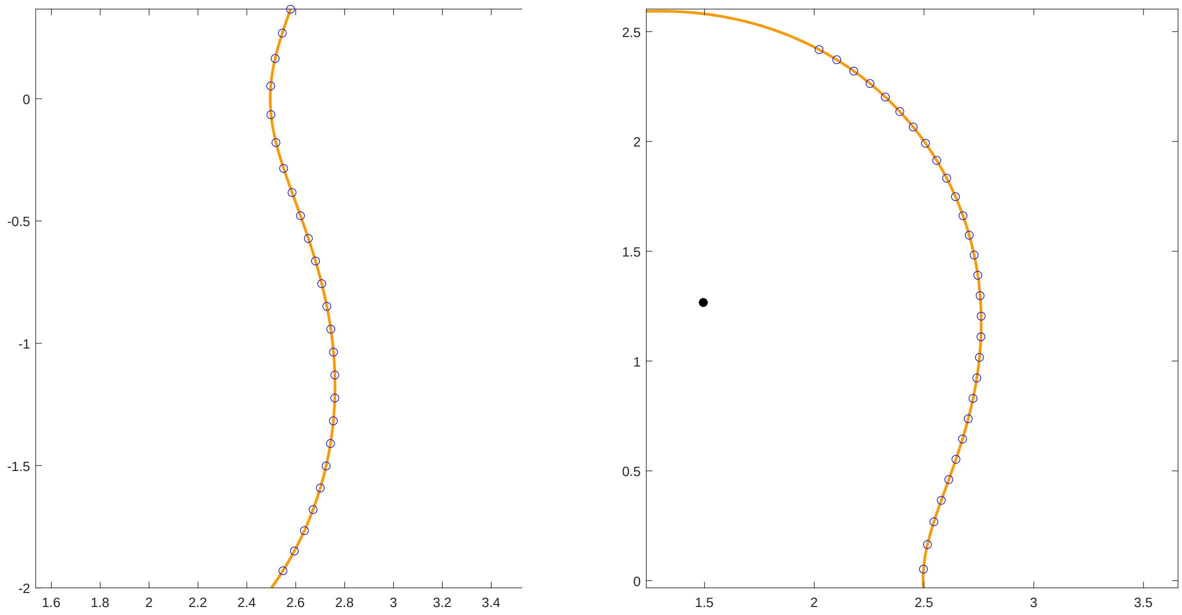

Fig. 3.3. Sections of the boundary of A computed by the ladder algorithm. fig:1

III.3.4 Real pseudospectra

We consider again the real matrix in (II.1.36) and $\varepsilon = 1$ and are interested in computing the boundary of the real ε-pseudospectrum

$$\Lambda_\varepsilon^{\mathbb{R}}(A) = \{\lambda \in \mathbb{C} \colon \lambda \in \Lambda(A + \varepsilon E) \quad \text{with } E \in \mathbb{R}^{n,n}, \|E\|_F \le 1\} \tag{3.3}$$

In order to do this we apply the ladder algorithm 8, that is we construct suitable control points external to the ε-pseudospectrum and compute the closest point on its boundary.

In Figure 3.4 we show a section of the real ε-pseudospectrum of A with $\varepsilon = 1$. Red points indicate control points computed by the ladder algorithm, while blue points represent the closest boundary points of $\Lambda_\varepsilon^{\mathbb{R}}(A)$ computed by the algorithm.

In Figure 3.5 we show the whole set of computed points on $\Lambda_\varepsilon^{\mathbb{R}}(A)$ (left picture) and the real ε-pseudospectrum together to the standard complex ε-pseudospectrum. Note that in the plot of the set $\Lambda_\varepsilon^{\mathbb{R}}(A)$, there are two missing real segments, which are reasily obtained interesecting the comples ε-pseudospectrum and the real axis.

III.3.5 Hamiltonian pseudospectra

Finally we show an example of real Hamiltonian ε-pseudospectrum,

$$\Lambda_\varepsilon^{\mathrm{Ham}(\mathbb{R}^{n,n})}(H) = \{\lambda \in \mathbb{C} \colon \lambda \in \Lambda(H + \varepsilon E) \quad \text{with } E \in \mathrm{Ham}(\mathbb{R}^{n,n}), \|E\|_F \leq 1\} \tag{3.4}$$

where $\mathrm{Ham}(\mathbb{R}^{n,n})$ denotes the set of $n \times n$ (with even n) of real Hamiltonian matrices.

Let

$$H = \begin{pmatrix} 1.0 & 1.6 & 1.2 & 0.4 \\ 2.2 & -0.6 & 0.4 & -4.4 \\ -4.0 & -7.4 & -1.0 & -2.2 \\ -7.4 & 6.0 & -1.6 & 0.6 \end{pmatrix} \tag{3.5}$$

and set $\varepsilon = 0.4$.

In Figure 3.6 we plot a section of the computed structured ε-pseudospectrum (blue points) together with the control points (in red) computed by the ladder algorithm.

In Figure 3.7 we show the whole set of computed boundary points, which identify the boundary of $\Lambda_\varepsilon^{\mathrm{Ham}(\mathbb{R}^{n,n})}(H)$.

In Figure 3.8 we show the classical complex ε-pseudospectrum together with the real structured Hamiltonian ε-pseudospectrum.

III.3.6 Toeplitz pseudospectra

sec:toepps

As an illustrative case we consider the case of Toeplitz matrices, for which - as far as we know - there are no available tools to draw structured pseudospectra and compute related quantities. We indicate by $\mathcal{T}_k$ the manifold of k-diagonal Toeplitz matrices and

$$\Lambda_\varepsilon^{\mathcal{T}_k}(A) = \{\lambda \in \mathbb{C} \colon \lambda \in \Lambda(A + \varepsilon E) \quad \text{with } E \in \mathcal{T}_k, \|E\|_F \leq 1\} \tag{3.6}$$

the associated structured ε-pseudospectrum.

As an illustrative example, consider the 12×12 penta-diagonal Toeplitz matrix

$$T = T_5(s_1, s_2, d, t_1, t_2), \in \mathcal{T}_5, \tag{3.7}$$

eq:example1

that is the matrix with elements

$$\begin{aligned} a_{i,i} &= d, && i = 1, \ldots, n \\ a_{i+1,i} &= s_1, && i = 1, \ldots, n-1 \\ a_{i+2,i} &= s_2, && i = 1, \ldots, n-2 \\ a_{i-1,i} &= t_1, && i = 1, \ldots, n-1 \\ a_{i-2,i} &= t_2, && i = 1, \ldots, n-2. \end{aligned}$$

with entries $d = -0.3$, $s_1 = -0.1$, $s_2 = -0.3$, $t_1 = 2$, $t_2 = 0.5$

We wish to compute the structured ε-pseudospectrum of T for $\varepsilon = 0.6$,

$$\Lambda_\varepsilon^{\mathcal{T}_5}(T) = \{\lambda \in \mathbb{C} \colon \lambda \in \Lambda(T + \varepsilon E) \quad \text{with } E \in \mathcal{T}_5, \|E\|_F \leq 1\} \tag{3.8}$$

The black points in Figure **??** are the spectra of 10^5 perturbed matrices obtained by adding the nominal matrix (3.7) random structured perturbations with Gaussian distributed entries of Frobenius norm ε, which therefore are internal to $\Lambda_\varepsilon^{\mathcal{T}_5}(A)$.

In Figure 3.10 we plot the points in the boundary of $\Lambda_\varepsilon^{\mathcal{T}_k}(A)$ computed by the ladder algorithm.

In Figure 3.13 we plot the control points (in red) and the the boundary points (in blue) of a section of the set $\Lambda_\varepsilon^{\mathcal{T}_5}(A)$ computed by the ladder algorithm.

In Figure 3.12 we show both the structured and the unstructured pseudospectra.

III.4 Notes

Pseudospectra. The standard reference for (complex unstructured) pseudospectra is the book by Trefethen & Embree (2005). The notion of pseudospectrum was coined by Trefethen (1992), but as noted there, the concept was used before under different names. The "approximate eigenvalues" of Varah (1967) appear to be among the earliest precursors. Interpreting approximate eigenvalues or pseudospectral values as exact eigenvalues of a perturbed matrix is in the spirit of backward error analysis as pioneered by Wilkinson (1965). In one of his last papers, Wilkinson (1986) considered pseudospectra under the tentative name "fundamental inclusion domains". Pseudospectra are of interest not only for matrices but more generally for linear operators on Hilbert or Banach spaces; see Trefethen (1997) and also the concise account in the book by Davies (2007), Chapter 9.

The motivating example of robust stability of a matrix (see Subsection III.1.1) was already studied by Van Loan (1985) and Hinrichsen & Pritchard (1986a,1986b,1990). They aim at determining the minimal norm of complex, real or structured perturbations that turn a stable matrix into an unstable one. This yields the (complex, real or structured) distance to instability, or stability radius, which equals the smallest perturbation size ε for which the ε-pseudospectral abscissa becomes non-negative; this will be taken up in the next chapter. We further refer to Qiu, Bernhardsson, Rantzer, Davison, Young & Doyle (1995) for the real stability radius.

The use of (complex) pseudospectra together with the Cauchy integral formula to better understand the transient behaviour of dynamical systems was suggested by Trefethen (1992). Remark III.1.2 follows this approach, emphasizing the role of the pseudospectral abscissa. There are close connections to the Kreiss matrix theorem (Kreiss 1962, LeVeque & Trefethen 1984 and Spijker 1991) and to further stability bounds as given, e.g., by Lubich & Nevanlinna (1991), Reddy & Trefethen (1992) and van Dorsselaer, Kraaijevanger & Spijker (1993); see also Eisner (2010) and references therein.

The basic Theorem 1.2 and its proof via the distance to singularity are in essence already given by Van Loan (1985) and also appear in Wilkinson (1986). Lemma 1.4 on extremal complex perturbations is closely related to Guglielmi & Overton (2011). Lemmas 1.5 and 1.6 on extremal real and structured perturbations appear to be new.

Pseudospectra of matrix pencils $A - \lambda B$ were studied by van Dorsselaer (1997) and, e.g., Ahmad, Alam & Byers (2010), structured pseudospectra for polynomial eigenvalue problems by Tisseur & Higham (2001), and pseudospectra for rectangular matrices by Wright & Trefethen (2002).

An extension of the concept of the ε-pseudospectrum of interest in control theory is the *spectral value set* that consists of the eigenvalues of all matrices $A + B\Delta(I - D\Delta)^{-1}C$ with given system matrices A, B, C, D of compatible dimensions and with varying (complex or real or structured) matrices Δ of 2-norm at most ε. This was first considered (for $D = 0$) by Hinrichsen and Kelb (1993); see also Karow (2003) for a detailed study and the book of Hinrichsen & Pritchard (2005), Chapter 5. Apart from eigenvalues of A, the unstructured complex spectral value set contains all $s \in \mathbb{C}$ for which the transfer matrix $G(s) = C(sI - A)^{-1}B + D$ has 2-norm at least ε^{-1}; see Guglielmi, Gürbüzbalaban & Overton (2013).

Algorithms for computing the pseudospectral abscissa and radius. The criss-cross algorithm of Burke, Lewis & Overton (2003) for computing the pseudospectral abscissa relies on Byers' lemma (Lemma 2.1), due to Byers (1988). This lemma is fundamental in that it relates singular values of a matrix shifted along a line parallel to the imaginary axis to the eigenvalues of a Hamiltonian matrix. This Hamiltonian connection has been put to important and enduring use in control systems starting with the work by Boyd, Balakrishnan & Kabamba (1989); see, e.g., Grivet-Talocia & Gustavsen (2015) and references therein. We will encounter Hamiltonian eigenvalue optimization problems in later chapters. Mengi & Overton (2005) extended the criss-cross algorithm to computing the pseudospectral radius (using radial and circular searches), and Lu & Vandereycken (2017) developed a criss-cross type algorithm for computing the real pseudospectral abscissa. Benner & Mitchell (2019) studied criss-cross algorithms for computing the spectral value set abscissa and radius.

The rank-1 iteration of Guglielmi & Overton (2011) for computing the pseudospectral abscissa and radius appears to be the first algorithm that uses the low-rank property of extremal perturbations as described by Lemmas 1.4–1.6. It motivated the rank-1 projected gradient flow algorithm of Guglielmi & Lubich (2011) for computing the pseudospectral abscissa and radius, which opened up an approach to a wide range of eigenvalue optimization and matrix nearness problems, as discussed throughout this book. In particular, this low-rank matrix differential equation approach can be used to compute extremal points of complex, real and structured pseudospectra; see Guglielmi & Lubich (2013) and Section II.2 for the real pseudospectral abscissa and radius, Guglielmi, Kressner & Lubich (2015) for Hamiltonian matrices, Section II.3 for general linear structures, and Guglielmi, Kressner & Lubich (2014) for the nonlinear structure of symplectic matrices.

Subspace acceleration approaches as used by Kressner & Vandereycken (2014) also have a much wider scope than merely computing the pseudospectral abscissa; see, e.g., Kangal, Meerbergen, Mengi & Michiels (2018) and Kressner, Lu & Vandereycken (2018) for their use in other eigenvalue optimization problems.

In retrospect it appears remarkable how the modest goal of computing the pseudospectral abscissa led to the discovery of diverse classes of efficient algorithms that find use in a wide variety of other problems.

Computing the boundary of complex, real and structured pseudospectra. Trefethen (1999) gave a survey of computing complex pseudospectra (as of 1999), which was then accompanied by the software package EigTool (Wright 2002, Wright & Trefethen 2001). The basic algorithm is based on a contour plot of the smallest singular value of $A - zI$ for z on a grid. Algorithms for tracing boundary curves of complex pseudospectra were developed by Brühl (1996) and Mezher & Philippe (2002), and Bekas & Gallopoulos (2001) combined curve-tracing and grids.

Karow, Kokiopoulou & Kressner (2010) discussed computational approaches to structured pseudospectra, including real, skew-symmetric, Hermitian, and Hamiltonian perturbations. This is implemented in the software package Seigtool (Structured EigTool) based on EigTool.

Guglielmi & Lubich (2012) exploited the low-rank property of extremal perturbations in curve-tracing algorithms that do not use singular values. Those algorithms are closely related to the tangential/transversal algorithm and the ladder algorithm described in Section III.3. The ladder algorithm can also be used to trace smooth sections of the boundary of real and structured pseudospectra.

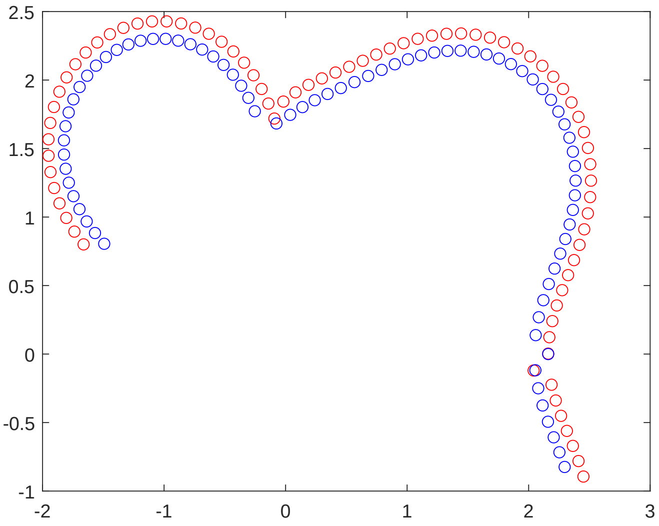

fig:R2

Fig. 3.4. Section of the boundary of the boundary of $\Lambda_\varepsilon^{\mathbb{R}}(A)$ computed by the ladder algorithm (red points indicate control points versus the computed boundary points in blue).

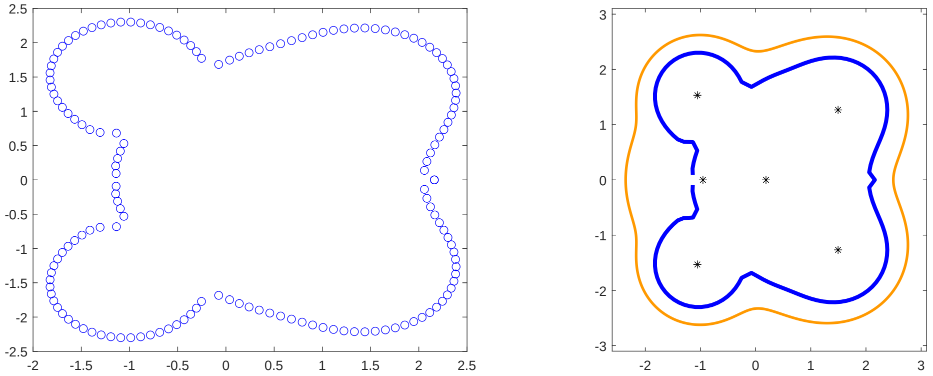

fig:R1

Fig. 3.5. Left picture: points of $\partial\Lambda_\varepsilon^{\mathbb{R}}(A)$ computed applying Algorithm 8 for the matrix (II.1.36). Right picture: real-structured pseudospectrum reconstructed by the compute points, compared to the unstructured pseudospectrum $\Lambda_\varepsilon(A)$ (in yellow). Right picture: zoom.

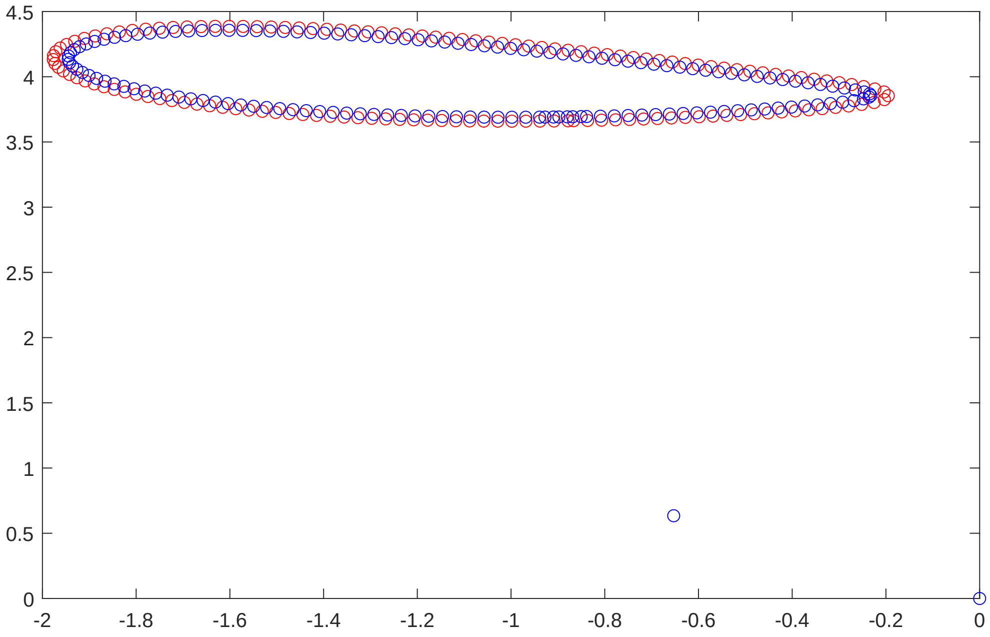

fig:H2

Fig. 3.6. A section of the computed boundary points (in blue) and control points (represented in red).

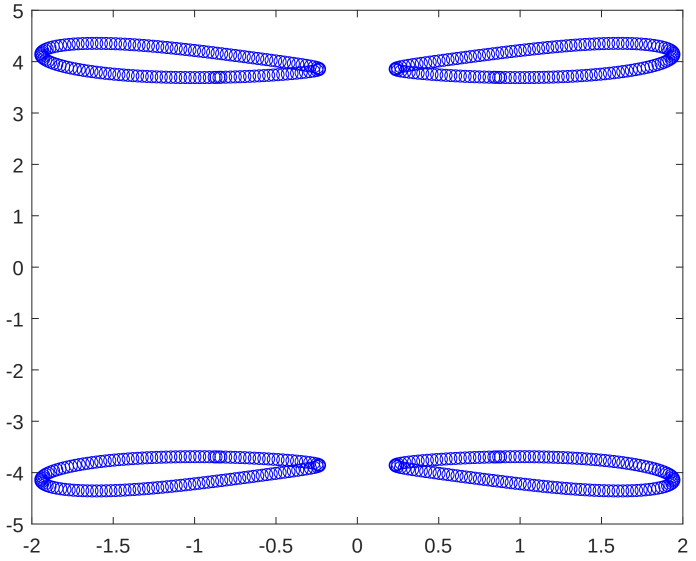

Fig. 3.7. The whole set of boundary points on $\Lambda_\varepsilon^{\mathrm{Ham}(\mathbb{R}^{n,n})}(H)$ computed by the ladder algorithm. $\boxed{\texttt{fig:H3}}$

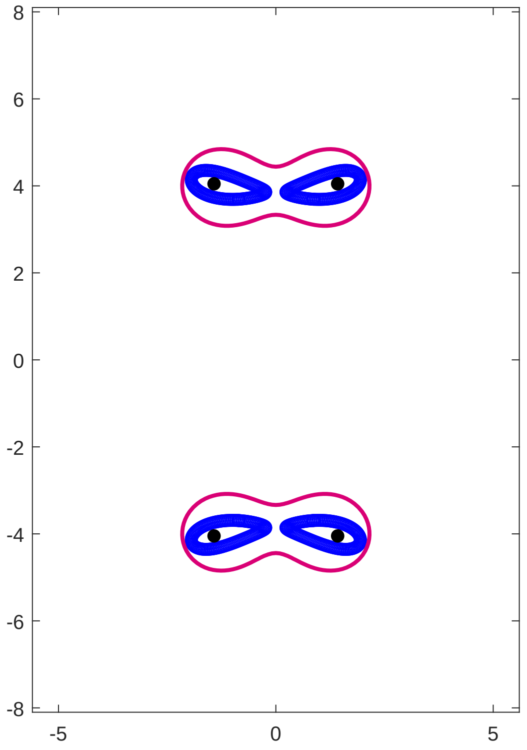

Fig. 3.8. The set $\Lambda_\varepsilon^{\mathrm{Ham}(\mathbb{R}^{n,n})}(H)$ versus the complex pseudospectrum $\Lambda_\varepsilon(H)$). $\boxed{\texttt{fig:H1}}$

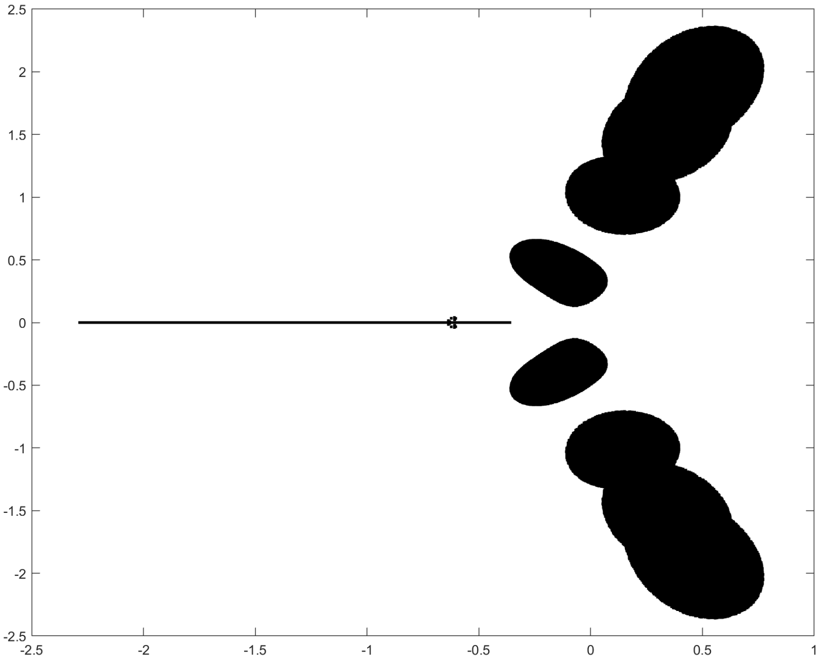

fig:T2

Fig. 3.9. A very dense sample of points internal to the structured pseudospectrum, from spectra of perturbed matrices.

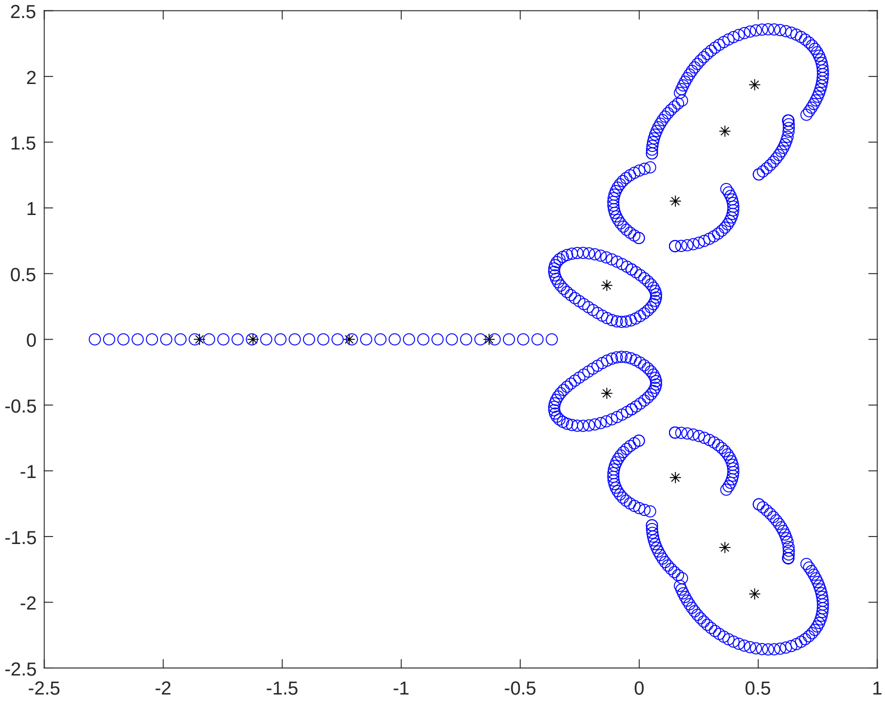

fig:T1

Fig. 3.10. Boundary points of the Toeplitz-structured pseudospectrum computed by the ladder algorithm.

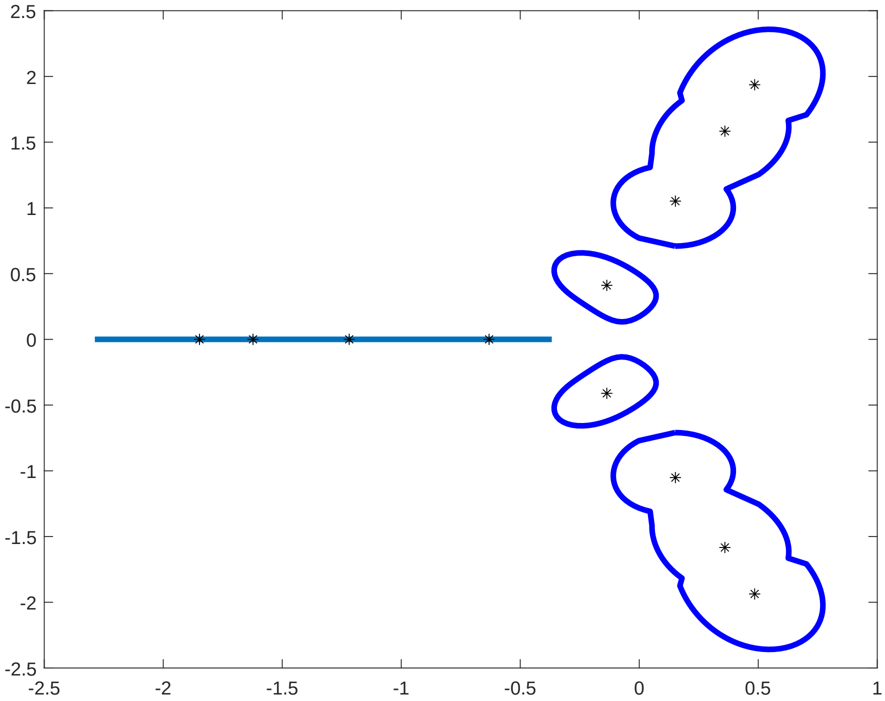

Fig. 3.11. Points on the boundary of the Toeplitz-structured pseudospectrum (blue points), with $\varepsilon = 0.6$, of matrix (3.7), computed by the ladder algorithm.

fig:T2

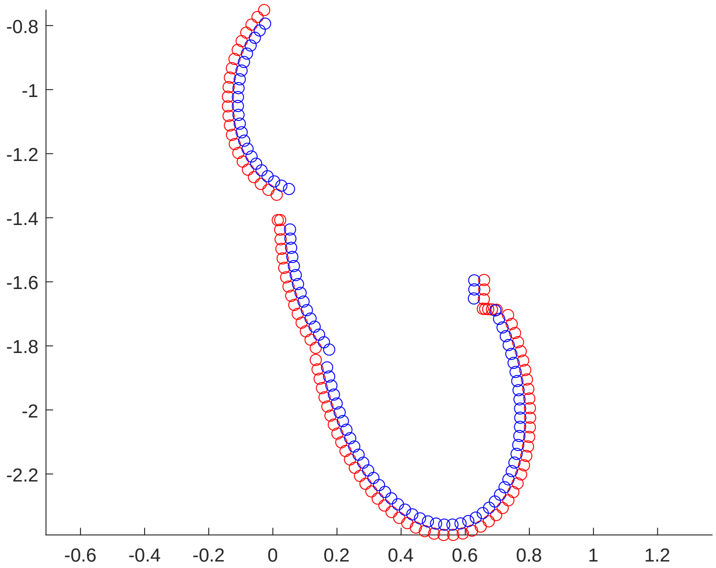

fig:T3

Fig. 3.12. A section of the computed boundary points (in blue) and control points (represented in red).

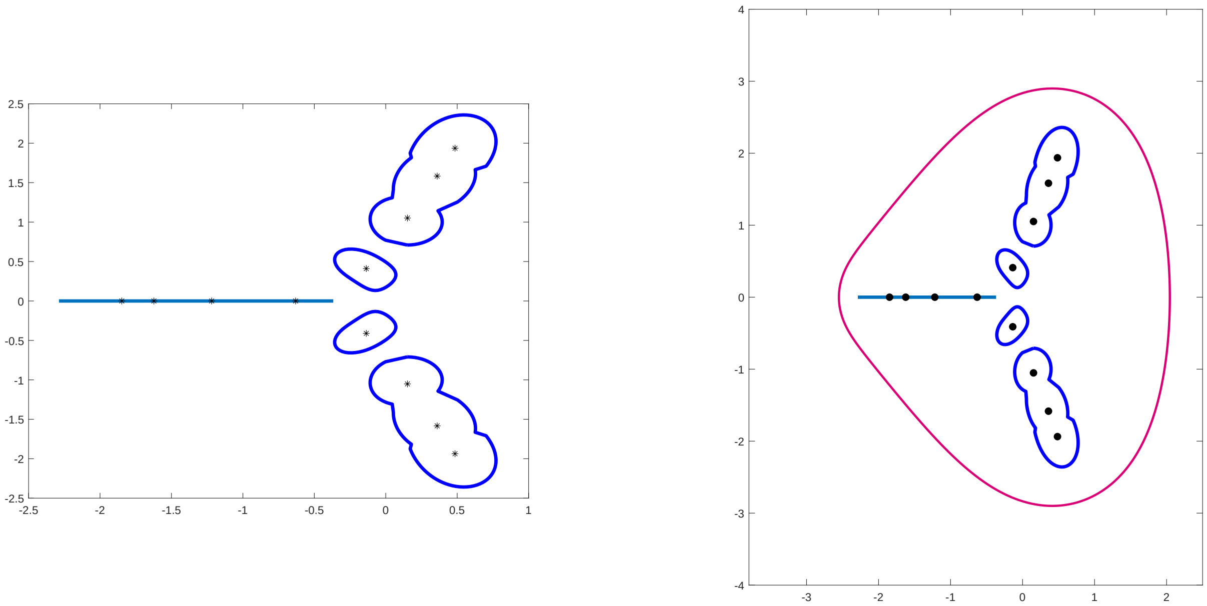

fig:T2

Fig. 3.13. Left picture: $\Lambda_\varepsilon^{\mathcal{T}_5}(A)$ reconstructed from the computed points. Right picture: the structured ε-pseudospectrum (blue curve) versus the unstructured ε-pseudospectrum (red curve).

Chapter IV.
Basic matrix nearness problems: a two-level approach

chap:two-level

In this chapter we consider a general approach to matrix nearness problems such as the following (and many more, as we will see in later chapters), where ‘nearest’ is understood as having the minimal distance in the Frobenius norm:

- **Distance to singularity.** *Given an invertible matrix, find the nearest singular matrix.* This problem is solved by a truncated singular value decomposition if general unstructured complex or real perturbations to the given matrix are admissible. In structured cases, e.g. sparse perturbations with a prescribed sparsity pattern, such a simple solution to this matrix nearness problem does not exist.
- **Distance to instability (stability radius).** *Given a matrix with all its eigenvalues having negative real part (a Hurwitz matrix), find the nearest matrix with some eigenvalue on the imaginary axis.* The perturbation to the given matrix can be restricted to be complex, real or structured. Similarly, given a matrix with all its eigenvalues having modulus smaller than 1 (a Schur matrix), find the nearest matrix with some eigenvalue on the complex unit circle.
- **Matrix stabilization.** *Given a matrix with some eigenvalues of positive real part, find the nearest matrix having no eigenvalues of positive real part.* The perturbation to the given matrix can be restricted to be complex, real or structured. Similarly, given a matrix with some eigenvalues having modulus larger than 1, find the nearest matrix with all eigenvalues in the complex unit disk.

The first two matrix nearness problems can be conveniently rephrased in terms of (complex, real or structured) pseudospectra: The distance to singularity is the smallest $\varepsilon > 0$ such that 0 is in the ε-pseudospectrum of the given matrix; the stability radius of a Hurwitz matrix is the smallest $\varepsilon > 0$ such that the ε-pseudospectrum has some point on the imaginary axis; and the stability radius of a Schur matrix is the smallest $\varepsilon > 0$ such that the ε-pseudospectrum has some point on the unit circle. More generally, for a given matrix having all eigenvalues in an open set Ω in the complex plane, the problem is to find the smallest $\varepsilon > 0$ such that the (complex, real or structured) ε-pseudospectrum has some point on the boundary of Ω.

The matrix stabilization problem is complementary to that. It is a spectral recovery problem where for a given matrix having some eigenvalues outside the closed set $\overline{\Omega}$ in the complex plane, the problem is to find the nearest (complex, real or structured) matrix that has all eigenvalues in $\overline{\Omega}$.

The two-level approach taken here uses an *inner iteration* to compute the solution of an eigenvalue optimization problem as considered in Chapter II for a fixed perturbation size ε, and then determines the optimal perturbation size $\varepsilon_\star$ in an *outer iteration*.

The algorithm is not guaranteed to find the global optimum of these nonsmooth and nonconvex optimization problems, but it computes a matrix with the desired spectral property which is locally nearest and often, as observed in our numerical experiments, has a distance close to the minimal distance. In any case it provides an upper bound to the minimal distance, and usually a very tight one. Running the algorithm with several different starting values reduces the risk of getting stuck in a local optimum.

IV.1 Problem setting and examples

We consider matrix nearness problems that are closely related to the eigenvalue optimization problems of Chapter II. We pose the problem in the structure space $\mathcal{S}$, which can be $\mathbb{C}^{n,n}$ or $\mathbb{R}^{n,n}$ or a subspace thereof as in Section II.3. For a given matrix $A \in \mathbb{C}^{n,n}$, let $\lambda(A) \in \mathbb{C}$ be a target eigenvalue of A. We again consider the smooth function $f(\lambda, \overline{\lambda})$ satisfying (II.1.2) that is to be minimized. For a prescribed real number r in the range of f we assume that

$$f(\lambda(A), \overline{\lambda}(A)) > r,$$

so that for sufficiently small $\varepsilon > 0$ we have $\phi(\varepsilon) > r$, where

$$\phi(\varepsilon) := \min_{\Delta \in \mathcal{S},\, \|\Delta\|_F = \varepsilon} f\left(\lambda\left(A + \Delta\right), \overline{\lambda}\left(A + \Delta\right)\right).$$

The objective now is to find the smallest $\varepsilon > 0$ such that $\phi(\varepsilon) = r$:

$$\varepsilon_\star = \min\bigl\{\varepsilon > 0 \,:\, \phi(\varepsilon) \le r\bigr\}. \tag{1.1}$$

eq:mnpb

Determining $\varepsilon_\star$ is a one-dimensional root-finding problem for the function ϕ that is defined by eigenvalue optimization problems as studied in Chapter II.

Example 1.1 (Stability radius of a Hurwitz matrix). With the function $f(\lambda, \overline{\lambda}) = -\frac{1}{2}(\lambda + \overline{\lambda}) = -\mathrm{Re}\,\lambda$ and $r = 0$ and the target eigenvalue $\lambda(M)$ chosen as an eigenvalue of largest real part of a matrix M, we arrive at the classical problem of computing the stability radius of a Hurwitz matrix A, i.e. a matrix with negative spectral abscissa $\alpha(A) = \max\{\mathrm{Re}\,\lambda \,:\, \lambda \text{ is an eigenvalue of } A\} < 0$:

$$\varepsilon_\star = \min\{\varepsilon > 0 \,:\, \alpha_\varepsilon^{\mathcal{S}}(A) = 0\},$$

where $\alpha_\varepsilon^{\mathcal{S}}(A) = \max\limits_{E \in \mathcal{S}, \|E\|_F = 1} \alpha(A + \varepsilon E)$ is the ε-pseudospectral abscissa with respect to the structure space $\mathcal{S}$ (see the previous chapter). For $\mathcal{S} = \mathbb{C}^{n,n}, \mathbb{R}^{n,n}$ or a strict subspace thereof, $\varepsilon_\star$ is called the complex, real or structured stability radius, respectively.

Example 1.2 (Stability radius of a Schur matrix). With $f(\lambda, \overline{\lambda}) = -\lambda\overline{\lambda} = -|\lambda|^2$ and $r = 1$ and the target eigenvalue $\lambda(M)$ chosen as an eigenvalue of largest modulus of a matrix M, we arrive at the problem of computing the stability radius of a Schur matrix A, i.e. a matrix with spectral radius $\rho(A) = \max\{|\lambda| \,:\, \lambda \text{ is an eigenvalue of } A\} < 1$:

$$\varepsilon_\star = \min\{\varepsilon > 0 \,:\, \rho_\varepsilon^{\mathcal{S}}(A) = 1\},$$

where $\rho_\varepsilon^{\mathcal{S}}(A) = \max\limits_{E \in \mathcal{S}, \|E\|_F = 1} \rho(A + \varepsilon E)$ is the ε-pseudospectral radius with respect to the structure space $\mathcal{S}$.

Example 1.3 (Structured distance to singularity). With $f(\lambda, \overline{\lambda}) = \lambda\overline{\lambda} = |\lambda|^2$ and $r = 0$ and the target eigenvalue $\lambda(M)$ chosen as an eigenvalue of smallest modulus of a matrix M, we arrive at the problem of computing the distance to singularity of an invertible matrix A under perturbations in $\mathcal{S}$,

$$\varepsilon_\star = \min\{\varepsilon > 0 \,:\, \delta_\varepsilon^{\mathcal{S}}(A) = 0\},$$

where $\delta_\varepsilon^{\mathcal{S}}(A) = \min\limits_{E \in \mathcal{S}, \|E\|_F = 1} \delta(A + \varepsilon E)$ and $\delta(M)$ is the smallest modulus of eigenvalues of a matrix M. Instead of eigenvalues of smallest modulus, we could take the smallest singular value.

IV.2 Two-level iteration

sec:two-level

Our approach is summarized by the following two-level method:

- **Inner iteration:** Given $\varepsilon > 0$, we aim to compute a matrix $E(\varepsilon) \in \mathcal{S}$ of unit Frobenius norm, such that $F_\varepsilon(E) = f\left(\lambda\left(A + \varepsilon E\right), \overline{\lambda}\left(A + \varepsilon E\right)\right)$ is minimized, i.e.

$$E(\varepsilon) = \arg \min_{E \in \mathcal{S}, \|E\|_F = 1} F_\varepsilon(E). \tag{2.1}$$

E-eps-21

- **Outer iteration:** We compute the smallest positive value $\varepsilon_\star$ with

$$\phi(\varepsilon_\star) = r, \tag{2.2}$$

eq:zero

  where $\phi(\varepsilon) = F_\varepsilon\left(E(\varepsilon)\right) = f\left(\lambda\left(A + \varepsilon E(\varepsilon)\right), \overline{\lambda}\left(A + \varepsilon E(\varepsilon)\right)\right)$.

IV.2.1 Inner iteration: Constrained gradient flow

The eigenvalue optimization problem (2.1) is precisely of the type studied in Chapter II. To compute $E(\varepsilon)$ for a given $\varepsilon > 0$, we make use of a constrained gradient system for the functional $F_\varepsilon(E)$ under the constraints of unit Frobenius norm of E and $E \in \mathcal{S}$. We use directly the gradient approach developed in Chapter II, possibly exploiting the low-rank structure of the optimizer $E(\varepsilon)$ by using a rank-constrained gradient system in the way described there.

IV.2.2 Outer iteration: Derivative for the Newton step

In the outer iteration we compute $\varepsilon_\star$, the smallest positive solution of the one-dimensional root-finding problem (1.17). This can be solved by a variety of methods, such as bisection. We aim for a locally quadratically convergent Newton-type method, which can be justified under regularity assumptions that appear to be usually satisfied (generically in the unstructured case $\mathcal{S} = \mathbb{C}^{n,n}$). If these assumptions are not met, we can always resort to bisection. The algorithm proposed in the next subsection in fact uses a combined Newton / bisection approach.

In the following, an important role is again played by the (structured) gradient $G_\varepsilon^{\mathcal{S}}(E) = \mathit{\Pi}^{\mathcal{S}}(2 f_{\overline{\lambda}} x y^*)$ as defined in Lemma II.1.1 and (II.3.4)–(II.3.5). Note that the orthogonal projection $\mathit{\Pi}^{\mathcal{S}}$ from $\mathbb{C}^{n,n}$ onto $\mathcal{S}$ is the identity map if $\mathcal{S} = \mathbb{C}^{n,n}$ and just takes the real part if $\mathcal{S} = \mathbb{R}^{n,n}$.

ass:E-eps

Assumption 2.1. For ε close to $\varepsilon_\star$ and $\varepsilon < \varepsilon_\star$, we assume the following for the optimizer $E(\varepsilon)$ of (2.1):

- The eigenvalue $\lambda(\varepsilon) = \lambda(A + \varepsilon E(\varepsilon))$ is a simple eigenvalue.
- The map $\varepsilon \mapsto E(\varepsilon)$ is continuously differentiable.
- The structured gradient $G(\varepsilon) = G_\varepsilon^{\mathcal{S}}(E(\varepsilon))$ is nonzero.

Under this assumption, the branch of eigenvalues $\lambda(\varepsilon)$ and its corresponding eigenvectors $x(\varepsilon), y(\varepsilon)$ with the scaling (II.1.5) are also continuously differentiable functions of ε in a left neighbourhood of $\varepsilon_\star$. We denote the eigenvalue condition number by

$$\kappa(\varepsilon) = \frac{1}{x(\varepsilon)^* y(\varepsilon)} > 0.$$

The following result gives us an explicit and easily computable expression for the derivative of $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon)) = f(\lambda(\varepsilon), \overline{\lambda}(\varepsilon))$ with respect to ε in terms of the (structured) gradient $G(\varepsilon)$.

phi-derivative

Theorem 2.2 (Derivative for the Newton iteration). *Under Assumption 2.1, the function ϕ is continuously differentiable in a left neighbourhood of $\varepsilon_\star$ and its derivative is given as*

$$\phi'(\varepsilon) = -\kappa(\varepsilon)\, \|G(\varepsilon)\|_F < 0. \qquad (2.3)$$

eq:dereps

Proof. By Lemma II.1.1 and (II.3.4) we obtain, indicating by $'$ differentiation w.r.t. ε and noting that $\frac{d}{d\varepsilon}(\varepsilon E(\varepsilon)) = E(\varepsilon) + \varepsilon E'(\varepsilon)$,

$$\frac{1}{\kappa(\varepsilon)} \frac{d}{d\varepsilon} F_\varepsilon(E(\varepsilon)) = \mathrm{Re}\big\langle G(\varepsilon), E(\varepsilon) + \varepsilon E'(\varepsilon) \big\rangle. \qquad (2.4)$$

eq:deriveps

By Theorem 1.5 for the unstructured complex case, by (II.2.8) for the unstructured real case, and by (II.4.14) for structured cases, we know that in the stationary point $E(\varepsilon)$, there exists a real $\mu(\varepsilon)$ such that

$$E(\varepsilon) = \mu(\varepsilon) G(\varepsilon). \qquad (2.5)$$

E-mu-G

Since $\|E(\varepsilon)\|_F = 1$ for all ε, we find $1 = |\mu(\varepsilon)|\, \|G(\varepsilon)\|_F$ (in particular $\mu(\varepsilon) \neq 0$) and

$$0 = \frac{1}{2}\frac{d}{d\varepsilon}\|E(\varepsilon)\|^2 = \mathrm{Re}\langle E(\varepsilon), E'(\varepsilon)\rangle = \mu(\varepsilon)\, \mathrm{Re}\langle G(\varepsilon), E'(\varepsilon)\rangle,$$

so that

$$\mathrm{Re}\langle G(\varepsilon), E'(\varepsilon)\rangle = 0.$$

Inserting this relation into (2.4) and using once again (2.5), we obtain

$$\frac{1}{\kappa(\varepsilon)}\, \phi'(\varepsilon) = \mathrm{Re}\big\langle G(\varepsilon), E(\varepsilon)\big\rangle = \frac{1}{\mu(\varepsilon)}\|E(\varepsilon)\|_F^2 = \frac{1}{\mu(\varepsilon)} = \mathrm{sign}(\mu(\varepsilon))\, \|G(\varepsilon)\|_F.$$

Since for $\varepsilon < \varepsilon_\star$, we have $\phi(\varepsilon) > \phi(\varepsilon_\star) = r$, and since the above formula shows that ϕ' cannot change sign, we must have $\phi'(\varepsilon) < 0$ and hence $\mu(\varepsilon) < 0$. This yields the stated result. □

IV.2.3 Outer iteration: Newton / bisection method

ewton-bisection

In view of Theorem 2.2, applying Newton's method to the equation $\phi(\varepsilon) = r$ yields the following iteration:

$$\varepsilon_{k+1} = \varepsilon_k + \frac{x(\varepsilon_k)^* y(\varepsilon_k)}{\|G(\varepsilon_k)\|_F}\left(\phi(\varepsilon_k) - r\right), \qquad (2.6)$$

CNM1

where the right-hand side uses the optimizer $E(\varepsilon_k)$ computed by the inner iteration in the k-th step.

Algorithm 9 implements a hybrid Newton / bisection method that maintains an interval known to contain the root, bisecting when the Newton step is outside the interval $[\varepsilon_{\mathrm{lb}}, \varepsilon_{\mathrm{ub}}]$.

Step **5** (in the **while** loop) gives the computational core of Algorithm 9; it implements the inner iteration and is not presented in detail since it depends on the possible structure of the matrix and on whether the low-rank structure is exploited. The inner iteration performs the algorithm to compute the extremal perturbation $E(\varepsilon_k)$, as described in Chapter II. As input to the k-th iteration we use the factors of the final matrix E computed for the previous value of ε (this explains the choice of the initial datum at step **5**).

The **while** loop after step **4** implements the outer iteration and makes use of a variable tolerance which decreases as k increases, when the method is expected to approach convergence. A typical choice of tol_0 is the norm of the difference of the first two iterates divided by 10.

The factor 10^{-2} between two subsequent tolerances is fruit of an empirical experimentation and is motivated by the fact that we expect convergence - on the average - in about $4 - 5$ iterates so that we reach the limit tolerance which is set here to 10^{-8}. These can naturally be considered as parameters of the code. Their choice here is only based on the experience with the numerical experiments we performed. When integrating numerically the gradient system with variable stepsize we are assured to fulfilll the termination

Algorithm 9: Outer iteration: Newton / bisection method

```
    Data: Matrix $A$, matrix type (real/complex, structured)
    $r > 0$, $\mathrm{tol}_0$ (initial tolerance), $k_{\max}$ (max number of iterations)
    $\varepsilon_{\mathrm{lb}}$ and $\varepsilon_{\mathrm{ub}}$ (starting values for the lower and upper bounds for $\varepsilon^\star$)
    Result: $\varepsilon_\star$ (upper bound for the stability radius)
    begin
1       Set $\lambda(0)$ target eigenvalue of $A$, $x(0)$ and $y(0)$ the corresponding left and right
          eigenvectors of unit norm with $x(0)^* y(0) > 0$.
2       Initialize $E(\varepsilon_0)$ according to the setting.
3       Initialize $\varepsilon_0$ according to the setting.
        Set $k = 0$.
4       Initialize lower and upper bounds: $\varepsilon_{\mathrm{lb}} = 0$, $\varepsilon_{\mathrm{ub}} = +\infty$.
        while $|\phi(\varepsilon_k) - \phi(\varepsilon_{k-1})| < \mathrm{tol}_k$ do
5           Compute $E(\varepsilon_k)$, $\phi(\varepsilon_k)$ by integrating the constrained gradient system with initial
              datum $E(\varepsilon_{k-1})$. (This is the inner iteration).
6           Update upper and lower bounds $\varepsilon_{\mathrm{lb}}, \varepsilon_{\mathrm{ub}}$.
7           if $\phi(\varepsilon_k) < r$ then
                Set $\varepsilon_{\mathrm{ub}} = \min(\varepsilon_{\mathrm{ub}}, \varepsilon_k)$.
            else
                Set $\varepsilon_{\mathrm{lb}} = \max(\varepsilon_{\mathrm{lb}}, \varepsilon_k)$.
8           Compute $G(\varepsilon_k)$.
9           Compute $\varepsilon_{k+1} = \varepsilon_k + \dfrac{x(\varepsilon_k)^* y(\varepsilon_k)}{\|G(\varepsilon_k)\|_F} (\phi(\varepsilon_k) - r)$.
10          Set $k = k + 1$.
11          if $\varepsilon_k \notin [\varepsilon_{\mathrm{lb}}, \varepsilon_{\mathrm{ub}}]$ then
                Set $\varepsilon_k = (\varepsilon_{\mathrm{lb}} + \varepsilon_{\mathrm{ub}})/2$.
            if $k = k_{\max}$ then
                goto 12.
            else
                Set $\mathrm{tol}_k = \max\{10^{-2}\,\mathrm{tol}_{k-1}, 10^{-8}\}$.
12      if $k < k_{\max}$ then
            Set $\varepsilon_\star = \varepsilon_k$. Return $\varepsilon_\star$.
        else
            Print *max number iterations reached*
```

alg_SR

condition $|\phi(\varepsilon_k) - \phi(\varepsilon_{k-1})| < \mathrm{tol}_k$ of the **while** loop because we are able to approximate the stationary point with prescribed accuracy. Due to the possible convergence of the inner method to a local instead of global minimum, the final value $\varepsilon_\star$ computed by Algorithm 9 might be larger than the minimal one.

IV.2.4 Outer iteration: Starting values

We make the following natural choice for the first initial datum E_0 (that is for the first value $\varepsilon = \varepsilon_0$): let G_0 be the structure-projected gradient of F_ε at $\varepsilon = 0$, i.e., $G_0 = \Pi^{\mathcal{S}}(2f_{\overline{\lambda}}xy^*)$ with the target eigenvalue and its left and right eigenvectors for the unperturbed matrix A. We set

$$E_0 = -\frac{G_0}{\|G_0\|},$$

which is the steepest descent direction at the unperturbed matrix for the functional $F_\varepsilon(E)$. At least for ε_0 not too large, this yields $F_{\varepsilon_0}(E_0) < F_0$, where F_0 is the value assumed by the functional for the unperturbed matrix.

By formula (2.6), with $E(0) = E_0$, we formally apply the first Newton step and set

$$\varepsilon_0 = \frac{x(0)^* y(0)}{\|G_0\|}(F_0 - r). \qquad (2.7)$$

eq:eps0F

While the above choice of starting values is reasonable when only a single trajectory is computed, it might in some problems be necessary to run several trajectories to reduce the risk of getting trapped in a local minimum.

IV.3 Complex, real and structured stability radii

sec:dist-instab

The two-level approach of the previous section applies directly to computing the (complex, real or structured) distance to instability of a Hurwitz-stable matrix A. We choose the target eigenvalue $\lambda(M)$ as an eigenvalue of largest real part (among these, the one with largest imaginary part), and we take $f(\lambda, \overline{\lambda}) = -\operatorname{Re}\lambda$ as the function to be minimized. The eigenvalue optimization problem (2.1) then becomes the maximization problem

$$E(\varepsilon) = \arg \max_{E\in\mathcal{S}, \|E\|_F=1} \operatorname{Re}\lambda(A + \varepsilon E), \qquad (3.1)$$

E-eps-stab-radius

and the optimal perturbation size $\varepsilon_\star$ is determined from Equation (1.17), which here reads

$$\operatorname{Re}\lambda(A + \varepsilon_\star E(\varepsilon_\star)) = 0. \qquad (3.2)$$

zero-stab-radius

In the inner iteration, the eigenvalue optimization problem (3.1) is of the class studied in Chapter II with $f(\lambda, \overline{\lambda}) = -\operatorname{Re}\lambda$ and is solved with the constrained gradient flow approach developed there, restricted to rank-1 perturbations in the complex unstructured case $\mathcal{S} = \mathbb{C}^{n,n}$ and to rank-2 perturbations in the real unstructured case $\mathcal{S} = \mathbb{R}^{n,n}$.

In the outer iteration we compute the optimal perturbation size $\varepsilon_\star$ by the Newton / bisection algorithm of Section IV.2.3 for $f(\lambda,\overline{\lambda}) = -\mathrm{Re}\,\lambda$, using the derivative formula of Theorem 2.2 with $G(\varepsilon) = -\Pi^{\mathcal{S}}(x(\varepsilon)y(\varepsilon)^*)$; in particular, $G = -xy^*$ in the complex untructured case, and $G = -\mathrm{Re}(xy^*)$ in the real unstructured case.

We consider the following numerical example: ...

IV.3.1 An illustrative example of distance to instability

IV.4 Matrix stabilization

sec:mat-stab

In this section we extend the two-level approach of Section IV.2 to the problem of moving all eigenvalues of a given matrix into a prescribed closed subset $\overline{\Omega}$ of the complex plane by a perturbation of minimal Frobenius norm. This spectral recovery problem is complementary to the robustness analysis in the previous section where the original matrix has all its eigenvalues inside Ω and it was required that one eigenvalue be driven to the boundary of Ω. While the approach presented here is conceptually applicable to the spectral recovery problem for quite general subsets Ω, we will focus our attention on the guiding problem of *Hurwitz stabilization* of an unstable matrix, which corresponds to the case where $\overline{\Omega} = \overline{\mathbb{C}}^-$ is the closed left complex half-plane:

Given a square matrix A that has some eigenvalues with positive real part, find a perturbation Δ of minimal Frobenius norm such that $A + \Delta$ has no eigenvalue with positive real part.

The perturbations Δ are restricted to lie in a structure space $\mathcal{S}$ that is $\mathbb{C}^{n,n}$ or $\mathbb{R}^{n,n}$ or a complex or real linear subspace of $\mathbb{C}^{n,n}$.

We describe two algorithmic approaches to this matrix stabilization problem, which are both of the two-level type considered in Section IV.2.

- The *exterior algorithm* moves eigenvalues that lie outside the closed target set $\overline{\Omega}$ where the eigenvalues of the perturbed matrix $A+\Delta$ should lie ($\overline{\Omega} = \overline{\mathbb{C}^-}$ for Hurwitz stability and $\overline{\Omega}$ is the closed unit disk for Schur stability). All eigenvalues outside $\overline{\Omega}$ are moved towards the boundary of Ω while increasing the perturbation size.
- The *interior algorithm* starts from a non-optimal perturbation Δ_0 such that $A + \Delta_0$ has all eigenvalues in Ω and moves some eigenvalue inside Ω to the boundary while reducing the perturbation size.

IV.4.1 Exterior two-level algorithm

subsec:ext-stab

Here we use the following eigenvalue optimization problem: For a given perturbation size $\varepsilon > 0$, find

$$\arg \min_{\Delta \in \mathcal{S},\, \|\Delta\|_F = \varepsilon} \sum_{i=1}^{n} f\left(\lambda_i\left(A+\Delta\right), \overline{\lambda}_i\left(A+\Delta\right)\right), \tag{4.1}$$

f-E-eps-stab

with

$$f\left(\lambda,\overline{\lambda}\right)=\frac{1}{2}\,\mathrm{dist}(\lambda,\overline{\mathbb{C}^-})^2=\frac{1}{2}\big((\mathrm{Re}\,\lambda)_+\big)^2=\frac{1}{8}\big((\lambda+\overline{\lambda})_+\big)^2, \tag{4.2}$$

eq:Hurw

where for $a\in\mathbb{R}$, $a_+ := \max\{a,0\}$.

The set $\{\lambda_i(A+\varDelta)\}_{i=1}^n$ of eigenvalues of the perturbed matrix $A+\varDelta$ is ordered by decreasing size of the real part. Note that only the eigenvalues with positive real part contribute to the sum in (4.1), so that the sum only extends from 1 to $m^+(A+\varDelta)$, where $m^+(M)$ is the number of eigenvalues of M with positive real part. As in Section IV.2, this leads us to the following two-level approach to the Hurwitz stabilization problem. Here we additionally introduce a small shift $\delta>0$ that aims for strict Hurwitz stability with all eigenvalues having real part not exceeding $-\delta$.

- **Inner iteration:** Given $\varepsilon>0$, we aim to compute a matrix $E(\varepsilon)\in\mathcal{S}$ of unit Frobenius norm such that

$$F_\varepsilon(E)=\frac{1}{2}\sum_{i=1}^n\big((\mathrm{Re}\,\lambda_i(A+\varepsilon E)+\delta)_+\big)^2 \tag{4.3}$$

Feps-stab

is minimized, i.e.

$$E(\varepsilon)=\arg\min_{E\in\mathcal{S},\|E\|_F=1}F_\varepsilon(E). \tag{4.4}$$

E-eps-stab

- **Outer iteration:** We compute the smallest positive value $\varepsilon_\star$ with

$$\phi(\varepsilon_\star)=0, \tag{4.5}$$

eq:zero-stab

where $\phi(\varepsilon)=F_\varepsilon\left(E(\varepsilon)\right)=\frac{1}{2}\sum_{i=1}^n\big((\mathrm{Re}\,\lambda_i(A+\varepsilon E(\varepsilon)+\delta)_+\big)^2$.

We remark that the existence of a zero of ϕ (i.e., stabilizability) is not guaranteed for arbitrary structure spaces $\mathcal{S}$, but it is when $\mathcal{S}$ equals $\mathbb{C}^{n,n}$ or $\mathbb{R}^{n,n}$ or more generally when $\mathcal{S}$ contains real multiples of the identity matrix I (since then some negative shift of the given matrix moves all its eigenvalues into the left half-plane).

As before, the inner iteration uses a norm- and structure-constrained gradient-flow differential equation, possibly further restricted to low-rank dynamics; the outer iteration uses a hybrid Newton / bisection method. We give details in the following subsections.

grad-flow-stab

Constrained gradient flow for minimizing $F_\varepsilon(E)$. Here we let $\varepsilon>0$ be fixed. Let $E(t)\in\mathcal{S}$, for t in an interval I, be a continuously differentiable path of matrices in the structure space $\mathcal{S}\subset\mathbb{C}^{n,n}$, and let the eigenvalues $\lambda_i(t)=\lambda_i(A+\varepsilon E(t))$ be simple for $i=1,\dots,n$ and all $t\in I$. The corresponding left and right eigenvectors x_i and y_i are assumed to be of unit norm and with $x_i^*y_i>0$. Then, applying Lemma VIII.1.1 we obtain

$$\frac{d}{dt}F_\varepsilon\big(E(t)\big)=\varepsilon\sum_{i=1}^n\big(\mathrm{Re}\,\lambda_i(A+\varepsilon E(t))+\delta\big)_+\,\frac{\mathrm{Re}(x_i(t)^*\dot{E}(t)y_i(t))}{x_i(t)^*y_i(t)}. \tag{4.6}$$

eq:derFeps

With the notation

$$\gamma_i(t):=\frac{\big(\mathrm{Re}\,\lambda_i(A+\varepsilon E(t))+\delta\big)_+}{x_i(t)^*y_i(t)}\geq 0.$$

Here we note that $\gamma_i(t) = 0$ for $i > m_\delta(A + \varepsilon E(t))$, which is the number of eigenvalues with real part greater than $-\delta$. We write (4.6) as

$$\frac{d}{dt} F_\varepsilon\big(E(t)\big) = \varepsilon \sum_{i=1}^{n} \gamma_i(t)\, \mathrm{Re}\big(x_i(t)^* \dot{E}(t) y_i(t)\big) = \varepsilon\, \mathrm{Re}\, \big\langle G_\varepsilon(E(t)), \dot{E}(t) \big\rangle$$

with the rescaled free gradient

$$G_\varepsilon(E) = \sum_{i=1}^{n} \gamma_i x_i y_i^* . \tag{4.7}$$

grad-stab

This is of rank at most $m_\delta(A + \varepsilon E)$. In the structured case where $E(t) \in \mathcal{S}$ for all t, and hence also its derivative is in $\mathcal{S}$ so that $\dot{E}(t) = \mathit{\Pi}^{\mathcal{S}} \dot{E}(t)$, we further obtain that with the projected gradient

$$G_\varepsilon^{\mathcal{S}}(E) = \mathit{\Pi}^{\mathcal{S}} G_\varepsilon(E) \in \mathcal{S},$$

we have

$$\frac{d}{dt} F_\varepsilon\big(E(t)\big) = \varepsilon\, \mathrm{Re}\, \big\langle G_\varepsilon^{\mathcal{S}}(E(t)), \dot{E}(t) \big\rangle. \tag{4.8}$$

eq:derFeps2

As in Section II.3, see (II.3.6), we consider the constrained gradient flow

$$\dot{E} = -G_\varepsilon^{\mathcal{S}}(E) + \mathrm{Re}\langle G_\varepsilon^{\mathcal{S}}(E), E \rangle E, \tag{4.9}$$

ode-E-S-stab

which again has the properties that

- the unit Frobenius norm is conserved along solutions $E(t)$;
- $F_\varepsilon(E(t))$ decays monotonically with growing t;
- stationary points E are real multiples of $G_\varepsilon^{\mathcal{S}}(E)$ provided that $G_\varepsilon^{\mathcal{S}}(E) \neq 0$.

Therefore, a stationary point E is a projection onto the structure space $\mathcal{S}$ of a matrix of rank at most $m_\delta = m_\delta(A + \varepsilon E)$. In particular, in the complex unstructured case the rank is at most m_δ, and in the real case at most $2m_\delta$.

c:low-rank-stab

Rank-constrained gradient flow in the complex and real unstructured cases. With an expected upper bound m of $m_{\delta,\varepsilon} = m_\delta(A + \varepsilon E(\varepsilon))$ at the minimizer $E(\varepsilon)$, which is of rank at most $m_{\delta,\varepsilon}$ in the complex unstructured case and at most $2m_{\delta,\varepsilon}$ in the real case, we can use a rank-constrained gradient flow in the same way as in Section II.2.3, where the chosen rank is now $r = m$ in the complex case and $r = 2m$ in the real case. We then consider the rank-r constrained gradient flow, with $\mathcal{S} = \mathbb{C}^{n,n}$ or $\mathbb{R}^{n,n}$,

$$\dot{E} = -P_E(G_\varepsilon^{\mathcal{S}}(E)) + \mathrm{Re}\langle E, P_E(G_\varepsilon^{\mathcal{S}}(E)) \rangle E, \tag{4.10}$$

ode-ErF-2-v2-stab

where $P_E(Z)$ is the orthogonal projection of $Z \in \mathbb{C}^{n,n}$ onto the tangent space at E of the manifold of (complex or real) $n \times n$-matrices of rank r. This differential equation is of the same type as in Section II.2.3 and is treated numerically in the same way as described there. (In the complex case, transposes of matrices are replaced by conjugate transposes.)

Iteration for ε. To solve the one-dimensional root-finding problem (4.4), we use a Newton / bisection method as in Section IV.2.3. We let $E(\varepsilon)$ of unit Frobenius norm be a (local) minimizer of the optimization problem (4.3) and we denote by $\lambda_i(\varepsilon)$, the eigenvalues and by x_i and y_i corresponding left and right eigenvectors of $A + \varepsilon E(\varepsilon)$, of unit norm and with positive inner product.

We denote by $\varepsilon_\star$ the smallest value of ε such that $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon)) = 0$. For a Newton-like algorithm we need an extra assumption that plays the same role as Assumption 2.1 in Section IV.2.

Assumption 4.1. For ε close to $\varepsilon_\star$ and $\varepsilon < \varepsilon_\star$, we assume the following:

- The eigenvalues of $A + \varepsilon E(\varepsilon)$ with positive real part are *simple* eigenvalues.
- The map $\varepsilon \mapsto E(\varepsilon)$ is continuously differentiable.
- The structured gradient $G(\varepsilon) := G_\varepsilon^{\mathcal{S}}(E(\varepsilon))$ is nonzero.

assumpt-stab

The following result extends Theorem 2.2 from one to several eigenvalues and is proved by the same arguments.

lem:der **Lemma 4.2 (Derivative for the Newton iteration).** *Under Assumption* 4.1*, the function* $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon))$ *is differentiable and its derivative equals*

$$\phi'(\varepsilon) = -\|G(\varepsilon)\|_F. \tag{4.11}$$

eq:derFdeps

With the derivative of ϕ at hand, we compute $\varepsilon_\star$ by a hybrid Newton / bisection algorithm as described in Section IV.2.3. In addition, since $F_\varepsilon(E(\varepsilon)) = 0$ for $\varepsilon \geq \varepsilon_\star$, we take a bisection step when all eigenvalues of $A + \varepsilon E(\varepsilon)$ have real part smaller than $-\delta$.

As an example consider the Grcar matrix of dimension $n = 6$,

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 1 & 1 & 0 \\ 0 & -1 & 1 & 1 & 1 & 1 \\ 0 & 0 & -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \tag{4.12}$$

The matrix has all (6) eigenvalues in the right complex half-plane (red points in Figure 4.1. We set $\delta = 0.1$, meaning we intend to push the whole spectrum on the left of the axis $\mathrm{Re}(z) = -\delta$.

In Figure 4.1 we show the paths of eigenvalues corresponding to the matrix $A+\varepsilon E(\varepsilon)$ for $\varepsilon \in [0, \varepsilon_\star]$ where $E(\varepsilon)$ indicates the extremizer of the functional computed at ε.

The behavior of $\varphi(\varepsilon)$ along the sequence ε_k selected by means of the Newton iteration is shown in Figure 4.2.

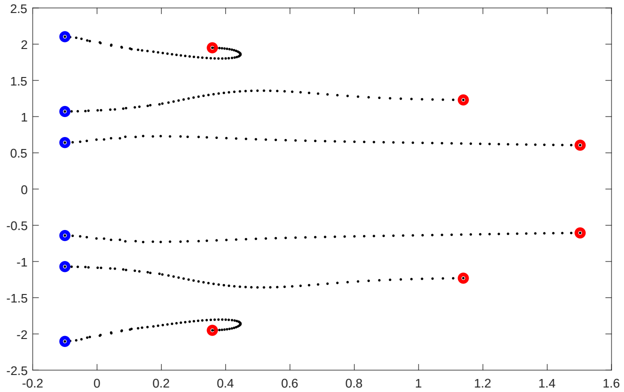

Fig. 4.1. Paths of the eigenvalues of the matrix $G + \varepsilon E(\varepsilon)$ for $\varepsilon \in [0, \varepsilon_\star]$.

fig:Stab1

IV.4.2 Interior two-level algorithm

subsec:int-stab

In the previous subsection, we worked with perturbed matrices that had some eigenvalues of positive real part, and the algorithm moved them to the left until it terminated with a matrix all of whose eigenvalues had real part at most $-\delta$. In an alternative approach, the given matrix A is first perturbed to a non-optimal matrix $A + \varepsilon_0 E_0$ with $\varepsilon_0 > 0$ and $E_0 \in \mathcal{S}$ of unit Frobenius norm such that all its eigenvalues have real part smaller than $-\delta$. For example, this can be achieved by a simple shift. We then reduce the perturbation size.

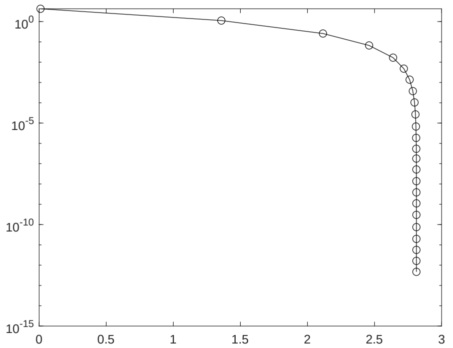

Fig. 4.2. The values $\varphi(\varepsilon_k)$ along the sequence $\{\varepsilon_k\}$ generated in the outer iteration.

fig:Stab2

Using a two-level iteration starting from $A+\varepsilon_0 E_0$, we aim to reduce the perturbation size to find $\varepsilon > 0$ for which $A+\varepsilon E$ has some eigenvalue of real part at least $-\delta$ for *every* matrix E of Frobenius norm 1. This differs from the problem of computing the distance to instability, where the aim was to find the smallest perturbation size ε for which $A+\varepsilon E$ (with A having only eigenvalues of negative real part) has eigenvalues of nonnegative real part for *some* matrix E of Frobenius norm 1.

As in the algorithm for computing the distance to instability of Section IV.3, the target eigenvalue $\lambda(M)$ of a matrix M is chosen as an eigenvalue of maximal real part. The function to be minimized is now $f(\lambda,\overline{\lambda}) = +\operatorname{Re}\lambda$, whereas it was $f(\lambda,\overline{\lambda}) = -\operatorname{Re}\lambda$

for computing the distance to instability of a Hurwitz matrix. Schematically, the interior two-level algorithm proceeds as follows.

- **Inner iteration:** Given $\varepsilon > 0$, we aim to compute a matrix $E(\varepsilon) \in \mathcal{S}$ of unit Frobenius norm such that the rightmost eigenvalue of $A + \varepsilon E$ is minimized, i.e.

$$E(\varepsilon) = \arg \min_{E \in \mathcal{S}, \|E\|_F = 1} \operatorname{Re} \lambda(A + \varepsilon E). \tag{4.13}$$

E-eps-stab-int

  (Note that for computing the distance to instability, we maximized the same functional; see (3.1).)
- **Outer iteration:** We compute $\varepsilon_\star > 0$ as the solution of the one-dimensional equation

$$\phi(\varepsilon_\star) = -\delta, \tag{4.14}$$

eq:zero-stab

  where $\phi(\varepsilon) = \operatorname{Re} \lambda(A + \varepsilon E(\varepsilon))$.

This yields the nearest perturbed matrix $A + \varepsilon_\star E(\varepsilon_\star)$ with all eigenvalues of real part at most $-\delta$ and the perturbation matrix in the structure space $\mathcal{S}$.

In the inner iteration, the eigenvalue optimization problem (4.13) is of the class studied in Chapter II and is solved with the constrained gradient flow approach developed there, restricted to rank-1 perturbations in the complex unstructured case $\mathcal{S} = \mathbb{C}^{n,n}$ and to rank-2 perturbations in the real unstructured case $\mathcal{S} = \mathbb{R}^{n,n}$.

In the outer iteration we compute the optimal perturbation size $\varepsilon_\star$ by the Newton / bisection algorithm of Section IV.2.3 for $f(\lambda, \overline{\lambda}) = \operatorname{Re} \lambda$, using the derivative formula of Theorem 2.2 with $G(\varepsilon) = \Pi^{\mathcal{S}}(x(\varepsilon)y(\varepsilon)^*)$; in particular, $G = xy^*$ in the complex untructured case, and $G = \operatorname{Re}(xy^*)$ in the real unstructured case.

IV.4.3 An illustrative example

sec:ill1

Consider the matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & -1 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 1 & 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 1 & -1 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & -1 & 1 & 1 & 0 & -1 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \tag{4.15}$$

ex:1

The matrix A has 6 eigenvalues with positive real part.

Exterior method. The stabilized matrix presents 7 eigenvalues on the parallel axis imaginary axis and abscissa $-\delta$; its distance from A is $\varepsilon_\star \approx 2.56$. For comparison, the stabilized matrix computed by the algorithm of Orbandexivry, Nesterov & Van Dooren (**?**) has a much larger distance 9.02 and all eigenvalues are located on the imaginary axis.

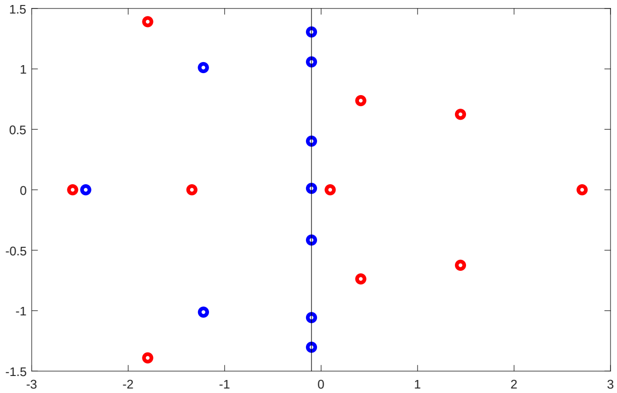

Fig. 4.3. Spectrum of the matrix (4.15) (red circles) and of the stabilized matrix $A + \varepsilon_\star E(\varepsilon_\star)$ (blue circles) in the exterior method.

fig1

Interior method. We start by computing a stable matrix, with spectrum on the left of the half-line Re$z = -\delta$, by using a method developed by Michael Overton, using a BFGS-type method on the penalized functional

$$\||X - A\||_F + \rho\,\alpha(X)$$

where $\alpha(X)$ denotes the spectral abscissa of X and ρ is a penalty parameter. We obtain in this way an initial matrix for our method, whose distance from A is approximately 3.14. After applying our interior type method we obtain a stabilized matrix which presents 3 eigenvalues on the parallel axis imaginary axis and abscissa $-\delta$; its distance from A is

$\varepsilon_\star \approx 2.94$, which is larger than the one computed by the exterior method, but still a nearby matrix. It is worth to remark that using the method proposed by Gillis and Sharma a closer matrix is computed, with approximate distance 1.92 and characterized by 8 eigenvalues aligned on the vertical axis Re $z = -\delta$.

Examples of large matrices

IV.5 Structured distance to singularity

Let $A \in \mathbb{C}^{n,n}$ be an invertible matrix, and let $\mathcal{S}$ be a complex- or real-linear subspace of $\mathbb{C}^{n,n}$ that defines the linear structure that is imposed on perturbations $\Delta \in \mathcal{S}$ to A. We consider the following structured matrix nearness problem.

Problem. *Given an invertible matrix A, find a structured perturbation $\Delta \in \mathcal{S}$ of minimal Frobenius norm such that $A + \Delta$ is singular.*

The norm of the minimizing $\Delta \in \mathcal{S}$ is called the $\mathcal{S}$-structured distance to singularity of the given matrix A. Unlike the complex or real unstructured distance to singularity, it cannot be obtained from a singular value decomposition of A. For its computation we use the two-level iteration of Section IV.2.

An obvious case of interest is when $\mathcal{S}$ is a space of complex or real matrices with a prescribed sparsity pattern. In the following we consider a different situation where a structured distance to singularity is of interest.

IV.5.1 Example: Nearest pair of polynomials with a common zero

We consider polynomials with real coefficients (alternatively, we could allow for complex coefficients). A pair of polynomials (p, q) is called *coprime* if p and q have no nontrivial common divisor, or equivalently, have no common zero. For a pair of polynomials (p, q) that is coprime, the distance to the nearest pair of polynomials with a common zero is of interest. In the following we measure the distance of pairs of polynomials by the Euclidean norm of the difference of the vectors of coefficients.

Problem. *Given a pair of polynomials that is coprime, find the nearest pair of polynomials with a nontrivial common divisor.*

Consider polynomials

$$\begin{aligned} p(z) &= a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 \\ q(z) &= b_m z^m + b_{m-1} z^{m-1} + \cdots + b_1 z + b_0 \end{aligned} \tag{5.1}$$

with real coefficients a_i and b_i. We may assume $m \le n$ and $a_n \neq 0$, and we can take $m = n$ if we allow for $b_n = 0$. So we assume $m = n$ in the following. With these polynomials we associate the Sylvester matrix of dimension $2n \times 2n$,

$$S(p,q) := \begin{pmatrix} a_n & & \dots & & a_0 & & & \\ & a_n & & \dots & & a_0 & & \\ & & \ddots & & & & \ddots & \\ & & & a_n & & \dots & & a_0 \\ b_n & & \dots & & b_0 & & & \\ & b_n & & \dots & & b_0 & & \\ & & \ddots & & & & \ddots & \\ & & & b_n & & \dots & & b_0 \end{pmatrix}. \tag{5.2}$$

eq:Sylpq

A result from commutative algebra (see, e.g. cite!) states:
The pair of polynomials (p, q) is coprime if and only if the associated Sylvester matrix $S(p, q)$ is invertible.

This allows us to reformulate the problem of nearest non-coprime polynomials as a matrix nearness problem, where the distance is measured by the Frobenius norm.

Problem. *Given an invertible Sylvester matrix, find the nearest singular Sylvester matrix.*

This amounts to the problem of computing the structured distance to singularity, where the structure is given by the subspace $\mathcal{S} \subset \mathbb{R}^{2n,2n}$ of Sylvester matrices of the form (5.2).

Let S be a Sylvester matrix. We define the *lower radius* of S as

$$\mu(S) = \min\{|\lambda| \,:\, \lambda \text{ is an eigenvalue of S}\},$$

and note that S is singular if and only if $\mu(S) = 0$. Furthermore, with the structured ε-pseudospectrum $\Lambda_\varepsilon^{\mathcal{S}}(S)$, let

$$\mu_\varepsilon^{\mathcal{S}}(S) = \min\{|\lambda| : \lambda \in \Lambda_\varepsilon^{\mathcal{S}}(S)\}, \tag{5.3}$$

eq:psa

which reduces to the lower radius $\mu(S)$ when $\varepsilon = 0$. We can then express the radius of coprimeness of the pair of polynomials (p, q) as

$$\rho_{\mathrm{co}}(p,q) = \frac{\varepsilon_\star}{\sqrt{n}} \quad \text{with} \quad \varepsilon_\star = \min\{\varepsilon > 0 \,:\, \mu_\varepsilon^{\mathcal{S}}(S) = 0\}.$$

(The division by $\sqrt{n}$ is done to account for the fact that each coefficient a_i and b_i appears n times in the Sylvester matrix, which yields a factor $\sqrt{n}$ in the Frobenius norm.)

Two-level iteration. We are thus in the situation of applying the two-level iteration of Section IV.2 with the functional $F_\varepsilon(E)$ (for $E \in \mathcal{S}$ of unit Frobenius norm) given as

$$F_\varepsilon(E) = \mu(S + \varepsilon E), \tag{5.4}$$

eq:FepSyl

which is of the form (1.3) with $f(\lambda, \overline{\lambda}) = \sqrt{\lambda\overline{\lambda}}$ and with the eigenvalue of smallest modulus as target eigenvalue.

To apply the gradient-based algorithm of Section II.3 in the inner iteration, we need the structured gradient, see II.3.4), which is the orthogonal projection onto the space $\mathcal{S}$ of Sylvester matrices of the (rescaled) gradient

$$G_\varepsilon(E) = \mathrm{Re}\Big(\frac{\lambda}{|\lambda|}\, xy^*\Big),$$

where x and y are the left and right eigenvectors, normalized to unit norm and with positive inner product, that are associated with the eigenvalue λ of smallest modulus of $S + \varepsilon E$, which is assumed to be simple. We note that the structured gradient is nonzero by Lemma II.3.5.

The orthogonal projection onto $\mathcal{S}$ is given in the following lemma.

lem:projS

Lemma 5.1 (Orthogonal projection onto the space of Sylvester matrices). *Let $S \in \mathcal{S} \subset \mathbb{R}^{2n\times 2n}$ and $Z \in \mathbb{C}^{2n,2n}$; the orthogonal projection $P_{\mathcal{S}}$ onto $\mathcal{S}$, with respect to the Frobenius inner product $\langle\cdot,\cdot\rangle$, is given by*

$$\Pi^{\mathcal{S}} Z \quad = \quad S(p,q), \tag{5.5}$$

where p and q are the polynomials with coefficients (for $k = 0,\ldots,n$)

$$a_{n-k} = \frac{1}{n}\sum_{l=1}^{n} \mathrm{Re}\left(Z_{l,l+k}\right), \qquad b_{n-k} = \frac{1}{n}\sum_{l=1}^{n} \mathrm{Re}\left(Z_{n+l,l+k}\right).$$

Proof. We have to find $\arg\min_{S\in\mathcal{S}} \|Z - S\|_F$. The result follows directly from the fact that for a complex vector $x \in \mathbb{C}^n$,

$$\mu_* = \arg\min_{\mu\in\mathbb{R}} \|x - \mu\mathbf{1}\|_F = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Re}(x_i),$$

where $\mathbf{1} = (1\,1\,\ldots\,1)^\top$. □

sec:exill

Numerical example. Consider the two polynomials of degree 3,

$$p(z) = z^3 + 2z^2 + 2z + 2, \quad q(z) = 2z^3 + z - 2, \tag{5.6}$$

ex:pq1

where p is constrained to be monic. Here $a = (1\,1\,2\,2)^\top$ and $b = (2\,0\,1\,-2)^\top$; the corresponding Sylvester matrix is given by

$$S(a,b) = \begin{pmatrix} 1 & 2 & 2 & 2 & 0 & 0 \\ 0 & 1 & 2 & 2 & 2 & 0 \\ 0 & 0 & 1 & 2 & 2 & 2 \\ 2 & 0 & 1 & -2 & 0 & 0 \\ 0 & 2 & 0 & 1 & -2 & 0 \\ 0 & 0 & 2 & 0 & 1 & -2 \end{pmatrix} \tag{5.7}$$

eq:illS

The structured pseudospectrum $\Lambda_\varepsilon^{\mathcal{S}}(S)$ for $\varepsilon = \frac{1}{2}$ is approximated by dense sampling on the set of admissible perturbations and is plotted in blue in Figure 5.1.

It turns out that for the value

$$\varepsilon = \varepsilon^\star = 0.618108064$$

the functional $F_\varepsilon(E)$ (see (5.4)) vanishes, while for $\varepsilon < \varepsilon^\star$ it holds $F_\varepsilon(E) > 0$. This gives $\rho_{\rm co}(p,q) = 0.356864857$.

The computed matrix $S + \varepsilon^\star E(\varepsilon^\star)$ has rank $2n-2$ due to a double semi-simple zero eigenvalue. The coefficients of the perturbed polynomials $\hat{p}$, $\hat{q}$ are shown (with 10-digit accuracy) in Table 5.1. The common complex conjugate zeros of $\hat{p}$, $\hat{q}$ are

$$z_{1,2} = -0.4008686595 \pm 1.03085391659\mathrm{i}.$$

Table 5.1. Coefficients of the perturbed polynomials $\hat{p}$, $\hat{q}$ in the example 5.6. t1

$\hat{a}_3 = 0.75744188$	$\hat{a}_2 = 2.10479150$	$\hat{a}_1 = 2.12724001$	$\hat{a}_0 = 1.83200184$
$\hat{b}_3 = 1.95430087$	$\hat{b}_2 = -0.06706025$	$\hat{b}_1 = 1.08084913$	$\hat{b}_0 = -1.99883585$

IV.6 Notes

An early survey of matrix nearness problems, with emphasis on the properties of symmetry, positive definiteness, orthogonality, normality, rank-deficiency and instability, was given by Higham (1989). This review is a source of continuing interest in view of its choice of topics and the references to the older literature.

Distance to instability (stability radius) under complex unstructured perturbations. Van Loan (1985) was apparently the first to address the question "How near is a stable matrix to an unstable matrix?". He considered both complex and real perturbations and came up with heuristic algorithms for approximating the smallest perturbations that shift an eigenvalue to the imaginary axis. His starting point was the characterization of the distance to instability under complex unstructured perturbations of the matrix A as

$$\beta(A) = \min_{\omega\in\mathbb{R}} \sigma_{\min}(A - \mathrm{i}\omega I) \tag{6.1}$$

stab-rad

and an intricate characterization of the distance to instability under real perturbations.

For the *complex* case, Byers (1988) showed that the Hamiltonian matrix

$$H(\sigma) = \begin{pmatrix} A & -\sigma I \\ \sigma I & A^* \end{pmatrix}$$

has a purely imaginary eigenvalue if and only if $\sigma \geq \beta(A)$; cf. Lemma III.2.1. Based on this result, he proposed a bisection method for computing the distance to the nearest complex matrix with an eigenvalue on the imaginary axis (the complex stability radius). Each step of the method requires the solution of an eigenvalue problem of the Hamiltonian matrix $H(\sigma)$ for varying $\sigma > 0$. Byers (1988) also gave an extension of the algorithm to compute the distance to the nearest complex matrix with an eigenvalue on the unit circle.

Conceptually related Hamiltonian eigenvalue methods by Boyd & Balakrishnan (1990) and Bruinsma & Steinbuch (1990) for the more general problem of computing the H^∞-norm of a transfer function also apply to computing the distance to stability. These methods converge locally quadratically.

He and Watson (1999) developed a method for computing the distance to instability that is better suited for large sparse matrices A. They use a method based on inverse iteration for singular values to compute a stationary point of the function $f(\omega) = \sigma_{\min}(A - \mathrm{i}\omega I)$. They then check whether the stationary point reached is a global minimum by solving an eigenvalue problem for $H(\sigma)$. An alternative method for large sparse matrices was devised by Kressner (2006) who worked with inverse iterations using sparse LU factorizations of imaginary shifts of Hamiltonian matrices $H(\sigma)$.

For $\sigma = \beta(A)$, the Hamiltonian matrix $H(\sigma)$ has an eigenvalue of even multiplicity on the imaginary axis. Generically, it is expected to be a defective double eigenvalue. Freitag & Spence (2011) used a Newton-based method to find the parameters σ and ω such that $H(\sigma) - \mathrm{i}\omega I$ has a zero eigenvalue corresponding to a two-dimensional Jordan block.

A different approach is to combine an algorithm for computing the ε-pseudospectral abscissa $\alpha_\varepsilon(A)$ (see Section III.2) with a root-finding algorithm for determining $\varepsilon_\star > 0$

such that $\alpha_{\varepsilon_\star}(A) = 0$. Then, $\varepsilon_\star$ is the distance to instability. This two-level approach can be used efficiently for large sparse matrices with the rank-1 iteration of Guglielmi & Overton (2011), with the subspace method of Kressner & Vandereycken (2014), and with the discretized rank-1 differential equation of Guglielmi & Lubich (2011). The latter, differential equation based approach is described here in Section IV.2. It extends in a direct way to computing the distance to instability under real or structured perturbations.

Distance to instability (stability radius) under real unstructured perturbations. Qiu, Bernhardsson, Rantzer, Davison, Young & Doyle (1995) characterized the real stability radius $r_{\mathbb{R}}(A)$ as

$$\frac{1}{r_{\mathbb{R}}(A)} = \sup_{\omega\in\mathbb{R}} \inf_{0<\gamma\le 1} \sigma_2 \begin{pmatrix} \mathrm{Re}\, M_\omega & -\gamma\, \mathrm{Im}\, M_\omega \\ \gamma^{-1}\mathrm{Im}\, M_\omega & \mathrm{Re}\, M_\omega \end{pmatrix} \qquad \text{with} \quad M_\omega = (A - \mathrm{i}\omega I)^{-1},$$

where $\sigma_2(\cdot)$ is the second largest singular value of a matrix. An algorithm for the computation of $r_{\mathbb{R}}(A)$ via this formula was proposed by Sreedhar, Van Dooren & Tits (1996). Based on a reformulation of this formula and using Byers' connection between singular values and eigenvalues of Hamiltonian matrices, Freitag & Spence (2014) also developed an algorithm to deal efficiently with this two-dimensional optimization.

In a different approach, Guglielmi & Manetta (2015) studied an algorithm that is well-suited also for large sparse matrices. It corresponds to the general two-level approach taken in Section IV.2. In the inner iteration, the algorithm computes the real ε-pseudospectral abscissa via rank-2 matrix differential equations of Guglielmi & Lubich (2013) (which are given there for both the matrix 2-norm and the Frobenius norm). In the outer iteration, it uses a combined Newton / bisection method to optimize the perturbation size ε to yield $\varepsilon_\star$ such that the real $\varepsilon_\star$-pseudospectral abscissa becomes zero. The Newton iteration used the simple derivative formula of Theorem 2.2 for the particular case of the real gradient $G = \mathrm{Re}(xy^*)$. A related method based on a real version of the iteration method of Guglielmi & Overton (2011) was proposed by Rostami (2015) and further developed and analysed by Guglielmi (2016).

Structured stability radii. In the control systems literature, Hinrichsen & Pritchard (1986a,1986b,1990) considered complex and real stability radii (i.e. distance to instability under complex and real perturbations) and also structured stability radii

$$r(A,B,C) = \min\{\|\Delta\| \, : \, A + B\Delta C \text{ has some eigenvalue of nonnegative real part}\},$$

where A is a Hurwitz-stable matrix and B and C are given matrices of compatible dimensions. The perturbation matrix Δ is assumed to be real or complex. Most of the algorithms mentioned above extend to this situation of range- and corange-restricted perturbations. For example, Hinrichsen, Kelb & Linnemann (1989) extended Byers' algorithm to compute the complex structured stability radius $r_{\mathbb{C}}(A,B,C)$.

We note that this notion of structured stability radius minimizes the norm of the parameter matrix Δ and not of the structured perturbation $B\Delta C$ of A. The latter would fit directly into the framework of Section IV.2, whereas controlling the norm of Δ requires some (minor) modifications to the algorithm.

We are not aware of algorithms for other structured stability radii (distance to instability under complex or real structured perturbations) in the literature, e.g. for perturbations with a given sparsity pattern and/or symmetry, or Toeplitz perturbations etc. The two-level algorithm of Section IV.2 addresses such problems with general linear structures.

Matrix stabilization. Finding the smallest (complex, real or structured) stabilizing perturbation to a given matrix is a harder problem than the complementary problem of finding the smallest destabilizing perturbation as discussed above. Several conceptually different algorithms for matrix stabilization with complex or real unstructured perturbations have been proposed in the literature.

A black-box approach is to consider the problem of finding the nearest stable matrix as a nonsmooth (but almost everywhere smooth), nonconvex, constrained optimization problem and apply general software for this class of problems, such as given by Curtis, Mitchell & Overton (2017).

Orbandexivry, Nesterov & Van Dooren (2013) presented a matrix stabilization algorithm that uses successive convex approximations. They started from Lyapunov's characterization of stability to reformulate the matrix stabilization problem as finding complex $n \times n$ matrices X and P that give

$$\inf_{X,P} \tfrac{1}{2}\|X - A\|_F^2 \quad \text{such that } P = P^* \text{ and } XP + PX^* \text{ are both positive definite.}$$

This nonconvex optimization problem is related to the convex problem of finding, for given X and P,

$$\inf_{H} \tfrac{1}{2}\|X + H - A\|_F^2 \quad \text{such that } H \text{ is in a suitable ellipsoid defined by } P \text{ and } X.$$

This update for X is complemented with a procedure that associates an admissible P to X. With an $O(n^5)$ complexity per iteration, the algorithm is limited to small matrices.

Gillis & Sharma (2017) showed that a real square matrix A is stable if and only if it can be written as the matrix of a dissipative Hamiltonian system, i.e. $A = (J - R)Q$, where J is skew-symmetric, R is positive semidefinite and Q is positive definite. This reformulation results in an equivalent nonconvex optimization problem with a convex feasible region onto which points can be projected easily. The authors proposed a projected gradient method (among other strategies) to solve the problem in the variables (J, R, Q), with $O(n^3)$ complexity per iteration. Gillis, Karow & Sharma (2019) made an analogous approach to Schur stablilization, based on their characterization of a Schur-stable matrix as being of the form $A = S^{-1}UBS$, where S is positive definite, U is orthogonal, and B is a positive semidefinite contraction. Choudhary, Gillis & Sharma (2020) extended the approach to finding the nearest matrix with eigenvalues in more general closed sets $\overline{\Omega}$ that are a finite intersection of disks, conical sectors and vertical strips.

Noferini & Poloni (2021) reformulated matrix stabilization as an optimization problem on the Riemannian manifold of orthogonal or unitary matrices. The problem is then solved using standard methods from Riemannian optimization. The problem of finding the nearest complex Hurwitz-stable matrix is shown to be equivalent to solving

$$\min_{U \in U(n)} \|L(U^* A U)\|_F^2,$$

where $L(Z)$ is the lower triangular matrix whose part below the diagonal coincides with that of Z, and the diagonal elements are changed to $L(Z)_{ii} = (\operatorname{Re} z_{ii})_+$. A related reformulation with orthogonal matrices is given for the real case. The approach is actually formulated for the problem of finding the nearest matrix with eigenvalues in an arbitrary prescribed closed set, thus including Hurwitz- and Schur-stability as special cases.

Guglielmi & Lubich (2017) studied an exterior two-level approach to matrix stabilization, with a gradient flow in the inner iteration and a combined Newton / bisection method in the outer iteration, as in Section IV.4.1 but with a different functional that aims at aligning a fixed number of eigenvalues on the imaginary axis. The interior two-level algorithm for matrix stabilization described in Section IV.4.2 is remarkably similar to the two-level algorithm for computing the (complex, real or structured) distance to instability. This interior algorithm has not appeared in the literature before, but it is related to algorithms for Hamiltonian matrix nearness problems and for the passivation of control systems proposed by Guglielmi, Kressner & Lubich (2015) and Fazzi, Guglielmi & Lubich (2021), respectively. In contrast to other methods in the literature, the exterior and interior two-level approaches of Section IV.4 can exploit sparsity of the given matrix A (in combination with low-rank perturbations) and they readily extend to matrix stabilization by perturbations with a prescribed linear structure, e.g. for perturbations with a given sparsity pattern.

The problem of finding the nearest nonnegative stable matrix was studied by Guglielmi & Protasov (2018) for the Frobenius norm, whereas Nesterov & Protasov (2020) considered the maximum norm.

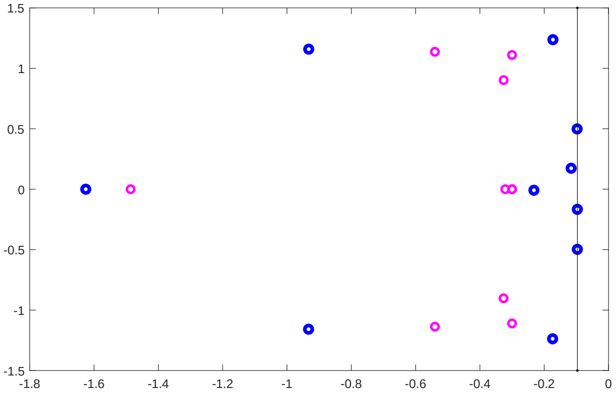

Fig. 4.4. Spectrum of the initial matrix stabilizing (4.15), computed by Overton's algorithm (magenta circles) and of the the matrix $A + \varepsilon_\star E(\varepsilon_\star)$ (blue circles) computed by the interior method.

fig1

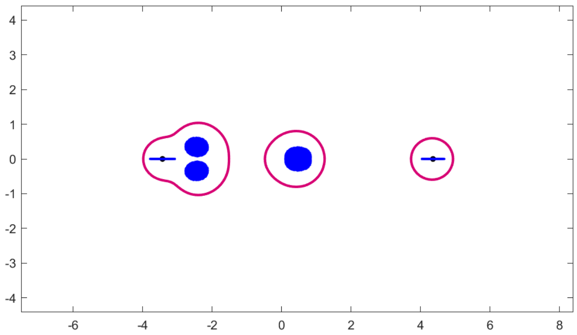

Fig. 5.1. The approximated structured Sylvester ε-pseudospectrum for $\varepsilon = \frac{1}{2}$ for Example (5.7) is filled with blue. The red curve represents the boundary of the set of eigenvalues obtained by considering arbitrary complex perturbations (that is omitting the constraint of real Sylvester structure) of norm bounded by $\frac{1}{2}$.

fig:illS

Chapter V.
Matrix nearness problems of different kinds

chap:mnp-mix

In this chapter we discuss extensions of the two-level approach of the previous chapter to matrix nearness problems that either

– require special attention to the structure, as in Hamiltonian matrix nearness problems,

or are of different types that are not covered by the framework of Chapter IV, be it because

– the functional in the associated eigenvalue optimization problem depends on eigenvectors, as in the Wilkinson problem of computing the distance to singularity of the eigenvalue condition number, which amounts to finding the nearest matrix with defective eigenvalues; or
– the nearness problem deals with matrix pencils, as in the problem of finding the nearest matrix pencil that is singular or additionally has a common null-vector; or
– it deals with eigenvalue problems that are nonlinear in the eigenvalues, as in computing the stability radius of linear delay differential equations; or
– it treats simultaneously both structured and unstructured perturbations, as in the problem of bounding transient behaviour of linear differential equations with structured perturbations to the matrix, which leads to the notion of structured ε-stability radius.

These items are exemplary, not exhaustive. They form the sections of this chapter. The sections can be read independently of each other, but we give fewer details in the later sections.

This chapter illustrates the versatility of the two-level approach that uses rank-constrained gradient flows in the inner iteration and combined Newton-type / bisection methods in the outer iteration. It enriches the toolbox for applications in various fields, such as those considered in the final chapters.

V.1 Hamiltonian matrix nearness problems

sec:Hamilton

In this section we apply the two-level algorithmic approach of the previous chapter to matrix nearness problems for real Hamiltonian matrices, where ‘nearest’ again refers to the smallest distance in the Frobenius norm:

- **Problem A.** *Given a Hamiltonian matrix with no eigenvalues on the imaginary axis, find a nearest Hamiltonian matrix having some purely imaginary eigenvalue.*

- **Problem B.** *Given a Hamiltonian matrix with some eigenvalues on the imaginary axis, find a nearest Hamiltonian matrix such that arbitrarily close to that matrix there exist Hamiltonian matrices with no eigenvalues on the imaginary axis.*

Such and related Hamiltonian matrix nearness problems arise in the passivation of linear time-invariant control systems and in the stabilization of gyroscopic systems; see the references in the notes at the end of this chapter.

We deal here with structured matrix nearness problems with the structure space $\mathcal{S}$ given as the space $\mathrm{Ham}(\mathbb{R}^{n,n})$ of *real Hamiltonian* matrices (with even dimension $n = 2d$), consisting of those matrices $A \in \mathbb{R}^{n,n}$ for which

$$JA \ \text{ is real symmetric, where } \ J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix}.$$

We note that the eigenvalues of a real Hamiltonian matrix lie symmetric to both the real axis and the imaginary axis: with any eigenvalue λ, also $\overline{\lambda}, -\lambda, -\overline{\lambda}$ are eigenvalues. In fact, if x is a left eigenvector of A to the eigenvalue λ, then Jx is a right eigenvector of A to the eigenvalue $-\overline{\lambda}$, since $AJx = J^{-1}(JA)Jx = -JA^\top J^\top Jx = -J(x^*A)^* = -\overline{\lambda}Jx$.

V.1.1 Problem A: Moving eigenvalues to the imaginary axis

subsec:ham-A

For a real Hamiltonian matrix M, we let in the following the target eigenvalue $\lambda(M)$ be the eigenvalue of minimal real part in the first quadrant $\{\lambda \in \mathbb{C} \,:\, \mathrm{Re}\,\lambda \ge 0,\, \mathrm{Im}\,\lambda \ge 0\}$. (If this eigenvalue is not unique, we choose the one with minimal imaginary part.) We follow the two-level approach of Section IV.2:

- **Inner iteration:** Given $\varepsilon > 0$, we aim to compute a matrix $E(\varepsilon) \in \mathcal{S} = \mathrm{Ham}(\mathbb{R}^{n,n})$ of unit Frobenius norm, such that $\mathrm{Re}\,\lambda(A+\varepsilon E)$ is minimized:

$$E(\varepsilon) = \arg \min_{E \in \mathrm{Ham}(\mathbb{R}^{n,n}),\, \|E\|_F = 1} \mathrm{Re}\,\lambda(A+\varepsilon E). \tag{1.1}$$

E-eps-ham

- **Outer iteration:** We compute the smallest positive value $\varepsilon_\star$ with

$$\phi(\varepsilon_\star) = 0, \tag{1.2}$$

eq:zero-ham

where $\phi(\varepsilon) = \mathrm{Re}\,\lambda(A+\varepsilon E(\varepsilon))$.

We note that the inner iteration aims to find a leftmost point, in the first quadrant, of the structured ε-pseudospectrum $\Lambda_\varepsilon^{\mathcal{S}}(A) = \{\lambda \in \mathbb{C} \,:\, \lambda \text{ is an eigenvalue of } A+\varepsilon E \text{ for some } E \in \mathcal{S} \text{ with } \|E\|_F = 1\}$.

We recall from Section II.3.2 that the orthogonal projection $\Pi_{\mathcal{S}}$ from $\mathbb{C}^{n,n}$ onto $\mathcal{S} = \mathrm{Ham}(\mathbb{R}^{n,n})$ is given by

$$\Pi^{\mathcal{S}} Z = J^{-1}\mathrm{Sym}(\mathrm{Re}\,JZ), \qquad Z \in \mathbb{C}^{n,n}. \tag{1.3}$$

Pi-Ham-recall

We already know from Section II.3.3 that if $\lambda(A+\varepsilon E(\varepsilon))$ is a simple eigenvalue, then the optimizer $E(\varepsilon)$ is a stationary point of the structure-constrained gradient flow given in (II.3.6), viz.,

$$\begin{aligned}&\dot{E} = -G_\varepsilon^{\mathcal{S}}(E) + \mathrm{Re}\langle G_\varepsilon^{\mathcal{S}}(E), E\rangle E\\ &\text{with the projected gradient } \; G_\varepsilon^{\mathcal{S}}(E) = \Pi^{\mathcal{S}}(xy^*),\end{aligned} \tag{1.4}$$

ode-E-S-recall

where x and y are again left and right eigenvectors of $A+\varepsilon E$ with $\|x\| = \|y\| = 1$ and $x^*y > 0$. With the given definitions and the orthogonality of J, this becomes

$$J\dot{E} = -\mathrm{Sym}(\mathrm{Re}\, Jxy^*) + \langle \mathrm{Sym}(\mathrm{Re}\, Jxy^*), JE\rangle JE. \tag{1.5}$$

ode-E-ham

In a stationary point, E is a real multiple of $G_\varepsilon^{\mathcal{S}}(E) = J^{-1}\mathrm{Sym}(\mathrm{Re}\, Jxy^*)$, which is of rank at most 4. The precise rank is as follows.

thm:rank-ham

Theorem 1.1 (Rank of optimizers). *For a real Hamiltonian matrix A and $\varepsilon > 0$, let $E \in \mathbb{R}^{n,n}$ with $\|E\|_F = 1$ be a stationary point of the differential equation* (1.5) *such that the eigenvalue $\lambda = \lambda(A+\varepsilon E)$ is simple and $\lambda \notin \mathrm{i}\mathbb{R}$. Then,*

- *E has rank 4 if $\lambda \notin \mathbb{R}$.*
- *E has rank 2 if $\lambda \in \mathbb{R}$.*

Proof. If x is a left eigenvector to the eigenvalue λ, then Jx is a right eigenvector to $-\overline{\lambda}$. For a real Hamiltonian matrix, we have in addition that $J\overline{x}$ and $\overline{y}$ are eigenvectors to $-\lambda$ and $\overline{\lambda}$, respectively.

Under the assumption $\lambda \notin \mathbb{R}\cup\mathrm{i}\mathbb{R}$ we have four different eigenvalues lying symmetric to the real as well as the imaginary axis. The corresponding right eigenvectors $Jx, J\overline{x}, y, \overline{y}$ are linearly independent, and so are their real and imaginary parts Jx_R, y_R, Jx_I, y_I. Therefore, the matrix $JG_\varepsilon^{\mathcal{S}}(E))$, which is proportional to JE, equals

$$\mathrm{Sym}(\mathrm{Re}\, Jxy^*) = \tfrac{1}{4}(Jx_R, y_R, Jx_I, y_I)\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}(Jx_R, y_R, Jx_I, y_I)^T,$$

which has rank 4. Hence, $G_\varepsilon^{\mathcal{S}}(E) = J^{-1}\mathrm{Sym}(J\,\mathrm{Re}\, xy^*)$ has rank 4, and so has its nonzero real multiple E.

If λ is real and nonzero, then Jx and y are real, and they are linearly independent as eigenvectors to $-\lambda$ and λ. It follows that $G_\varepsilon^{\mathcal{S}}(E)$ and hence E are of rank 2. □

V.1.2 Inner iteration: rank-4 dynamics

c:rank-four-dyn

In view of Theorem 1.1, we restrict the gradient flow (1.4) to Hamiltonian rank-4 matrices, i.e., JE will be constrained to lie in the manifold of symmetric rank-4 matrices and hence can be represented as

$$JE = USU^\top,$$

where $U \in \mathbb{R}^{n\times 4}$ has orthonormal columns and $S \in \mathbb{R}^{4\times 4}$ is symmetric and invertible. Like in Section II.2.3, the orthogonal projection onto the tangent space at JE is given as

$$P_{JE}(Z) = Z - (I - UU^\top)Z(I - UU^\top), \qquad Z \in \mathbb{R}^{n\times n}.$$

We consider the rank-4 projected gradient flow

$$J\dot{E} = P_{JE}\Big(-\mathrm{Sym}(\mathrm{Re}\, Jxy^*) + \langle \mathrm{Sym}(\mathrm{Re}\, Jxy^*), JE\rangle JE\Big), \tag{1.6}$$

ode-E-4

where x and y are again left and right eigenvectors of $A + \varepsilon E$ with $\|x\| = \|y\| = 1$ and $x^*y > 0$ to the target eigenvalue $\lambda(A + \varepsilon E)$. Similarly to Sections II.1.7 and II.2.4, we have the following properties.

m:monotone-ham

Theorem 1.2 (Monotonicity). *Let $E(t)$ of unit Frobenius norm be a solution to the differential equation* (1.6)*. If the eigenvalue $\lambda(t) = \lambda(A + \varepsilon E(t))$ is simple, then*

$$\frac{d}{dt}\,\mathrm{Re}\,\lambda(t) \le 0. \tag{1.7}$$

eq:mon-ham

Proof. We abbreviate $G = G_\varepsilon^{\mathcal{S}}(E) = J^{-1}\mathrm{Sym}(\mathrm{Re}\, Jxy^*)$ and obtain from (II.3.4) with $f(\lambda, \overline{\lambda}) = \frac{1}{2}(\lambda + \overline{\lambda}) = \mathrm{Re}\,\lambda$ and $\kappa = 1/(x^*y) > 0$, using the orthogonality of J,

$$\begin{aligned} \frac{1}{\varepsilon\kappa}\,\frac{d}{dt}\,\mathrm{Re}\,\lambda(t) &= \langle G, \dot{E}\rangle = \langle JG, J\dot{E}\rangle \\ &= \langle JG, P_{JE}\big(-JG - \langle JG, JE\rangle E\big)\rangle \\ &= \Big(-\|P_{JE}(JG)\|_F^2 + \big(\mathrm{Re}\,\langle P_{JE}(JG), JE\rangle\big)^2\Big) \le 0, \end{aligned} \tag{1.8}$$

c-s-1-ham

where we used $P_{JE}(JE) = JE$ in the last equality, and $\|JE\|_F = 1$ and the Cauchy–Schwarz inequality in the final inequality. □

thm:stat-ham

Theorem 1.3 (Stationary points). *Let E be a real Hamiltonian rank-4 matrix of unit Frobenius norm and suppose that $P_{JE}(JG_\varepsilon^{\mathcal{S}}(E)) \neq 0$. If E is a stationary point of the projected differential equation* (1.6)*, then E is already a stationary point of the differential equation* (1.4)*.*

Proof. The proof is similar to the proof of Theorem II.2.4. We show that E is a real multiple of $G_\varepsilon^{\mathcal{S}}(E)$. By (II.4.14), E is then a stationary point of the differential equation (II.3.6), which is the same as (1.4).

For a stationary point E of (1.6), we must have equality in the estimate of the previous proof, which shows that $P_{JE}(JG)$ (with $G = G_\varepsilon^{\mathcal{S}}(E)$) is a nonzero real multiple of E. Hence, in view of $P_{JE}(JE) = JE$, we can write G as

$$G = \mu E + W, \qquad \text{where } \mu \neq 0 \text{ and } P_{JE}(JW) = 0.$$

With $JE = USU^\top$ as above, we then have

$$JW = JW - P_{JE}(JW) = (I - UU^\top)JW(I - UU^\top).$$

Since G is of rank at most 4 and real Hamiltonian, it can be written in the form $JG = XRX^\top$, where $X \in \mathbb{R}^{n,4}$ has orthonormal columns and $R \in \mathbb{R}^{4,4}$. So we have

$$XRX^\top = \mu USU^\top + (I - UU^\top)JW(I - UU^\top).$$

Multiplying from the right with U yields $X(RX^\top U) = \mu US$, which shows that X has the same range as U. Hence, JG has the same range as JE, which implies that $P_{JE}(JG) = JG$. Since we already know that $P_{JE}(JG)$ is a nonzero real multiple of $P_{JE}(JE) = JE$, it follows that G is the same real multiple of E. Hence E is a stationary point of (1.4). □

We further remark that for *real* simple eigenvalues λ we have an analogous rank-2 dynamics.

A robust integrator. The following time-stepping method is an adaptation to (1.6) of the low-rank integrator of Ceruti & Lubich (2021), similarly to the integrator in Section II.2.5. It first updates the basis matrix U with orthonormal columns and then computes an update of the symmetric 4×4 matrix S by a Galerkin approximation to the differential equation (1.6) in the updated basis. This integrator is robust to the presence of small singular values, which appear in the case of a target eigenvalue near the real axis, where the rank degenerates from 4 to 2.

One time step of integration from time t_k to $t_{k+1} = t_k + h$ starting from a factored rank-4 matrix $JE_k = U_k S_k U_k^\top$ of unit Frobenius norm computes an updated rank-r factorization $JE_{k+1} = U_{k+1} S_{k+1} U_{k+1}^\top$ of unit Frobenius norm as follows.

1. Update the basis matrix $U_k \to U_{k+1}$:
   Integrate from $t = t_k$ to $t_{k+1} = t_k + h$ the $n \times r$ matrix differential equation

   $$\dot K(t) = -JG_\varepsilon^{\mathcal{S}}(K(t)U_k^\top)U_k, \qquad K(t_k) = U_k S_k.$$

   Perform a QR factorization $K(t_{k+1}) = U_{k+1}R_{k+1}$ and compute the $r \times r$ matrix $M = U_{k+1}^\top U_k$.

2. Update the symmetric matrix $S_k \to S_{k+1}$:
   Integrate from $t = t_k$ to t_{k+1} the $r \times r$ matrix differential equation

   $$\dot S(t) = -U_{k+1}^\top JG_\varepsilon^{\mathcal{S}}(U_{k+1}S(t)U_{k+1}^\top)U_{k+1}, \qquad S(t_k) = \frac{MS_kM^\top}{\|MS_kM^\top\|_F},$$

   and set $S_{k+1} = S(t_{k+1})/\|S(t_{k+1})\|_F$.

The differential equations in the substeps are solved approximately by a step of some standard numerical integrator, e.g. the explicit Euler method or a low-order explicit Runge–Kutta method such as the second-order Heun method. The stepsize selection is done as in Section II.2.5, using an Armijo-type line search.

V.1.3 Non-imaginary eigenvalues close to coalescence on the imaginary axis

sec:eps

This theoretical section serves as a preparation for the algorithm of the outer iteration that will be presented in the next section. Let $E(\varepsilon)$ of unit Frobenius norm be a local minimizer of the optimization problem (1.1). We let $\lambda(\varepsilon)$ be the eigenvalue of smallest positive real part (and nonnegative imaginary part) of the Hamiltonian matrix

$$M(\varepsilon) := A + \varepsilon E(\varepsilon)$$

and $x(\varepsilon)$ and $y(\varepsilon)$ are corresponding left and right eigenvectors normalized by $\|x(\varepsilon)\| = \|y(\varepsilon)\| = 1$ and $x(\varepsilon)^* y(\varepsilon) > 0$. We let $\varepsilon_\star$ be the smallest value of ε such that

$$\phi(\varepsilon) = \text{Re}\,\lambda(\varepsilon)$$

becomes zero:

$$\phi(\varepsilon) > 0 \quad \text{for } 0 < \varepsilon < \varepsilon_\star \qquad \text{and} \qquad \phi(\varepsilon) = 0 \quad \text{for } \varepsilon \geq \varepsilon_\star \text{ near } \varepsilon_\star.$$

Under Assumption IV.2.1, the function ϕ is continuously differentiable in a left neighbourhood of $\varepsilon_\star$ and its derivative is given by Theorem IV.2.2. In the following we show that under further assumptions, the function ϕ has a square-root behavior $\phi(\varepsilon) \sim \sqrt{\varepsilon_\star - \varepsilon}$ as $\varepsilon \nearrow \varepsilon_\star$.

assumpt-epsstar

Assumption 1.4. We assume that the limit $M(\varepsilon_\star) := \lim_{\varepsilon \nearrow \varepsilon_\star} M(\varepsilon)$ of the Hamiltonian matrices exists and that the purely imaginary eigenvalue $\lambda(\varepsilon_\star) = \lim_{\varepsilon \nearrow \varepsilon_\star} \lambda(\varepsilon)$ of $M(\varepsilon_\star)$ has algebraic multiplicity two and is defective (that is, the zero singular value of $M(\varepsilon_\star) - \lambda(\varepsilon_\star) I$ is simple).

By definition of $\varepsilon_\star$, the eigenvalue $\lambda(\varepsilon_\star)$ is on the imaginary axis and has even multiplicity because of the symmetry of the eigenvalues with respect to the imaginary axis. Here we assume multiplicity two. The defectivity appears to be generic (we have no proof for this but observed defectivity in all our numerical experiments).

Under Assumption 1.4, the eigenvalue $\lambda(\varepsilon_\star)$ of $M(\varepsilon_\star)$ is non-derogatory, that is, only a single Jordan block corresponds to this eigenvalue, and hence its left and right eigenspaces are of dimension 1. Since $\lambda(\varepsilon_\star)$ is a defective eigenvalue, left and right eigenvectors at $\varepsilon_\star$ are orthogonal to each other: $x(\varepsilon_\star)^* y(\varepsilon_\star) = 0$.

We need the following result.

thm:yJx

Theorem 1.5 (Eigenvectors at coalescence). *Let $M(\varepsilon)$, $\varepsilon \in [\varepsilon_0, \varepsilon_\star]$, be a continuous path of real Hamiltonian matrices, and $\lambda(\varepsilon)$ be a path of eigenvalues of $M(\varepsilon)$ that are simple and not purely imaginary for $\varepsilon < \varepsilon_\star$ and satisfy Assumption 1.4 at $\varepsilon_\star$. Under a nondegeneracy condition on eigenvectors of $M(\varepsilon)$ stated in* (1.20) *below, there exist left and right eigenvectors $x(\varepsilon)$ and $y(\varepsilon)$ to the eigenvalue $\lambda(\varepsilon)$, normalized to unit norm and with $x(\varepsilon)^* y(\varepsilon) > 0$ for $\varepsilon < \varepsilon_\star$, which depend continuously on ε in the closed interval $[\varepsilon_0, \varepsilon_\star]$. In particular, the eigenvectors converge for $\varepsilon \nearrow \varepsilon_\star$. In the limit we have*

$$y(\varepsilon_\star) = \pm J x(\varepsilon_\star),$$

where the sign depends on $x(\varepsilon)$ for ε near $\varepsilon_\star$.

The important fact here is that $y(\varepsilon_\star)$ is not just a complex multiple of $Jx(\varepsilon_\star)$, as would easily be obtained from the symmetry of eigenvalues with respect to the imaginary axis, but that it is a *real* multiple.

Proof. By a result of Paige & Van Loan (1981), Theorem 5.1, the Hamiltonian matrix $M(\varepsilon)$ with no imaginary eigenvalues (for $\varepsilon < \varepsilon_\star$) admits a real Schur-Hamiltonian decomposition, that is, there exists an orthogonal symplectic real matrix $S(\varepsilon)$ (i.e., $S(\varepsilon)^\top S(\varepsilon) = I$ and $S(\varepsilon)^\top JS(\varepsilon) = J$) for $\varepsilon > \varepsilon_\star$) that transforms $M(\varepsilon)$ to a block triangular Hamiltonian matrix

$$M_0(\varepsilon) = S(\varepsilon)^{-1} M(\varepsilon) S(\varepsilon) = \begin{pmatrix} F(\varepsilon) & H(\varepsilon) \\ 0 & -F(\varepsilon)^\top \end{pmatrix}, \tag{1.9}$$

where $H(\varepsilon)$ is symmetric and $F(\varepsilon)$ is upper quasi-triangular.

For the eigenvalue $\lambda(\varepsilon)$, the left and right eigenvectors of $M_0(\varepsilon)$ are related to those of $M(\varepsilon)$ by

$$x_0(\varepsilon) = S(\varepsilon)^\top x(\varepsilon), \qquad y_0(\varepsilon) = S(\varepsilon)^{-1} y(\varepsilon). \tag{1.10} \qquad \boxed{\texttt{x-x0}}$$

We assume that $x(\varepsilon)$ and $y(\varepsilon)$ are normalized to norm 1 and such that $x(\varepsilon)^* y(\varepsilon) > 0$ for $\varepsilon > \varepsilon_\star$, and hence we have also

$$x_0(\varepsilon) \text{ and } y_0(\varepsilon) \text{ are of norm 1 and } \; x_0(\varepsilon)^* y_0(\varepsilon) > 0. \tag{1.11} \qquad \boxed{\texttt{x0y0pos}}$$

We observe that the lower half of the right eigenvector $y_0(\varepsilon)$ to the block triangular matrix $M_0(\varepsilon)$ consists only of zeros and we split the eigenvectors into the upper and the lower $n/2$-dimensional subvectors as

$$y_0(\varepsilon) = \begin{pmatrix} -p(\varepsilon) \\ 0 \end{pmatrix}, \qquad x_0(\varepsilon) = \begin{pmatrix} -s(\varepsilon) \\ r(\varepsilon) \end{pmatrix}. \tag{1.12} \qquad \boxed{\texttt{yxpsr}}$$

By the Hamiltonian symmetry, left and right eigenvectors associated with the eigenvalue $-\overline{\lambda(\varepsilon)}$, with positive inner product, are $\widetilde{x}_0(\varepsilon) = Jy_0(\varepsilon)$ and $\widetilde{y}_0(\varepsilon) = Jx_0(\varepsilon)$, and so we have

$$\widetilde{y}_0(\varepsilon) = \begin{pmatrix} r(\varepsilon) \\ s(\varepsilon) \end{pmatrix}, \qquad \widetilde{x}_0(\varepsilon) = \begin{pmatrix} 0 \\ p(\varepsilon) \end{pmatrix}. \tag{1.13}$$

By compactness, there exists a sequence (ε_n) with $\varepsilon_n \nearrow \varepsilon_\star$ as $n \to \infty$ such that $x_0(\varepsilon_n)$, $y_0(\varepsilon_n)$ and $S(\varepsilon_n)$ converge to vectors $x_{0,\star}$, $y_{0,\star}$ of norm 1 and an orthogonal symplectic real matrix $S_\star$. By the continuity of $M(\cdot)$ and $\lambda(\cdot)$ at $\varepsilon_\star$, the limit vectors $x_{0,\star}$, $y_{0,\star}$ are then left and right eigenvectors corresponding to the purely imaginary eigenvalue $\lambda(\varepsilon_\star)$ of $M(\varepsilon_\star)$.

By Assumption 1.4, the left and right eigenspaces to $\lambda(\varepsilon_\star)$ are one-dimensional, and so we have that for some complex ξ, η of unit modulus,

$$\lim_{n\to\infty} \widetilde{y}_0(\varepsilon_n) = -\eta \lim_{n\to\infty} y_0(\varepsilon_n), \qquad \lim_{n\to\infty} \widetilde{x}_0(\varepsilon_n) = \xi \lim_{n\to\infty} x_0(\varepsilon_n). \tag{1.14} \qquad \boxed{\texttt{y0-x0-limits}}$$

We thus obtain

$$\lim_{n\to\infty} s(\varepsilon_n) = 0 \tag{1.15} \qquad \boxed{\texttt{s-to-zero}}$$

and

$$\lim_{n\to\infty} r(\varepsilon_n) = \eta \lim_{n\to\infty} p(\varepsilon_n), \qquad \lim_{n\to\infty} p(\varepsilon_n) = \xi \lim_{n\to\infty} r(\varepsilon_n), \tag{1.16} \qquad \boxed{\texttt{eq:limits}}$$

so that

$$\xi = \bar{\eta}. \tag{1.17}$$

By (1.11),

$$s(\varepsilon)^* p(\varepsilon) \ \text{ is real and positive for } \varepsilon > \varepsilon_\star, \tag{1.18} \qquad \boxed{\texttt{sp-pos}}$$

and in particular, $s(\varepsilon) \neq 0$ for $\varepsilon > \varepsilon_\star$ (but recall (1.15)).

Moreover, from the fact that $x_0(\varepsilon)$ is a left eigenvalue of $M_0(\varepsilon)$, we infer that (omitting the argument ε in the next few lines) $s^* F = \lambda s^*$ and $-s^* H - r^* F^\top = \lambda r^*$. Multiplying the second equation with s from the right and using the first equation then yields $-s^* H s = (\lambda + \overline{\lambda})\, r^* s$, which shows that

$$s(\varepsilon)^* r(\varepsilon) \ \text{ is real for } \varepsilon > \varepsilon_\star. \tag{1.19} \qquad \boxed{\texttt{sr-real}}$$

Under the nondegeneracy condition

$$\liminf_{\varepsilon \nearrow \varepsilon_\star} \left| \left(\frac{s(\varepsilon)}{\|s(\varepsilon)\|} \right)^* \frac{r(\varepsilon)}{\|r(\varepsilon)\|} \right| > 0, \tag{1.20} \qquad \boxed{\texttt{nondeg-condition}}$$

which states that the normalizations of the vectors s and r are not asymptotically orthogonal, we conclude that there is a subsequence (ε_n') of (ε_n) such that the normalized sequence $\bigl(s(\varepsilon_n')/\|s(\varepsilon_n')\|\bigr)$ is convergent and (on noting that $\|r(\varepsilon_n)\| \to 1$ because of (1.12) and (1.15))

$$\lim_{n\to\infty} \frac{s(\varepsilon_n')^* r(\varepsilon_n')}{\|s(\varepsilon_n')\|} \neq 0. \tag{1.21} \qquad \boxed{\texttt{sp-conv}}$$

As (1.16) implies that this nonzero limit equals

$$\lim_{n\to\infty} \frac{s(\varepsilon_n')^* r(\varepsilon_n')}{\|s(\varepsilon_n')\|} = \eta \lim_{n\to\infty} \frac{s(\varepsilon_n')^* p(\varepsilon_n')}{\|s(\varepsilon_n')\|}$$

and the two limits in this formula are real by (1.18) and (1.19), it follows that η is real and hence η equals 1 or -1. In view of (1.18) and (1.20), we actually have

$$\eta = \lim_{\varepsilon \nearrow \varepsilon_\star} \operatorname{sign}(s(\varepsilon)^* r(\varepsilon)) = \pm 1, \tag{1.22} \qquad \boxed{\texttt{eta-sr}}$$

which depends only on the left eigenvector $x_0(\varepsilon)$. As a consequence, we obtain from (1.14) that

$$y_{0,\star} = -\eta J x_{0,\star} = \mp J x_{0,\star}. \tag{1.23} \qquad \boxed{\texttt{eq:limy0}}$$

By (1.10) we have

$$x(\varepsilon) = \bigl(S(\varepsilon)^\top\bigr)^{-1} x_0(\varepsilon) = -J S(\varepsilon) J x_0(\varepsilon), \qquad y(\varepsilon) = S(\varepsilon) y_0(\varepsilon) = \pm S(\varepsilon) J x_0(\varepsilon) \tag{1.24}$$

and therefore the limits $x_\star = \lim_{n\to\infty} x(\varepsilon_n)$ and $y_\star = \lim_{n\to\infty} y(\varepsilon_n)$ exist and satisfy

$$y_\star = \pm J x_\star. \tag{1.25}$$

eq:limy

We now use once again that by Assumption 1.4, the left and right eigenspaces to $\lambda(\varepsilon_\star)$ are one-dimensional. Hence $x_\star$ is a complex multiple of the unique left eigenvector $x(\varepsilon_\star)$ of norm 1 for which the first nonzero entry is positive. If we choose the eigenvectors $x(\varepsilon)$ such that their corresponding entry is also nonnegative, then we find that every convergent subsequence $(x(\varepsilon_n))$ converges to the same limit $x(\varepsilon_\star)$ as $n \to \infty$, and hence $x(\varepsilon)$ converges to $x(\varepsilon_\star)$ as $\varepsilon \nearrow \varepsilon_\star$. To the left eigenvector $x(\varepsilon)$, there corresponds a unique right eigenvector $y(\varepsilon)$ of norm 1 that satisfies $x(\varepsilon)^* y(\varepsilon) > 0$ for $\varepsilon > \varepsilon_\star$. By (1.25), the limit of every convergent subsequence $(y(\varepsilon_n))$ converges to $\pm J \lim_{n\to\infty} x(\varepsilon_n) = \pm J x(\varepsilon_\star)$, and hence the limit $y(\varepsilon_\star) := \lim_{\varepsilon\nearrow\varepsilon_\star} y(\varepsilon)$ exists, is a right eigenvector of $M(\varepsilon_\star)$ to the eigenvalue $\lambda(\varepsilon_\star)$, and it equals

$$y(\varepsilon_\star) = \lim_{\varepsilon\nearrow\varepsilon_\star} y(\varepsilon) = \pm J \lim_{\varepsilon\nearrow\varepsilon_\star} x(\varepsilon) = \pm J x(\varepsilon_\star),$$

which completes the proof. □

Remark 1.6. If additionally $M(\varepsilon)$, $\varepsilon \in [\varepsilon_0, \varepsilon_\star]$, is continuously differentiable and $\operatorname{Re} x(\varepsilon_\star)^* M'(\varepsilon_\star) y(\varepsilon_\star) \neq 0$, then Theorem 1.5 implies that the eigenvalue $\lambda(\varepsilon)$ approaches the imaginary axis in normal direction (i.e. horizontally in the complex plane). This is because then we have, with $\kappa(\varepsilon) = 1/(x(\varepsilon)^* y(\varepsilon)) > 0$,

$$\begin{aligned} \operatorname{Im} \frac{\lambda'(\varepsilon)}{\kappa(\varepsilon)} &= \operatorname{Im} x(\varepsilon)^* M'(\varepsilon) y(\varepsilon) \to \operatorname{Im} x(\varepsilon_\star)^* M'(\varepsilon_\star) y(\varepsilon_\star) \\ &= \operatorname{Im} (Jx(\varepsilon_\star))^* J M'(\varepsilon_\star) y(\varepsilon_\star) = \pm \operatorname{Im} y(\varepsilon_\star)^* J M'(\varepsilon_\star) y(\varepsilon_\star) = 0 \end{aligned}$$

by the symmetry of $JM'(\varepsilon_\star)$. By assumption,

$$\operatorname{Re} \frac{\lambda'(\varepsilon)}{\kappa(\varepsilon)} = \operatorname{Re} x(\varepsilon)^* M'(\varepsilon) y(\varepsilon) \to \operatorname{Re} x(\varepsilon_\star)^* M'(\varepsilon_\star) y(\varepsilon_\star) \neq 0.$$

Hence, $\operatorname{Im} \lambda'(\varepsilon) \,/\, \operatorname{Re} \lambda'(\varepsilon) \to 0$ as $\varepsilon \nearrow \varepsilon_\star$. This can, however, not be concluded when $M'(\varepsilon)$ has no limit at $\varepsilon_\star$ and $\|M'(\varepsilon)\| \to \infty$ as $\varepsilon \nearrow \varepsilon_\star$.

We are now in a position to characterize the asymptotic behaviour of the function $\phi(\varepsilon) = \operatorname{Re} \lambda(\varepsilon)$ as $\varepsilon \nearrow \varepsilon_\star$, in the situation described at the beginning of this section.

thm:sqrt

Theorem 1.7 (Square root asymptotics). *Under Assumptions IV.2.1 and 1.4 and the nondegeneracy condition* (1.20)*, and under the further condition that the eigenvalue $\lambda(\varepsilon)$ of the Hamiltonian matrix $M(\varepsilon) = A + \varepsilon E(\varepsilon)$ does not approach the imaginary axis tangentially as $\varepsilon \nearrow \varepsilon_\star$, we have*

$$\operatorname{Re} \lambda(\varepsilon) = \gamma \sqrt{\varepsilon_\star - \varepsilon}\, (1 + o(1)) \quad \textit{as } \varepsilon \nearrow \varepsilon_\star$$

for some positive constant γ.

Proof. We split the proof into four parts (a)-(d).
(a) For $\varepsilon < \varepsilon_\star$, let $x(\varepsilon)$ and $y(\varepsilon)$ be left and right eigenvectors of $M(\varepsilon) = A + \varepsilon E(\varepsilon)$ to the simple eigenvalue $\lambda(\varepsilon)$, of unit norm and with $x(\varepsilon)^* y(\varepsilon) > 0$ for $\varepsilon < \varepsilon_\star$ and normalized such that their limits for $\varepsilon \nearrow \varepsilon_\star$ exist according to Theorem 1.5. We consider the nonnegative function

$$\vartheta(\varepsilon) := \frac{1}{\kappa(\varepsilon)} = x(\varepsilon)^* y(\varepsilon) > 0 \ \text{ for } \varepsilon \in (\varepsilon_0, \varepsilon_\star), \qquad \vartheta(\varepsilon_\star) = 0.$$

To compute the derivative of ϑ, we use Theorem VIII.1.5 for the left and right eigenvectors of norm 1 and with positive inner product,

$$\begin{aligned} x'(\varepsilon)^* &= -x(\varepsilon)^* M'(\varepsilon) Z(\varepsilon) + \mathrm{Re}\big(x(\varepsilon)^* M'(\varepsilon) Z(\varepsilon) x(\varepsilon)\big) x(\varepsilon)^* \\ y'(\varepsilon) &= -Z(\varepsilon) M'(\varepsilon) y(\varepsilon) + \mathrm{Re}\big(y(\varepsilon)^* Z(\varepsilon) M'(\varepsilon) y(\varepsilon)\big) y(\varepsilon), \end{aligned}$$

where $Z(\varepsilon)$ is the group inverse of $N(\varepsilon) := M(\varepsilon) - \lambda(\varepsilon) I$. Since Theorem VIII.1.4 shows that $x(\varepsilon)^* Z(\varepsilon) = 0$ and $Z(\varepsilon) y(\varepsilon) = 0$, these formulas imply

$$\vartheta'(\varepsilon) = \mathrm{Re}\Big(x(\varepsilon)^* M'(\varepsilon) Z(\varepsilon) x(\varepsilon) + y(\varepsilon)^* Z(\varepsilon) M'(\varepsilon) y(\varepsilon)\Big)\vartheta(\varepsilon). \tag{1.26}$$

eq:derdelta

By Theorem VIII.1.4, the group inverse is related to the pseudoinverse $N(\varepsilon)^\dagger$ by the formulas

$$\begin{aligned} Z(\varepsilon) &= \frac{1}{\vartheta(\varepsilon)^2} \widehat{Z}(\varepsilon) \\ \widehat{Z}(\varepsilon) &= \big(\vartheta(\varepsilon) I - y(\varepsilon) x(\varepsilon)^*\big) N(\varepsilon)^\dagger \big(\vartheta(\varepsilon) I - y(\varepsilon) x(\varepsilon)^*\big). \end{aligned} \tag{1.27}$$

By Assumption 1.4, the second smallest singular value $\sigma_{n-1}(\varepsilon)$ of $N(\varepsilon)$ does not converge to zero. Therefore, $N(\varepsilon)^\dagger$ has a finite limit as $\varepsilon \nearrow \varepsilon_\star$. We thus have

$$\begin{aligned} \widehat{Z}(\varepsilon) &= y(\varepsilon) x(\varepsilon)^* N(\varepsilon)^\dagger y(\varepsilon) x(\varepsilon)^* + O(\vartheta(\varepsilon)) \qquad (1.28) \\ &= \nu(\varepsilon) y(\varepsilon) x(\varepsilon)^* + O(\vartheta(\varepsilon)) \end{aligned}$$

with the factor

$$\nu(\varepsilon) := x(\varepsilon)^* N(\varepsilon)^\dagger y(\varepsilon),$$

Furthermore, we set

$$\mu(\varepsilon) := x(\varepsilon)^* M'(\varepsilon) y(\varepsilon).$$

We insert the expression for the group inverse $Z(\varepsilon)$ into (1.26) and note the identities $N^\dagger(\varepsilon) x(\varepsilon) = 0$ and $y(\varepsilon)^* N^\dagger(\varepsilon) = 0$, which follow from $x(\varepsilon)^* N(\varepsilon) = 0$ and $N(\varepsilon) y(\varepsilon) = 0$, respectively. We then obtain

$$\begin{aligned} \vartheta'(\varepsilon)\vartheta(\varepsilon) &= \mathrm{Re}\Big(x(\varepsilon)^* M'(\varepsilon) \widehat{Z}(\varepsilon) x(\varepsilon) + y(\varepsilon)^* \widehat{Z}(\varepsilon) M'(\varepsilon) y(\varepsilon)\Big) \\ &= \mathrm{Re}\Big(2\nu(\varepsilon)\mu(\varepsilon) + O(\vartheta(\varepsilon)\mu(\varepsilon))\Big). \end{aligned} \tag{1.29}$$

vartheta-prime

(b) We now study the limit behaviour of $\nu(\varepsilon)$ as $\varepsilon \nearrow \varepsilon_\star$. By Theorem 1.5, the limits of the left and right eigenvectors for $\varepsilon \nearrow \varepsilon_\star$ exist and satisfy $y(\varepsilon_\star) = \pm J x(\varepsilon_\star)$ and further $x(\varepsilon_\star)^* y(\varepsilon_\star) = 0$. Since $JN(\varepsilon_\star)$ is a hermitian matrix, we therefore obtain

$$\begin{aligned}\nu(\varepsilon_\star) &= x(\varepsilon_\star)^* N(\varepsilon_\star)^\dagger y(\varepsilon_\star) = x(\varepsilon_\star)^* (JN(\varepsilon_\star))^\dagger J y(\varepsilon_\star)\\ &= \mp\, x(\varepsilon_\star)^* (JN(\varepsilon_\star))^\dagger x(\varepsilon_\star) \in \mathbb{R}.\end{aligned}$$

We next show that $\nu(\varepsilon_\star) \neq 0$. Since $x(\varepsilon_\star) \perp y(\varepsilon_\star)$ and since $x(\varepsilon_\star)$ spans the nullspace of $N(\varepsilon_\star)^*$ (by the defectivity condition in Assumption 1.4), we obtain

$$y(\varepsilon_\star) \in \mathrm{Ker}(N(\varepsilon_\star)^*)^\perp = \mathrm{Range}\,(N(\varepsilon_\star))\,.$$

Hence there exists z_1 such that $y(\varepsilon_\star) = N(\varepsilon_\star) z_1$. Assume, in a proof by contradiction, $N(\varepsilon_\star)^\dagger y(\varepsilon_\star) \perp x(\varepsilon_\star)$, which means

$$N(\varepsilon_\star)^\dagger y(\varepsilon_\star) \in \mathrm{Range}\,(N(\varepsilon_\star))\,.$$

Hence there exists z_2 such that $N(\varepsilon_\star)^\dagger y(\varepsilon_\star) = N(\varepsilon_\star) z_2$.

Multiplying this equation with $N(\varepsilon_\star)^2$ we obtain (omitting the argument $\varepsilon_\star$ in the following)

$$N^3 z_2 = N^2 N^\dagger y = N\, N N^\dagger N z_1 = N N z_1 = N y = 0.$$

The null-space of N^3 is two-dimensional by Assumption 1.4 and contains the two nonzero vectors y and $N^\dagger y$, since $Ny = 0$ and $N^2 N^\dagger y = 0$. Note that $N^\dagger y \neq 0$ because otherwise y would be in the nullspace of $N^\dagger$, which is the nullspace of N^*, which contradicts the above observation that $y \neq 0$ is in the orthogonal complement of the nullspace of N^*. Moreover, y and $N^\dagger y$ are linearly independent, since otherwise the relation $y = cN^\dagger y$ would yield, on multiplication with N, that

$$0 = Ny = cNN^\dagger y = cNN^\dagger N z_1 = cNz_1 = cy,$$

which contradicts $y \neq 0$. Therefore, the null-space of N^3 is *spanned* by y and $N^\dagger y$, and since we have shown that it contains z_2, we obtain

$$z_2 = c_1 y + c_2 N^\dagger y.$$

Multiplying this equation with N then gives

$$N^\dagger y = N z_2 = c_2 N N^\dagger N z_1 = c_2 N z_1 = c_2 y,$$

which contradicts the linear independence of y and $N^\dagger y$. We have thus led the assumption $N(\varepsilon_\star)^\dagger y(\varepsilon_\star) \perp x(\varepsilon_\star)$ to a contradiction. Therefore, $\nu(\varepsilon_\star) \neq 0$. So we have shown that

$$\nu_\star := \lim_{\varepsilon \nearrow \varepsilon_\star} \nu(\varepsilon) \quad \text{exists and is real and nonzero.} \qquad (1.30)$$

nu-star

(c) We next study the limit behaviour of $\mu(\varepsilon)$ as $\varepsilon \nearrow \varepsilon_\star$. By Theorem VIII.1.1 we have

$$\mu(\varepsilon) = \lambda'(\varepsilon)\vartheta(\varepsilon).$$

By Theorem IV.2.2 with the objective function $\phi(\varepsilon) = \mathrm{Re}\,\lambda(\varepsilon)$ and the gradient $Z(\varepsilon) = \mathrm{Sym}(\mathrm{Re}\, Jx(\varepsilon)y(\varepsilon)^*)$, we thus have

$$\mathrm{Re}\,\mu(\varepsilon) = -\|Z(\varepsilon)\|_F.$$

Since $y(\varepsilon_\star) = \pm Jx(\varepsilon_\star)$ by Theorem 1.5, we have in the limit $\varepsilon \nearrow \varepsilon_\star$ that

$$G(\varepsilon_\star) = \mathrm{Sym}(\mathrm{Re}\, Jx(\varepsilon_\star)y(\varepsilon_\star)^*) = \pm\mathrm{Re}\, y(\varepsilon_\star)y(\varepsilon_\star)^* \neq 0.$$

By assumption, $\lambda(\varepsilon)$ does not approach the imaginary axis tangentially, and hence we have

$$\frac{|\mathrm{Im}\,\mu(\varepsilon)|}{|\mathrm{Re}\,\mu(\varepsilon)|} = \frac{|\mathrm{Im}\,\lambda'(\varepsilon)|}{|\mathrm{Re}\,\lambda'(\varepsilon)|} \leq C \tag{1.31}$$

Im-mu

for some constant C independent of $\varepsilon \in (\varepsilon_0, \varepsilon_\star)$. This implies that $|\mathrm{Im}\,\mu(\varepsilon)|$ is bounded independently of ε. In the following we let

$$\rho_\star := -\lim_{\varepsilon \nearrow \varepsilon_\star} \mathrm{Re}\,\mu(\varepsilon) = \|G(\varepsilon_\star)\|_F > 0. \tag{1.32}$$

mu-G

(d) From (1.30)–(1.32) we conclude that the right-hand side of (1.29) has a nonzero finite real limit as $\varepsilon \nearrow \varepsilon_\star$. So we have

$$\frac{d}{d\varepsilon}\vartheta(\varepsilon)^2 = 2\vartheta'(\varepsilon)\vartheta(\varepsilon) = -4\rho_\star\nu_\star(1 + o(1)) \quad \text{as } \varepsilon \nearrow \varepsilon_\star.$$

Integrating this relation and using $\vartheta(\varepsilon_\star)^2 = 0$ yields

$$\vartheta(\varepsilon)^2 = (\varepsilon_\star - \varepsilon)\, 4\rho_\star\nu_\star(1 + o(1)).$$

This further allows us to conclude that $\nu_\star$ is not only nonzero and real but actually positive. We recall that $\vartheta(\varepsilon) > 0$ for $\varepsilon < \varepsilon_\star$ and take the square root to obtain

$$\vartheta(\varepsilon) = \sqrt{\varepsilon_\star - \varepsilon}\, 2\sqrt{\rho_\star\nu_\star}\,(1 + o(1)). \tag{1.33}$$

vartheta-asymptote

On the other hand, since $\mu(\varepsilon) = \lambda'(\varepsilon)\vartheta(\varepsilon)$, we find

$$\mathrm{Re}\,\lambda'(\varepsilon) = \frac{\mathrm{Re}\,\mu(\varepsilon)}{\vartheta(\varepsilon)} = \frac{-\rho_\star}{\vartheta(\varepsilon)}(1 + o(1)).$$

Using (1.33) and setting $\gamma = \sqrt{\rho_\star/\nu_\star}$, this yields

$$\mathrm{Re}\,\lambda'(\varepsilon) = -\frac{\gamma}{2\sqrt{\varepsilon_\star - \varepsilon}}(1 + o(1)), \tag{1.34}$$

phi-prime

and integration then implies the stated result for $\mathrm{Re}\,\lambda(\varepsilon)$. □

V.1.4 Outer iteration: Square root model and bisection

c:outer-it-sqrt

For a small positive parameter δ, we aim to find ε_δ as the minimal number such that

$$\operatorname{Re}\lambda(\varepsilon_\delta) = \delta. \tag{1.35}$$

eps-delta

We can use the Newton/bisection method of Section IV.2.3 to compute ε_δ, but for small δ, this can lead to many Newton step rejections and bisection steps. In the situation of Theorem 1.7, using a square root model of $\phi(\varepsilon) = \operatorname{Re}\lambda(\varepsilon)$ appears more appropriate. The algorithm falls back to simple bisection if the local square root model fails.

For $\varepsilon \nearrow \varepsilon_\star$, we have in the expected situation of Theorem 1.7 the square-root behaviour for $\phi(\varepsilon) = \operatorname{Re}\lambda(\varepsilon)$ (and by (1.34) for $\phi'(\varepsilon)$)

$$\begin{aligned} \phi(\varepsilon) &= \gamma\sqrt{\varepsilon_\star - \varepsilon}\,(1 + o(1)) \\ \phi'(\varepsilon) &= -\frac{\gamma}{2\sqrt{\varepsilon_\star - \varepsilon}}\,(1 + o(1)). \end{aligned} \tag{1.36}$$

For an iterative process, given ε_k, we use that $\phi'(\varepsilon_k) = -\kappa(\varepsilon)\|Z(\varepsilon)\|_F$ by Theorem IV.2.2 and solve (1.36) for γ and $\varepsilon_\star$, ignoring the $o(1)$ terms. We denote the solution as γ_k and $\widehat{\varepsilon}_k$, i.e.,

$$\gamma_k = \sqrt{-2\phi(\varepsilon_k)\phi'(\varepsilon_k)}, \qquad \widehat{\varepsilon}_k = \varepsilon_k - \frac{\phi(\varepsilon_k)}{2\phi'(\varepsilon_k)}. \tag{1.37}$$

As a substitute for the equation $\phi(\varepsilon) = \delta$, we solve the equation $\gamma_k\sqrt{\widehat{\varepsilon}_k - \varepsilon_{k+1}} = \delta$ for ε_{k+1}, which yields

$$\varepsilon_{k+1} = \widehat{\varepsilon}_k + \delta^2/\gamma_k^2. \tag{1.38}$$

Algorithm 10 is based on these formulas. Here, tol is a tolerance that controls the desired accuracy of the computed optimal ε (not to be chosen too small).

Algorithm 10: Basic algorithm for computing the optimal perturbation for small δ

alg:problemA

```
Data: $\delta$, tol, $\theta$ (default 0.8), and $\varepsilon_0$ (such that $\phi(\varepsilon_0) > \mathrm{tol}$)
Result: $\widehat{\varepsilon}_\delta$, $E(\widehat{\varepsilon}_\delta)$
begin
    Set Reject = False and $k = 0$
    while $|\phi(\varepsilon_k) - \delta| \geq \mathrm{tol}$ do
        if $\mathrm{Reject} = \mathrm{False}$ then
            Set $\widetilde{\varepsilon} = \varepsilon_k,\ \widetilde{\theta} = \theta$
            Compute $\gamma_k$ and $\widehat{\varepsilon}_k$ by (1.37)
            Set $\varepsilon_{k+1} = \widehat{\varepsilon}_k + \delta^2/\gamma_k^2$
        else
            Set $\varepsilon_{k+1} = \widetilde{\theta}\,\varepsilon_k + (1 - \widetilde{\theta})\,\widetilde{\varepsilon}$
            Set $\widetilde{\theta} = \theta\widetilde{\theta}$
        Set $k = k + 1$
        Compute $\phi(\varepsilon_k)$ by solving the rank-4 differential equation (1.6) with
          initial datum $E(\varepsilon_{k-1})$ into a stationary point $E(\varepsilon_k)$ as in Section V.1.2
        Compute $\phi'(\varepsilon_k)$ by (IV.2.3)
        if $\phi(\varepsilon_k) < \mathrm{tol}$ then
            Set $\mathrm{Reject} = \mathrm{True}$
        else
            Set $\mathrm{Reject} = \mathrm{False}$
    Return $\widehat{\varepsilon}_\delta = \varepsilon_k$
```

V.1.5 Problem B: Eigenvalues leaving the imaginary axis

subsec:ham-B

We describe two complementary approaches to Problem B. The first approach moves eigenvalues on the imaginary axis, and the second approach moves eigenvalues off the imaginary axis.

Moving eigenvalues on the imaginary axis to coalescence. Because of the symmetry of eigenvalues of Hamiltonian matrices with respect to the imaginary axis, paths of eigenvalues can leave the imaginary axis only at multiple eigenvalues. Given a Hamiltonian matrix with some simple eigenvalues on the imaginary axis, it is thus of interest to find its distance to the nearest matrix where two previously adjacent eigenvalues coalesce. This problem is addressed by an extension of the two-level approach considered before.

Let A be a real Hamiltonian matrix with a pair $\lambda_1(A)$ and $\lambda_2(A)$ of adjacent eigenvalues on the imaginary axis, with $\mathrm{Im}\,\lambda_2(A) > \mathrm{Im}\,\lambda_1(A)$. In the inner iteration we determine, for a fixed perturbation size $\varepsilon > 0$, a real Hamiltonian matrix E of Frobenius norm 1 such that the functional

$$F_\varepsilon(E) = \mathrm{Im}\,\lambda_2(A + \varepsilon E) - \mathrm{Im}\,\lambda_1(A + \varepsilon E)$$

is minimized. As in previous sections, this minimization is carried out with a constrained gradient flow. Along a path $E(t)$ of real Hamiltonian matrices with simple eigenvalues $\lambda_k(t) = \lambda_k(A + \varepsilon E(t))$ $(k = 1, 2)$ on the imaginary axis, corresponding left and right eigenvectors $x_k(t), y_k(t)$ of unit norm with positive inner product, and the eigenvalue condition numbers $\kappa_k(t) = 1/(x_k(t)^* y_k(t))$ we find (omitting the ubiquitous argument t)

$$\begin{aligned}\frac{d}{dt} F_\varepsilon(E(t)) &= \mathrm{Im}(\dot\lambda_2 - \dot\lambda_1) = \mathrm{Im}(\kappa_2 x_2^* \dot E y_2 - \kappa_1 x_1^* \dot E y_1)\\ &= \langle \mathrm{Im}(\kappa_2 x_2 y_2^* - \kappa_1 x_1 y_1^*), \dot E\rangle\\ &= \langle G_\varepsilon(E), \dot E\rangle\end{aligned}$$

with the Hamiltonian gradient

$$G_\varepsilon(E) = \mathit{\Pi}_{\mathcal{S}}\, \mathrm{Im}(\kappa_2 x_2 y_2^* - \kappa_1 x_1 y_1^*) = J^{-1}\mathrm{Sym}\big(\mathrm{Im}(\kappa_2 J x_2 y_2^* - \kappa_1 J x_1 y_1^*)\big),$$

which has rank at most 8. The corresponding norm-constrained gradient system is then again

$$\dot E = -G_\varepsilon(E) + \langle G_\varepsilon(E), E\rangle E,$$

along which $F_\varepsilon(E(t))$ decreases monotonically. In a stationary point, E is a real multiple of $G_\varepsilon(E)$, which is of rank at most 8. We can then solve numerically the rank-8 constrained gradient system into a stationary point in the same way as we did with the rank-4 system in Section V.1.2. In the outer iteration we aim to determine the smallest zero $\varepsilon_\star$ of $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon))$, where $E(\varepsilon)$ is the minimizer corresponding to the perturbation size ε. This is again done by a Newton/bisection algorithm (or using a square root model and bisection) as discussed before.

We note, however, that a coalescence on the imaginary axis does not guarantee that the coalescent eigenvalues can be moved off the imaginary axis by an arbitrarily small further perturbation; see Mehrmann & Xu (2008), Theorem 3.2. Moreover, mere coalescence on the imaginary axis does not give an answer to the related problem of finding a smallest perturbation that moves the adjacent imaginary eigenvalues to a prescribed positive distance δ to the imaginary axis.

Moving non-imaginary eigenvalues of perturbed Hamiltonian matrices back to coalescence on the imaginary axis. In a complementary approach, we first perturb the given real Hamiltonian matrix A, which is assumed to have some eigenvalues on the imaginary axis, to another Hamiltonian matrix $A_0 = A + \varepsilon_0 E_0$ (with $\|E_0\|_F = 1$) that has no eigenvalues on the imaginary axis, but which is not the one that is closest to A. With a sufficiently large perturbation size ε_0, this is always possible; just take $A_0 = \mathrm{blockdiag}(B, -B^\top)$, where $B \in \mathbb{R}^{d,d}$ is an arbitrary matrix having no purely imaginary eigenvalues. For example, one might choose B as the left upper block of A, if this has no imaginary eigenvalues, or else slightly shifted to have no imaginary eigenvalue. We remark that in our numerical experiments, the choice of A_0 was not a critical issue. Starting from A_0, we reduce the perturbation size to $\varepsilon < \varepsilon_0$ and in this way drive eigenvalues back to the imaginary axis.

We aim to find the largest perturbation size ε for which $A+\varepsilon E$ has some eigenvalue on the imaginary axis for *every* matrix E of Frobenius norm 1. This differs from Problem A, where the aim was to find the smallest perturbation size ε for which $A+\varepsilon E$ (with A having no purely imaginary eigenvalues) has eigenvalues on the imaginary axis for *some* matrix E of Frobenius norm 1.

As in Section V.1.1, the target eigenvalue $\lambda(M)$ of a Hamiltonian matrix M is taken as an eigenvalue of minimal real part in the first quadrant. In the inner iteration we use a rank-4-constrained gradient system to compute, for a given perturbation size $\varepsilon > 0$, a real Hamiltonian matrix $E(\varepsilon)$ of Frobenius norm 1 such that $\operatorname{Re}\lambda(A+\varepsilon E)$ is (locally) *maximized* (as opposed to *minimized* for Problem A):

$$E(\varepsilon) = \arg \max_{E \in \mathrm{Ham}(\mathbb{R}^{n,n}), \|E\|_F = 1} \operatorname{Re}\lambda(A+\varepsilon E). \qquad (1.39)$$

E-eps-ham-max

The details of the algorithm are nearly identical to Section V.1.1, except that the sign of the right-hand sides of the differential equations (1.4), (1.5) and (1.6) is switched, or in other words, we go backward in time with the same differential equations.

In the outer iteration we compute, for a given small $\delta > 0$, the perturbation size ε_δ as the largest ε with $\operatorname{Re}\lambda(A+\varepsilon E(\varepsilon)) = \delta$, in the same way as in Section V.1.4.

sec:symplectic

V.2 Nearest defective real matrix

sec:defective

Let A be a *real* $n \times n$ matrix with n distinct eigenvalues. A classical problem, known as Wilkinson problem, is that of determining the nearest matrix to A with a *defective* multiple eigenvalue, that is, the Jordan canonical form has a non-diagonal block. Note that a Jordan block is non-diagonal if and only if the left and right eigenvectors of the Jordan block are orthogonal to each other.

Here we restrict the problem to the space of real matrices. We are interested in computing the following distance:

$$w_{\mathbb{R}}(A) \quad = \quad \inf\bigl\{ \|\Delta\|_F \,:\, \Delta \in \mathbb{R}^{n,n} \text{ is such that } A+\Delta \text{ is defective} \bigr\}, \qquad (2.1)$$

i.e. the Frobenius-norm distance of A to the set of defective real matrices. This can be interpreted as the distance to singularity of the eigenvalue condition number $\kappa = 1/(x^*y)$, where x and y are left and right eigenvectors of unit norm and with positive inner product. This notion of eigenvalue condition number is due to Wilkinson (1965).

V.2.1 Two-level approach

For $0 < \varepsilon < w_{\mathbb{R}}(A)$ we introduce the functional $F_\varepsilon(E)$ (for matrices $E \in \mathbb{C}^{n,n}$, which will later be restricted to be real and of unit Frobenius norm) as follows: Let $\lambda = \lambda(A+\varepsilon E)$ be a simple target eigenvalue of $A+\varepsilon E$ and let x and y be corresponding left and right eigenvectors, respectively, normalized to unit norm and with positive inner product. We set

$$F_\varepsilon(E) = x^* y > 0. \tag{2.2}$$

F-eps-def

In contrast to previous sections, the functional to be minimized now depends on eigenvectors instead of eigenvalues.

We follow the two-level approach of Section IV.2:

- **Inner iteration:** Given $\varepsilon > 0$, we aim to compute a matrix $E(\varepsilon) \in \mathbb{R}^{n,n}$ of unit Frobenius norm that minimizes F_ε:

$$E(\varepsilon) = \arg \min_{E \in \mathbb{R}^{n,n}, \|E\|_F = 1} F_\varepsilon(E). \tag{2.3}$$

E-eps-def

- **Outer iteration:** For a small threshold $\delta > 0$, we compute the smallest positive value ε_δ with

$$\phi(\varepsilon_\delta) = \delta, \tag{2.4}$$

zero-def

where $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon)) = x(\varepsilon)^* y(\varepsilon)$, where $x(\varepsilon)$ and $y(\varepsilon)$ are left and right eigenvectors of $A + \varepsilon E(\varepsilon)$ associated with $\lambda(\varepsilon)$, of unit norm and with positive inner product.

Provided that these computations succeed, we then expect that $\Delta A_\delta = \varepsilon_\delta E(\varepsilon_\delta) \in \mathbb{R}^{n,n}$ makes $A + \Delta A$ close to a defective matrix and that, for an appropriate choice of the target eigenvalue, the limit $\varepsilon_\star = \lim_{\delta \searrow 0} \varepsilon_\delta$ exists and is equal to the distance $w_{\mathbb{R}}(A)$ of the matrix A to the set of defective real matrices. These steps are detailed in the following subsections.

V.2.2 Constrained gradient flow for the inner iteration

As in Section II.2, we begin by computing the free (complex) and real gradients of the functional F_ε.

gradient-mp-cnv

Lemma 2.1 (Free gradient). *Let $E(t) \in \mathbb{C}^{n,n}$, for real t near t_0, be a continuously differentiable path of matrices, with the derivative denoted by $\dot{E}(t)$. Assume that $\lambda(t)$ is a simple eigenvalue of $A + \varepsilon E(t)$ depending continuously on t, with corresponding left and right eigenvectors $x(t)$ and $y(t)$, respectively, which are taken to be of unit norm and with positive inner product. Let $\kappa(t) = 1/(x(t)^* y(t))$. Then, $F_\varepsilon(E(t)) = x(t)^* y(t)$ is continuously differentiable w.r.t. t and*

$$\frac{\kappa(t)}{\varepsilon} \frac{d}{dt} F_\varepsilon(E(t)) = \mathrm{Re}\big\langle G_\varepsilon(E(t)), \dot{E}(t) \big\rangle, \tag{2.5}$$

eq:deriv-def

where the rescaled gradient of F_ε is a matrix of rank at most 2, given by

$$G_\varepsilon(E) = x x^* Z^* + Z^* y y^*. \tag{2.6}$$

freegrad-def

Here, Z is the group inverse of $A + \varepsilon E - \lambda I$ (see Chapter VIII) for the eigenvalue $\lambda = \lambda(A + \varepsilon E)$, and x and y are the left and right normalized eigenvectors with positive inner product.

Proof. By (VIII.1.5) and using that $Zy=0$, $x^*Z=0$, we get

$$\begin{aligned}\frac{d}{dt}(x^*y) &= \mathrm{Re}\big(\dot{x}^*y + x^*\dot{y}\big) = \\ &= \varepsilon\,(x^*y)\,\mathrm{Re}\big(x^*\dot{E}Zx + y^*Z\dot{E}y\big) \\ &= \varepsilon\,(x^*y)\,\mathrm{Re}\,\langle xx^*Z^* + Z^*yy^*, \dot{E}\rangle = \varepsilon\,(x^*y)\,\mathrm{Re}\,\langle G_\varepsilon(E), \dot{E}\rangle\ ,\end{aligned}$$

from which (2.5) follows. □

Real gradient. For a path of *real* matrices $E(t) \in \mathbb{R}^{n,n}$, also $\dot{E}(t)$ is real, and hence the right-hand side of (2.5) becomes $\big\langle \mathrm{Re}\,G_\varepsilon(E(t)), \dot{E}(t)\big\rangle$ with the real inner product. With the real gradient

$$G_\varepsilon^{\mathbb{R}}(E) := \mathrm{Re}\,G_\varepsilon(E) = \mathrm{Re}(xx^*Z^* + Z^*yy^*), \tag{2.7}$$

real-grad-def

which is a matrix of rank at most 4, we then have

$$\frac{\kappa(t)}{\varepsilon}\,\frac{d}{dt}F_\varepsilon(E(t)) = \big\langle G_\varepsilon^{\mathbb{R}}(E(t)), \dot{E}(t)\big\rangle. \tag{2.8}$$

eq:deriv-real-def

With this real gradient, we now follow closely the programme of Section II.2.

Norm-constrained real gradient flow. We consider the gradient flow on the manifold of matrices in $\mathbb{R}^{n,n}$ of unit Frobenius norm,

$$\dot{E} = -G_\varepsilon^{\mathbb{R}}(E) + \langle G_\varepsilon^{\mathbb{R}}(E), E\rangle E. \tag{2.9}$$

ode-E-def

Monotonicity. Assuming simple eigenvalues along the trajectory, we again have the monotonicity property of Theorem II.1.4 and (II.2.7),

$$\frac{d}{dt}F_\varepsilon(E(t)) = -\|G_\varepsilon^{\mathbb{R}}(E) - \langle G_\varepsilon^{\mathbb{R}}(E), E\rangle E\|_F^2 \le 0. \tag{2.10}$$

eq:pos-def

Stationary points. Also the characterization of stationary points as given in Theorem II.1.5 extends with the same proof: Let $E \in \mathbb{R}^{n,n}$ with $\|E\|_F = 1$ be such that the eigenvalue $\lambda(A+\varepsilon E)$ is simple and $G_\varepsilon^{\mathbb{R}}(E) \neq 0$. Then,

$$\begin{gathered}E \text{ is a stationary point of the differential equation (3.6)} \quad \boxed{\texttt{stat-def}} \\ \text{if and only if } E \text{ is a real multiple of } G_\varepsilon^{\mathbb{R}}(E).\end{gathered} \tag{2.11}$$

stat-def

Hence, in this situation the rank of an optimizer of (2.3) is at most 4.

This raises the question as to whether the real gradient $G_\varepsilon^{\mathbb{R}}(E) = \mathrm{Re}\,G_\varepsilon(E) = \mathrm{Re}(xx^*Z^* + Z^*yy^*)$ can be the zero matrix. We first note that if $x^*y = 1$, then $G_\varepsilon(E) = 0$, because then $x = y$ and we have $Zy = 0$ and $x^*Z = 0$. Hence, for a normal matrix $A + \varepsilon E$ the gradient is always zero.

The following result excludes a vanishing real gradient if $x^*y < 1$ and if the eigenvalue λ is not purely real.

th:ReS

Theorem 2.2 (Non-vanishing real gradient). *Assume that the real matrix B has a pair of simple complex conjugate eigenvalues λ and $\overline{\lambda}$. Let x and y be left and right eigenvectors of unit norm with positive inner product associated with λ and assume that $x^*y < 1$. Let Z be the group inverse of $B - \lambda I$ and let $G = xx^*Z^* + Z^*yy^*$ as in* (2.6)*. Then,* $\operatorname{Re} G \neq 0$.

Proof. The proof is done by leading the assumption $\operatorname{Re} G = 0$ to a contradiction. So let us assume that G is purely imaginary. In part (a) of the proof we show that then $\{x, \overline{x}\}$ and $\{y, \overline{y}\}$ span the same 2-dimensional invariant subspace. In part (b) we work on this subspace and derive a contradiction.

(a) By definition of the matrix G, its range is given by

$$\operatorname{Ran}(G) = \operatorname{span}\{x, Z^*y\}\,.$$

If G is purely imaginary, then $\operatorname{Ran}(\overline{G}) = \operatorname{Ran}(G)$ and hence $\overline{x} \in \operatorname{Ran}(G)$, i.e., we have $\overline{x} = \alpha x + \beta Z^*y$. A left premultiplication with y^* allows us to conclude $\alpha = 0$, because (i) $y^*\overline{x} = 0$ by the bi-orthogonality of left and right eigenvectors corresponding to different eigenvalues (here λ and $\overline{\lambda}$), (ii) $Zy = 0$, a property of the group inverse Z, and (iii) $x^*y \neq 0$ as λ is a simple eigenvalue. This implies $\overline{x} \propto Z^*y$. Analogously, we obtain $\overline{y} \propto Zx$. So we have

$$Zx = \gamma\overline{y}, \qquad Z^*y = \eta\overline{x}, \tag{2.12}$$

eq:d1

with $\gamma \neq 0$ and $\eta \neq 0$.

Since y is in the null-space of $B - \lambda I$ and B is real, it follows that $\overline{y}$ is a right eigenvector of $B - \lambda I$ to the eigenvalue $\mu = -2\mathrm{i}\operatorname{Im}\lambda \neq 0$. For the group inverse Z of $B - \lambda I$, this implies that $\overline{y}$ is a right eigenvector of Z to the eigenvalue $\nu = 1/\mu$. Analogously we find that $\overline{x}$ is a left eigenvector of Z to the eigenvalue ν. So we have

$$Z\overline{y} = \nu\overline{y}, \qquad Z^*\overline{x} = \overline{\nu}\,\overline{x}, \tag{2.13}$$

eq:d2

with $\nu \neq 0$. Equations (2.12)–(2.13) show that $Z(\gamma^{-1}x - \nu^{-1}\overline{y}) = 0$, and since y spans the null-space of Z, we find $y \propto \gamma^{-1}x - \nu^{-1}\overline{y}$. Analogously we obtain $x \propto \eta^{-1}y - \overline{\nu}^{-1}\overline{x}$. We conclude that

$$\operatorname{span}\{y, \overline{y}\} = \operatorname{span}\{x, \overline{x}\}\ . \tag{2.14}$$

span-x-y

(b) The two-dimensional space $Y := \operatorname{span}\{y, \overline{y}\}$ is thus an invariant subspace of both B and $B^\top$. We choose the real orthonormal basis (q_1, q_2) of Y that is obtained by normalizing the orthogonal vectors $y + \overline{y}$ and $i(y - \overline{y})$. We extend this basis of Y to a real orthonormal basis $Q = (q_1, \ldots, q_n)$ of $\mathbb{C}^n$. From (2.14) we infer the block-diagonal structure

$$\widetilde{B} = Q^\top BQ = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}.$$

The 2×2 matrix

$$B_1 = \begin{pmatrix} \varrho & \sigma \\ -\,\tau & \varrho \end{pmatrix}$$

is such that $\varrho = \mathrm{Re}(\lambda)$ and $\sigma > 0, \tau > 0$ with $\sigma\tau = \mathrm{Im}(\lambda)^2 > 0$ so that B_1 has eigenvalues λ and $\overline{\lambda}$.

If $\sigma = \tau$ then B_1 is normal, which implies that the pair of right and left eigenvectors associated with λ, say $\widetilde{y}, \widetilde{x}$ (scaled to have unit norm and positive inner product) is such that $\widetilde{x}^*\widetilde{y} = 1$. Since $y = Q\widetilde{y}$ and $x = Q\widetilde{x}$, the orthogonality of Q implies $x^*y = 1$, which contradicts the assumption $x^*y < 1$. So we must have $\sigma \neq \tau$.

By the properties of the group inverse we have that

$$\widetilde{Z} = Q^\top Z Q = \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix},$$

where Z_1 is the group inverse of $B_1 - \lambda I$ and Z_2 is the inverse of the nonsingular matrix $B_2 - \lambda I$. The following formula for the group inverse is verified by simply checking the three conditions in Definition VIII.1.3:

$$Z_1 = \begin{pmatrix} \frac{\mathrm{i}}{4\sqrt{\sigma\tau}} & -\frac{1}{4\tau} \\ \frac{1}{4\sigma} & \frac{\mathrm{i}}{4\sqrt{\sigma\tau}} \end{pmatrix}.$$

It follows that also $Q^\top G Q$ is block diagonal so that we write

$$\widetilde{G} = Q^\top G Q = \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix} \qquad \text{with} \qquad G_1 = \widetilde{x}_1\widetilde{x}_1^* Z_1^* + Z_1^*\widetilde{y}_1\widetilde{y}_1^*,$$

where $\widetilde{x}_1 \in \mathbb{C}^2$ and $\widetilde{y}_1 \in \mathbb{C}^2$ are the projections onto $\mathrm{span}(e_1, e_2)$ (the subspace spanned by the first two vectors of the canonical basis) of the eigenvectors of $\widetilde{B}$ associated with λ, that is $\widetilde{y} = Q^\top y$ and $\widetilde{x} = Q^\top x$,

$$\begin{aligned} \widetilde{y} &= \nu_y^{-1} \begin{pmatrix} \mathrm{i}\frac{\sqrt{\sigma}}{\sqrt{\tau}} & 1 & 0 & \dots & 0 \end{pmatrix}^\top \\ \widetilde{x} &= \nu_x^{-1} \begin{pmatrix} -\mathrm{i}\frac{\sqrt{\tau}}{\sqrt{\sigma}} & 1 & 0 & \dots & 0 \end{pmatrix}^\top \end{aligned}$$

with $\nu_y = \sqrt{\frac{\sigma}{\tau} + 1}$ and $\nu_x = \sqrt{\frac{\tau}{\sigma} + 1}$ chosen such that $\widetilde{x}$ and $\widetilde{y}$ are of unit norm with positive inner product. Finally we obtain

$$G_1 = \begin{pmatrix} 0 & \frac{\tau - \sigma}{2\sigma(\sigma+\tau)} \\ \frac{\tau - \sigma}{2\tau(\sigma+\tau)} & 0 \end{pmatrix},$$

which is real and cannot vanish due to the fact that $\sigma \neq \tau$.

Recalling that Q is real, if G were purely imaginary then G_1 would be purely imaginary as well, which would give a contradiction. □

V.2.3 Rank-4 constrained gradient flow

With the real gradient of rank $r = 4$, the rank-r constrained gradient flow and its discretization as described in Sections II.2.4 and II.2.5 can be used directly for the present minimization problem. In the same way as in Theorem II.2.4 we find that the stationary points are the same as for the gradient flow (2.9) provided that $P_E G_\varepsilon^{\mathbb{R}}(E) \neq 0$, which holds true if $G_\varepsilon^{\mathbb{R}}(E) \neq 0$ and λ is not an eigenvalue of A.

V.2.4 Outer iteration, update of ε

We proceed as in Section IV.2 and make an analogous assumption:

ass:E-eps-def **Assumption 2.3.** For ε close to $\varepsilon_\star$ and $\varepsilon < \varepsilon_\star$, we assume the following for the optimizer $E(\varepsilon)$ of (2.3):

- The eigenvalue $\lambda(\varepsilon) = \lambda(A + \varepsilon E(\varepsilon))$ is a simple eigenvalue.
- The map $\varepsilon \mapsto E(\varepsilon)$ is continuously differentiable.
- The real gradient $G^{\mathbb{R}}(\varepsilon) = G^{\mathbb{R}}_\varepsilon(E(\varepsilon))$ is nonzero.

In the same way as in Theorem IV.2.2, we calculate under this assumption the derivative of $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon)) = x(\varepsilon)^* y(\varepsilon)$, where $x(\varepsilon)$ and $y(\varepsilon)$ are left and right eigenvectors of $A + \varepsilon E(\varepsilon)$ associated with $\lambda(\varepsilon)$, of unit norm and with positive inner product. Here we obtain (with $' = d/d\varepsilon$)

$$\phi'(\varepsilon) = -\phi(\varepsilon)\, \|G^{\mathbb{R}}(\varepsilon)\|_F < 0. \qquad (2.15)$$ eq:dereps-def

Starting from $\varepsilon > 0$ such that $\phi(\varepsilon) > \delta$, we want to compute the smallest root $\varepsilon_\delta > 0$ of the equation $\phi(\varepsilon) = \delta$. It is of interest to study the behaviour of $\phi(\varepsilon)$ as ε approaches $\varepsilon_\star = \lim_{\delta \searrow 0} \varepsilon_\delta$, where eigenvalues coalesce to form a Jordan block. We make the following generic assumption.

npt-epsstar-def **Assumption 2.4.** We assume the following in the limit $\varepsilon \nearrow \varepsilon_\star$:

- The eigenvalue $\lambda(\varepsilon)$ coalesces with only one other eigenvalue as $\varepsilon \nearrow \varepsilon_\star$ to form a Jordan block.
- The limits $x_\star = \lim_{\varepsilon \nearrow \varepsilon_\star} x(\varepsilon)$, $y_\star = \lim_{\varepsilon \nearrow \varepsilon_\star} y(\varepsilon)$, and $E_\star = \lim_{\varepsilon \nearrow \varepsilon_\star} E(\varepsilon)$ exist.

We note that if the limit matrix $E_\star$ exists and the matrix $A + \varepsilon_\star E_\star$ is non-derogatory, i.e., for each distinct eigenvalue there is only one Jordan block, then the existence of the limits $x_\star$ and $y_\star$ of left and right eigenvectors is ensured by a theorem of Conway & Halmos (1980). On the other hand, if the limits $x_\star$ and $y_\star$ of left and right eigenvectors exist, then also the limit matrix $E_\star$ exists by (2.11) and (2.7).

thm:sqrt-def **Theorem 2.5 (Square root asymptotics).** *Under Assumptions 2.3 and 2.4 and the non-degeneracy condition that $\gamma \ge 0$ defined in (2.16) below is nonzero, we have*

$$\phi(\varepsilon) = \gamma \sqrt{\varepsilon_\star - \varepsilon}\, (1 + o(1)) \quad \textit{as } \varepsilon \nearrow \varepsilon_\star.$$

Proof. The result follows if we can show that $\phi(\varepsilon)\phi'(\varepsilon)$ has a finite nonzero limit as $\varepsilon \nearrow \varepsilon_\star$. By (2.15) we have

$$\phi(\varepsilon)\phi'(\varepsilon) = -\phi(\varepsilon)^2\, \|\mathrm{Re}\, G(\varepsilon)\|_F.$$

We recall from (2.6) that

$$G(\varepsilon) = x(\varepsilon)x(\varepsilon)^* Z(\varepsilon)^* + Z(\varepsilon)^* y(\varepsilon) y(\varepsilon)^*,$$

where $Z(\varepsilon)$ is the group inverse of $N(\varepsilon) := A + \varepsilon E(\varepsilon) - \lambda(\varepsilon)I$. By Assumption 2.4, the rank of the matrix $N(\varepsilon)$ remains equal to $n-1$ also as $\varepsilon \nearrow \varepsilon_\star$. Therefore, the Moore-Penrose pseudoinverse

$$B(\varepsilon) := N(\varepsilon)^\dagger$$

has a finite limit $B_\star$ as $\varepsilon \nearrow \varepsilon_\star$, and by the second part of Assumption 2.4 also

$$\beta(\varepsilon) := x(\varepsilon)^* B(\varepsilon) y(\varepsilon)$$

has a finite limit $\beta_\star$. By Theorem VIII.1.4, we have for $\varepsilon \nearrow \varepsilon_\star$

$$\begin{aligned}\phi(\varepsilon)^2 Z(\varepsilon) &= (\phi(\varepsilon)I - y(\varepsilon)x(\varepsilon)^*)B(\varepsilon)(\phi(\varepsilon)I - y(\varepsilon)x(\varepsilon)^*)\\ &\to y_\star x_\star^* B_\star y_\star x_\star^* = \beta_\star y_\star x_\star^*\end{aligned}$$

and therefore by (2.6),

$$\begin{aligned}\phi(\varepsilon)^2\,\mathrm{Re}\,G(\varepsilon) &= \mathrm{Re}\Big(x(\varepsilon)x(\varepsilon)^*\phi(\varepsilon)^2\,Z(\varepsilon)^* + \phi(\varepsilon)^2\,Z(\varepsilon)^* y(\varepsilon)y(\varepsilon)^*\Big)\\ &\to \mathrm{Re}\big(2\overline{\beta}_\star x_\star y_\star^*\big),\end{aligned}$$

so that finally using (2.15),

$$\phi(\varepsilon)\phi'(\varepsilon) = -\phi(\varepsilon)^2 \|\mathrm{Re}\,G(\varepsilon)\|_F \to -\tfrac{1}{2}\gamma^2 := -2\|\mathrm{Re}\big(\overline{\beta}_\star x_\star y_\star^*\big)\|_F. \tag{2.16}$$

gamma-def

The stated result then follows in the same way as in part (d) of the proof of Theorem 1.7 provided that $\gamma \neq 0$ as is assumed. □

In view of the expected square root behaviour of Theorem 2.5, we use an outer iteration based on a square root model and bisection as described in Section V.1.4. If δ is not too small, a classical Newton iteration might also be used.

V.3 Nearest singular matrix pencil

matrix-pencils

Let A and B be complex $n \times n$ matrices. The matrix pencil $\{A - \mu B \;:\; \mu \in \mathbb{C}\}$, or equivalently the pair (A, B), is called *singular* if

$$A - \mu B \;\text{ is singular for all } \mu \in \mathbb{C}. \tag{3.1}$$

mp-sing

This notion is fundamental in the theory of linear differential-algebraic equations $B\dot{y}(t) = Ay(t) + f(t)$ (which might arise as linearizations of nonlinear differential-algebraic equations $F(y, \dot{y}) = 0$ near a stationary point). If the matrix pencil (A, B) is singular, then there exists no initial value $y(0)$ such that the corresponding initial value problem has a unique solution.

A necessary condition for (A, B) to be singular is clearly that both A and B are singular matrices (i.e. non-invertible). A sufficient condition for (A, B) to be singular is that

A and B have a common nonzero vector in their null-spaces. While this is a special case of particular interest, the Kronecker normal form of a matrix pencil (see Gantmacher **?**) shows that a common null-vector is not a necessary condition for a matrix pencil to be singular.

Given a matrix pencil (A, B) that is not singular, it is of interest to know how far it is from a singular matrix pencil. In this section we fix B, which is assumed to be a singular matrix, and consider structured perturbations $\Delta A \in \mathcal{S}$ to A, where the structure space $\mathcal{S}$ is an arbitrary complex- or real-linear subspace of $\mathbb{C}^{n,n}$, as considered in Section II.3. For example, $\mathcal{S}$ might be a space of real or complex matrices with a given sparsity pattern. We consider the following two structured matrix nearness problems:

- **Problem 1.** *Find $\Delta A \in \mathcal{S}$ of minimal Frobenius norm such that $(A + \Delta A, B)$ is a singular matrix pencil.*
- **Problem 2.** *Find $\Delta A \in \mathcal{S}$ of minimal Frobenius norm such that $A + \Delta A$ and B have a common nonzero vector in their null-spaces.*

If A itself is also in $\mathcal{S}$, then both problems have a solution, because the trivial choice $\Delta A = -A$ gives us the singular matrix pencil $(0, B)$ and because the set of perturbations ΔA yielding singular matrix pencils $(A+\Delta A, B)$ is closed. We will approach both problems by a two-level method in the spirit of Chapter IV. We note that this approach would equally allow us to treat analogous matrix nearness problems where also B is perturbed (see the notes at the end of this chapter), but for ease of presentation we have chosen not to do so here. Moreover, in the applications of interest to dynamical systems on networks, B is typically an adjacency matrix or related fixed matrix depending only on the network topology and hence is not subject to perturbations.

V.3.1 Distance to structured singular matrix pencils

Given $n + 1$ distinct complex numbers $\mu_0, \dots, \mu_n$, the fundamental theorem of algebra together with the fact that a matrix is singular if and only if its determinant vanishes, shows that a matrix pencil (A, B) is singular if and only if

$$\text{the } n+1 \text{ matrices } \; A - \mu_k B \; \text{ are singular for } k = 0, \dots, n. \tag{3.2}$$

mp-sing-d

Our numerical approach for Problem 1 is based on this criterion. The choice of the numbers $\mu_0, \dots, \mu_n$ did not appear to be a critical issue in our numerical experiments. We had good experience with the choice $\mu_k = re^{2\pi k \mathrm{i}/(n+1)}$ of modulus $r = \|A\|_F / \|B\|_F$.

For $\varepsilon > 0$ we introduce the functional F_ε (of matrices $E \in \mathcal{S}$ of unit Frobenius norm) in the following way:

$$F_\varepsilon(E) = \frac{1}{2} \sum_{k=0}^{n} |\lambda(A + \varepsilon E - \mu_k B)|^2, \tag{3.3}$$

F-eps-mp

where $\lambda(M)$ is an eigenvalue of smallest modulus of the matrix M. (Alternatively, we might take the sum of the smallest singular values of these $n + 1$ matrices.)

We follow the two-level approach of Section IV.2:

- **Inner iteration:** Given $\varepsilon > 0$, we aim to compute a matrix $E(\varepsilon) \in \mathcal{S}$ of unit Frobenius norm that minimizes F_ε:

$$E(\varepsilon) = \arg \min_{E \in \mathcal{S}, \|E\|_F = 1} F_\varepsilon(E). \tag{3.4}$$

E-eps-mp

- **Outer iteration:** We compute the smallest positive value $\varepsilon_\star$ with

$$\phi(\varepsilon_\star) = 0, \tag{3.5}$$

zero-mp

where $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon))$.

Provided that these computations succeed, we then have that $\varDelta A_\star = \varepsilon_\star E(\varepsilon_\star) \in \mathcal{S}$ is a solution to Problem 1 above, and $\varepsilon_\star$ is the distance of the matrix pencil (A, B) to the set of structured singular matrix pencils of the form $(A + \varDelta A, B)$ with $\varDelta A \in \mathcal{S}$.

V.3.2 Constrained gradient flow for the inner iteration

radient-flow-mp

The programme of Section II.3.3 extends to the current situation as follows.

Structured gradient. We consider a path of matrices $E(t) \in \mathcal{S}$ and we assume that the eigenvalues $\lambda_k(t) = \lambda(A + \varepsilon E(t) - \mu_k B)$ are simple eigenvalues. As in Section II.3.3 and Lemma II.1.1, we find that

$$\frac{1}{\varepsilon}\,\frac{d}{dt} F_\varepsilon(E(t)) = \mathrm{Re}\big\langle G_\varepsilon^{\mathcal{S}}(E(t)), \dot{E}(t) \big\rangle \tag{3.6}$$

eq:deriv-S-mp

with the rescaled structured gradient

$$G_\varepsilon^{\mathcal{S}}(E) = \varPi^{\mathcal{S}} G_\varepsilon(E) \quad \text{with} \quad G_\varepsilon(E) = \sum_{k=0}^{n} \alpha_k\, x_k y_k^*, \tag{3.7}$$

gradient-S-mp

where x_k and y_k are left and right eigenvectors corresponding to λ_k, chosen of unit norm and with positive inner product, and $\alpha_k = \lambda_k / (x_k^* y_k)$. Moreover, $\varPi^{\mathcal{S}}$ is again the orthogonal projection onto the structure space $\mathcal{S}$; see Section II.3.2.

Structure- and norm-constrained gradient flow. We consider the gradient flow on the manifold of matrices in $\mathcal{S}$ of unit Frobenius norm,

$$\dot{E} = -G_\varepsilon^{\mathcal{S}}(E) + \mathrm{Re}\langle G_\varepsilon^{\mathcal{S}}(E), E \rangle E. \tag{3.8}$$

ode-E-S-mp

Monotonicity. Assuming simple eigenvalues along the trajectory, we again obtain the monotonicity property of Theorem II.1.4,

$$\frac{d}{dt} F_\varepsilon(E(t)) \le 0. \tag{3.9}$$

eq:pos-S-mp

Stationary points. Also the characterization of stationary points as given in Theorem II.1.5 extends with the same proof: Let $E \in \mathcal{S}$ with $\|E\|_F = 1$ be such that the eigenvalues $\lambda(A + \varepsilon E - \mu_k B)$ are simple for $k = 0, \dots, n$ and $G_\varepsilon^{\mathcal{S}}(E) \neq 0$. Then,

$$\begin{array}{l} E \text{ is a stationary point of the differential equation (3.6)} \\ \text{if and only if } E \text{ is a real multiple of } G_\varepsilon^{\mathcal{S}}(E). \end{array} \tag{3.10}$$

stat-S-mp stat-S-mp

However, in contrast to Section II.3.3, this now does not imply that non-degenerate optimizers are projections of rank-1 matrices onto the structure space $\mathcal{S}$, because $G_\varepsilon^{\mathcal{S}}(E)$ is no longer a projected rank-1 matrix. So we cannot work with rank-1 matrices here. In the inner iteration we therefore follow the structure- and norm-constrained gradient flow (3.8) into a stationary point.

V.3.3 Outer iteration, updating ε

We proceed as in Section IV.2 and make an analogous assumption: For ε close to $\varepsilon_\star$ and $\varepsilon < \varepsilon_\star$, we assume the following for the optimizer $E(\varepsilon)$ of (3.4):

- The eigenvalues $\lambda_k(\varepsilon) = \lambda(A + \varepsilon E(\varepsilon) - \mu_k B)$ for $k = 0, \dots, n$ are simple eigenvalues.
- The map $\varepsilon \mapsto E(\varepsilon)$ is continuously differentiable.
- The structured gradient $G^{\mathcal{S}}(\varepsilon) = G_\varepsilon^{\mathcal{S}}(E(\varepsilon))$ is nonzero.

In the same way as in Theorem IV.2.2, we obtain under this assumption a simple expression for the derivative of $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon))$, which here becomes (with $' = d/d\varepsilon$)

$$\phi'(\varepsilon) = -\|G^{\mathcal{S}}(\varepsilon)\|_F < 0. \tag{3.11}$$

eq:dereps-mp

This expression can be used in a Newton / bisection method. Since the eigenvalues $\lambda_k(\varepsilon)$ that define $\phi(\varepsilon) = F_\varepsilon(E(\varepsilon))$ (see (3.3)) are assumed to be simple, $\phi(\varepsilon)$ can be expected to behave asymptotically for $\varepsilon \nearrow \varepsilon_\star$ as

$$\phi(\varepsilon) \approx c\,(\varepsilon_\star - \varepsilon)^2.$$

The unknown quantities c and $\varepsilon_\star$ can be estimated using $\phi(\varepsilon)$ and $\phi'(\varepsilon)$, which is known from (3.11) for $\varepsilon < \varepsilon_\star$. This gives

$$\varepsilon_\star \approx \varepsilon - 2\frac{\phi(\varepsilon)}{\phi'(\varepsilon)}, \qquad c \approx \frac{\phi'(\varepsilon)^2}{4\phi(\varepsilon)}.$$

For $\varepsilon = \varepsilon_k < \varepsilon_\star$, we thus obtain the Newton-type iteration

$$\varepsilon_{k+1} = \varepsilon_k - 2\frac{\phi(\varepsilon_k)}{\phi'(\varepsilon_k)}, \tag{3.12}$$

eq:Newton

which yields a locally quadratically convergent iteration from the left (if instead $\varepsilon_k > \varepsilon_\star$ occurs, then we should use bisection, which would give a linear reduction of the error from the right).

V.3.4 Distance to structured singular matrix pencils with common null-vectors

We aim to find a perturbation $\Delta A \in \mathcal{S}$ of minimal Frobenius norm such that there exists a nonzero vector v that is in the null-spaces of both $A + \Delta A$ and B. To this end we introduce, for $\varepsilon > 0$, a functional F_ε (of matrices $E \in \mathcal{S}$ of unit Frobenius norm) as follows: Let $\lambda = \lambda(A + \varepsilon E)$ be an eigenvalue of minimal modulus of $A + \varepsilon E$ and let y be a corresponding (right) eigenvector of unit norm. With $\beta = 1/\|B\|_2^2$, we set

$$F_\varepsilon(E) = \frac{1}{2}|\lambda|^2 + \frac{\beta}{2}\|By\|_2^2. \tag{3.13}$$

F-eps-mp-cnv

With this functional, which now depends on both an eigenvalue and an eigenvector, we again follow the two-level approach (3.4)–(3.5). Provided that these computations succeed, we then have that $\Delta A_\star = \varepsilon_\star E(\varepsilon_\star) \in \mathcal{S}$ is a solution to Problem 2 above, and $\varepsilon_\star$ is the distance of the matrix pencil (A, B) to the set of matrix pencils of the form $(A + \Delta A, B)$ with $\Delta A \in \mathcal{S}$ for which $A + \Delta A$ and B have a common nonzero null-vector.

For the inner iteration, the programme of Section V.3.2 carries over to the present situation. The only additional difficulty is in calculating the gradient, for which we proceed as in the proof of Lemma II.1.1 and now use the formulas for derivatives of both eigenvalues and eigenvectors as given in the appendix. This straightforward though slightly lengthy calculation yields the following result, which provides us once again with a gradient of rank 1.

gradient-mp-cnv **Lemma 3.1 (Free gradient).** *Let $E(t) \in \mathbb{C}^{n,n}$, for real t near t_0, be a continuously differentiable path of matrices, with the derivative denoted by $\dot E(t)$. Assume that $\lambda(t)$ is a simple eigenvalue of $A + \varepsilon E(t)$ depending continuously on t. Then, $F_\varepsilon(E(t))$ of* (3.13) *is continuously differentiable w.r.t. t and we have*

$$\frac{1}{\varepsilon}\frac{d}{dt}F_\varepsilon(E(t)) = \mathrm{Re}\,\big\langle G_\varepsilon(E(t)), \dot E(t)\big\rangle, \tag{3.14}$$

eq:deriv

where the rescaled gradient of F_ε is the rank-1 matrix

$$G_\varepsilon(E) = uy^* \in \mathbb{C}^{n,n} \quad \textit{with } u = \kappa\lambda x - \beta Z^*\big(B^*B - \|By\|_2^2\, I_n\big)y, \tag{3.15}$$

eq:freegrad

with the eigenvalue $\lambda = \lambda(A + \varepsilon E)$ and the corresponding left and right normalized eigenvectors x and y with positive inner product, with $\kappa = 1/(x^*y) > 0$ and with the group inverse Z of $A + \varepsilon E - \lambda I$.

With this rank-1 gradient (and its projection $G_\varepsilon^{\mathcal{S}}(E) = \Pi^{\mathcal{S}} G_\varepsilon(E)$ onto the structure space $\mathcal{S}$), the full programme of Section II.3 carries over to the present situation, including the rank-1 differential equation (II.3.10) and its discretization, which were not available for the (high-rank) gradient of Section V.3.2.

For the outer iteration, we again use a Newton / bisection method as in Section IV.2. Here we have again $\phi'(\varepsilon) = -\|G_\varepsilon^{\mathcal{S}}(E(\varepsilon))\|_F$. Note that the factor $\kappa(\varepsilon)$ does not appear in this formula because it is already included in $G_\varepsilon(E)$.

V.4 Stability radii for delay differential equations

sec:delay

V.4.1 A simple example

Consider scalar DDEs

$$\dot{x}(t) = ax(t) + bx(t-1) \tag{4.1}$$

eq:abdde

with $a, b \in \mathbb{R}$. Looking for solutions $x(t) = ce^{\lambda t}$ gives the characteristic equation

$$\lambda - a - b\,\mathrm{e}^{-\lambda} = 0. \tag{4.2}$$

eq:charab

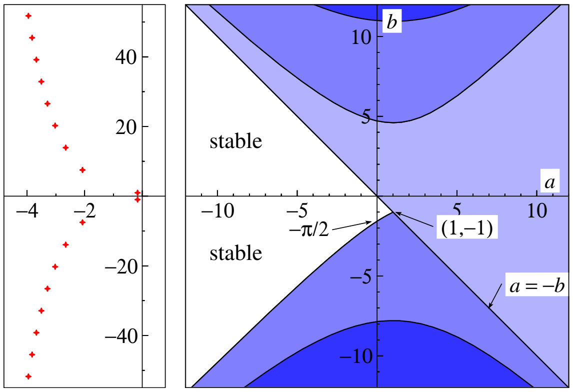

fig:stabr

Fig. 4.1. Region of the (a, b)-real plane such that the solution of (4.1) is asymptotically stable (white set in the right illustration); roots of the characteristic equation (left) for $a = 0.5, b = -1$

A main feature of this problem is that the entire function $\lambda - a - b\,\mathrm{e}^{-\lambda}$ (also called quasi-polynomial) has infinitely many roots and stability can be proved if and only if all the roots lie on the complex left-plane or - in other words - the rightmost root has negative real part. This is what happens when $a = 0.5, b = -1$ as shown in Figure V.4.1 (left illustration).

Clearly stability is more robust for problems s.t. (a, b) is far from the boundary.

The stability of a linear system of delay equations

$$\begin{array}{rcl} \dot{x}(t) & = & A_1 x(t) + A_2 x\left(t - \tau\right), \qquad t > 0 \\ x(t) & = & g(t), \qquad t \in [-\tau, 0] \end{array} \tag{4.3}$$

with A_1, A_2 given $n \times n$ matrices and $\tau > 0$ constant delay, can be analyzed in a similar way.

Inserting solutions of the form $x(t) = v\,\mathrm{e}^{\lambda t}$ (with $v \in \mathbb{C}^n$) we get :

$$\left(\lambda I - A_1 - A_2 \mathrm{e}^{-\lambda\tau}\right) v = 0 \quad \Longleftrightarrow \quad \det\left(\lambda I - A_1 - A_2 \mathrm{e}^{-\lambda\tau}\right) = 0, \tag{4.4}$$

eq:charsys

a nonlinear eigenvalue problem.

If the infinitely many eigenvalues strictly lie within the complex left half-plane every solution of (4.3) is asymptotically stable, independently of the initial data.

The main difficulty here is that in general the matrices A_1 and A_2 do not commute, so that cannot be transformed simultaneously to a simple (diagonal) form, in contrast to the case of linear ODEs.

Suitable algorithms for the numerical computation of characteristic roots of linear systems are available (e.g Engelborghs and Roose, 2002, Breda, Maset and Vermiglio, 2005, Jarlebring, Meerbergen and Michiels, 2010 (based on Krylov solvers)).

Understanding robust stability of these systems is an important task.

V.4.2 The nonlinear eigenvalue problem

We consider the following class of nonlinear eigenvalue problems,

$$\left(\sum_{i=0}^{m} A_i f_i(\lambda) \right) v = 0, \qquad \lambda \in \mathbb{C}, v \in \mathbb{C}^n, \tag{4.5}$$

eq:nonlin

where $A_0, \dots, A_m$ are given $n \times n$ matrices and the functions $f_0, \dots, f_m$ are assumed to be *entire*, such that

$$f_i(\overline{\lambda}) = \overline{f_i(\lambda)}, \; 0 \le i \le m.$$

As usual we denote the spectrum by $\mathit{\Lambda}$, i.e.

$$\mathit{\Lambda} := \left\{ \lambda \in \mathbb{C} : \; \det \left(\sum_{i=0}^{m} A_i f_i(\lambda) \right) = 0 \right\} \tag{4.6}$$

and are interested in the effect of bounded perturbations $\mathit{\Delta} A_i$ of A_i, i.e. in studying

$$\left(\sum_{i=0}^{m} (A_i + \mathit{\Delta} A_i) f_i(\lambda) \right) v = 0, \qquad \lambda \in \mathbb{C}, v \in \mathbb{C}^n. \tag{4.7}$$

pert-init

We let

$$\mathit{\Delta} := \begin{pmatrix} \mathit{\Delta} A_0 \\ \vdots \\ \mathit{\Delta} A_m \end{pmatrix}.$$

In analogy to the classical definition of ε-pseudospectrum of a matrix, we allow the perturbations to be *complex*.

Introducing weights $w_i > 0, i = 0, \dots, m$, we make use of the norm:

$$\| \mathit{\Delta} \| := \sqrt{\sum_{i=0}^{m} w_i^2 \| \mathit{\Delta} A_i \|_F^2}$$

In order to define the associated ε-pseudospectrum we consider bounded perturbations,

$$\|\Delta\| \leq \varepsilon$$

Note that taking $w_i = +\infty$ implies that the matrix A_i is unperturbed.

Definition 4.1 **(ε-pseudospectrum).** The complex set

$$\Lambda_\varepsilon = \bigcup_{\|\Delta\| \leq \varepsilon} \left\{ \lambda \in \mathbb{C} : \ \det\left(\sum_{i=0}^{m} (A_i + \Delta A_i) f_i(\lambda) \right) = 0 \right\}$$

is the set of eigenvalues associated to all possible perturbed problems.

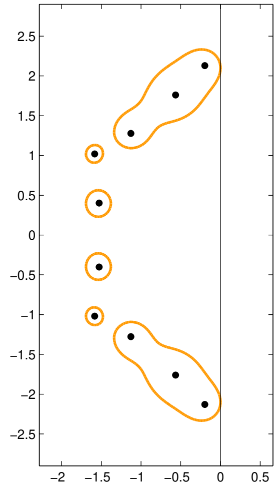

Fig. 4.2. Rightmost roots of an example of system (4.4) and ε-pseudospectrum tangential to the imaginary axis (showing $d = \varepsilon$).

In analogy the linear case the ε-pseudospectral abscissa is defined by

$$\alpha_\varepsilon := \sup \left\{ \mathrm{Re}(\lambda) : \ \lambda \in \Lambda_\varepsilon \right\}. \tag{4.8}$$

defpsa

where in this case - due to the infinitely many eigenvalues - the $\max$ is replaced by the sup.

Hurwitz stability is associated with the requirement that the spectrum be located in the open left half plane and bounded away from the imaginary axis.

Then the *distance to instability* (stability radius) of a stable system is expressed as

$$d := \inf \{\varepsilon > 0 : \; \alpha_\varepsilon \geq 0\} \, .$$

Our aim is to construct methods able to approximately compute α_ε and d.

Some important assumptions For nonlinear eigenvalue problems the pseudospectral abscissa may be equal to infinity (as in differential-algebraic equations), or a globally rightmost point of the pseudospectrum may not exist.

Assumption 4.2. We assume the following:

(i) For all $r \in \mathbb{R}$ the set $\Lambda_\varepsilon \cap \{\lambda \in \mathbb{C} : \; \mathrm{Re}(\lambda) \geq r\}$ is bounded.
(ii) For an arbitrary but fixed $\varepsilon > 0$, $\alpha_\varepsilon < +\infty$.

As an example which does not fulfil assumption (i) consider the following neutral equation

$$\dot{x}(t) = \dot{x}(t-\tau) - 2x(t) - x(t-\tau) \tag{4.9}$$

eq:neut

Its characteristic equation is given by

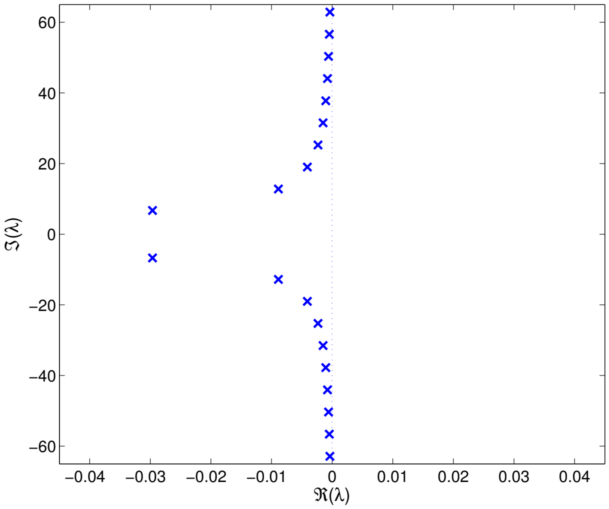

Fig. 4.3. Rightmost roots of the neutral equation.

$$\lambda \left(1 - \mathrm{e}^{-\lambda}\right) + 2 + \mathrm{e}^{-\lambda} = 0$$

which has a sequence of roots approaching the imaginary axis with increasing imaginary part.

Here the spectral abscissa is equal to zero, yet there is *no characteristic root with zero real part*.

When considering systems of the form $A_0\dot{x}(t) = A_1x(t) + A_2x\,(t-\tau)$, with A_0 singular (DDAEs), it is possible to have eigenvalues at $+\infty$, in analogy to the ODE case.

As an illustration of such a situation where $\alpha_0 = +\infty$ we consider the eigenvalue problem

$$\left(\begin{pmatrix} -1 & 2 \\ -1 & 1 \end{pmatrix} + \lambda \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \mathrm{e}^{-\lambda} \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}\right) v = 0. \qquad (4.10)$$

quadeig

The shaded area in Figure V.4.2 corresponds to $\varepsilon = 0.2$ and weights $(w_1, w_2) = (1, 1)$. The component not connected to the eigenvalue $\lambda = -1$ can be interpreted as a result of perturbations of the "eigenvalue at infinity".

fig:ddae

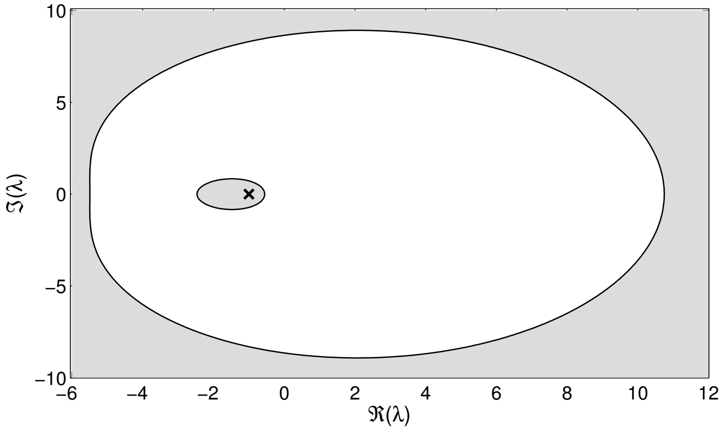

Fig. 4.4. ε-pseudospectrum of the eigenvalue problem (4.10).

V.4.3 General methodology for Hurwitz stability

We consider the special case

$$f(\lambda, \overline{\lambda}) = -\frac{\lambda + \overline{\lambda}}{2} = -\mathrm{Re}\lambda$$

where the target eigenvalue is the rightmost, and assume (4.5) fulfils assumptions (i) and (ii). Moreover we also assume that $w_i < \infty$ for $i = 0, \ldots, m$ (otherwise the matrices associated to infinite weights are not perturbed). As usual we introduce perturbations

$$\varDelta A_0 = \varepsilon E_0, \quad \varDelta A_1 = \varepsilon E_1, \quad \ldots \quad \varDelta A_m = \varepsilon E_m$$

with

$$\left\| \begin{pmatrix} E_0 \\ E_1 \\ \vdots \\ E_m \end{pmatrix} \right\|^2 = \left\langle \begin{pmatrix} w_0 E_0 \\ w_1 E_1 \\ \vdots \\ w_m E_m \end{pmatrix}, \begin{pmatrix} w_0 E_0 \\ w_1 E_1 \\ \vdots \\ w_m E_m \end{pmatrix} \right\rangle = 1$$

and indicate by $\mathcal{S}_1$ the unit ball of the norm $\| \cdot \|$; then we look for

$$\max_{(E_0 \ E_1 \ \ldots E_m)^\top \in \mathcal{S}_1} \Big\{ \mathrm{Re}\lambda : \ \det\big(f_0(\lambda)(A_0 + \varepsilon E_0) + \ldots + f_m(\lambda)(A_m + \varepsilon E_m) \big) = 0 \Big\}.$$

With the functional to minimize, $F_\varepsilon\left(E_0, E_1, \ldots, E_m\right) = -\mathrm{Re}\left(\lambda\right)$, λ being the rightmost (target) eigenvalue of the eigenvalue problem (4.16) we construct smooth matrix valued functions $\{E_i(t)\}_{i=0,1,\ldots,m}$ such that $F_\varepsilon\left(E_0(t), E_1(t), \ldots, E_m(t)\right)$ is decreasing, $\lambda(t)$ being the rightmost root of

$$\det\Big(f_0(\lambda)(A_0 + \varepsilon E_0(t)) + f_1(\lambda)(A_1 + \varepsilon E_1(t)) + \ldots + f_m(\lambda)(A_m + \varepsilon E_m(t)) \Big) = 0$$

In this way we obtain - by applying the standard first order perturbation result (VIII.1.1)

$$\begin{aligned} \frac{1}{\varepsilon \kappa(t)} \frac{d}{dt} F_\varepsilon\left(E_0(t), \ldots, E_m(t))\right) &= -x(t)^* \left(f_0(\lambda(t)) \dot{E}_1(t) + \ldots + f_m(\lambda(t)) \dot{E}_m(t) \right) y(t) \\ &= -\Big(\Big\langle \overline{f_0(\lambda(t))} x(t) y(t)^*, \dot{E}_0(t) \Big\rangle + \ldots \\ &\quad + \Big\langle \overline{f_m(\lambda(t))} x(t) y(t)^*, \dot{E}_m(t) \Big\rangle \Big) \end{aligned}$$

where $(\lambda(t), x(t), y(t))$ is an eigen-triplet of (4.5) with perturbed matrices $A_i + \varepsilon E_i(t)$ and $\kappa(t) = 1/(x(t)^* y(t))$.

This gives us the free gradient of the functional $F_\varepsilon(E_0, E_1, \ldots, E_m)$:

$$G_0 = \ - f_0(\overline{\lambda}) x y^*, \qquad G_1 = \ - f_1(\overline{\lambda}) x y^*, \qquad , G_m = \ - f_m(\overline{\lambda}) x y^*$$

which are the *directions* for $E_0, E_1, \ldots, E_m$ of maximal increase of F_ε.

Constrained gradient flow

Differentiating the norm constraint yields

$$\frac{d}{dt} \left\| \begin{pmatrix} E_0 \\ E_1 \\ \vdots \\ E_m \end{pmatrix} \right\|^2 = 2\mathrm{Re} \left\langle \begin{pmatrix} w_0 \dot{E}_0 \\ w_1 \dot{E}_1 \\ \vdots \\ w_m \dot{E}_m \end{pmatrix}, \begin{pmatrix} w_0 E_0 \\ w_1 E_1 \\ \vdots \\ w_m E_m \end{pmatrix} \right\rangle = 0 \qquad (4.11) \qquad \boxed{\texttt{eq:normcon}}$$

Lemma 4.3. *Let* $G = (G_0 \;\; G_1 \quad \ldots \; G_m)^\top$ *and* $Z = (Z_0 \;\; Z_1 \quad \ldots \; Z_m)^\top$. *A solution to the optimization problem*

$$
\begin{array}{rcl}
Z_\star & = & \arg \max\limits_{Z \in \mathbb{C}^{(m+1)n,n}} \operatorname{Re} \langle G, Z \rangle \\[2ex]
\text{subj.to} & & \operatorname{Re} \left\langle \begin{pmatrix} w_0 E_0 \\ w_1 E_1 \\ \vdots \\ w_m E_m \end{pmatrix}, \begin{pmatrix} w_0 Z_0 \\ w_1 Z_1 \\ \vdots \\ w_m Z_m \end{pmatrix} \right\rangle = 0 \\[2ex]
\text{and} & & \|Z\|_F = 1 \qquad \text{(for uniqueness)}
\end{array}
$$

is given by

$$
\mu Z_\star = \begin{pmatrix} G_0 - \eta E_0 \\ G_1 - \eta E_1 \\ \vdots \\ G_m - \eta E_m \end{pmatrix}
$$

with

$$
\eta = \operatorname{Re} \left\langle \begin{pmatrix} w_0 G_0 \\ w_1 G_1 \\ \vdots \\ w_m G_m \end{pmatrix}, \begin{pmatrix} w_0 E_0 \\ w_1 E_1 \\ \vdots \\ w_m E_m \end{pmatrix} \right\rangle \tag{4.12}
$$

eq:eta

and μ *a normalization factor.*

Proof. Analogous to the one of Lemma (II.1.3).

In order to preserve the norm the gradient has to be projected and an easy calculation allows us to write the constrained gradient flow

$$
\left\{
\begin{array}{rcl}
\dot{E}_0 & = & -G_0 + \eta E_0 \\
\dot{E}_1 & = & -G_1 + \eta E_1 \\
\vdots & & \\
\dot{E}_m & = & -G_m + \eta E_m
\end{array}
\right. \tag{4.13}
$$

ode-E

with η given by (4.12).

Theorem 4.4 (Stationary points). *Let* $E_0(t), E_1(t), \ldots, E_m(t)$ *be the solution of* (4.13) *passing through* $(\widehat{E}_0, \widehat{E}_1, \ldots, \widehat{E}_m)$ *with* λ *simple rightmost root of the characteristic equation. Then the following are equivalent:*

1. $\dfrac{d}{dt} F_\varepsilon \left(\widehat{E}_0, \widehat{E}_1, \ldots, \widehat{E}_m \right) = 0$
2. $\left(\widehat{E}_0, \widehat{E}_1, \ldots, \widehat{E}_m \right)^\top$ is a stationary point of the differential equation (4.13). stat-S
3. $\widehat{E}_i$ is a real multiple of G_i, $i = 0, 1, \ldots, m$.

Since $G_i = -f_i(\overline{\lambda})xy^*$ all extremizers $\widehat{E}_i$ have rank-1, which motivates the search of rank-1 ODEs with the same qualitative properties of (4.13).

Rank-1 ODEs

As we have seen in Chapter II the idea is that to project the ODE to the rank-1 manifold $\mathcal{M}_1$. In full analogy we obtain the projected system in $\mathcal{M}_1$,

$$\left\{\begin{array}{rcl} \dot{Y}_1 & = & P_{Y_0}G_0 - \eta Y_0 \\ \dot{Y}_2 & = & P_{Y_1}G_1 - \eta Y_1 \\ \vdots & & \\ \dot{Y}_m & = & P_{Y_m}G_m - \eta Y_m \end{array}\right. \tag{4.14}$$

ode-E-1-delay

with $Y_i = \sigma_i u_i v_i^*$ of rank-1 with $\|u_i\| = \|v_i\| = |\sigma_i = 1$ for all i, P_Y the ortogonal projection onto the tangent plane at $Y \in \mathcal{M}_1$, and

$$\eta = \mathrm{Re} \left\langle \begin{pmatrix} w_0 P_{Y_0} G_0 \\ w_1 P_{Y_1} G_1 \\ \vdots \\ w_m P_{Y_m} G_m \end{pmatrix}, \begin{pmatrix} w_0 Y_0 \\ w_1 Y_1 \\ \vdots \\ w_m Y_m \end{pmatrix} \right\rangle$$

the coupling quantity.

Similarly to the prototype case considered in Chapter II, it can be shown that the projected ODE preserves the norm, the monotonicity and the stationary points of the unprojected one, which again determines a significant reduction in CPU time and storage.

V.4.4 Stability radii

For the computation of stability radii we still make use of the two-step methodology:

(i) **Inner iteration:** Given $\varepsilon > 0$ compute $E_0 1(\varepsilon), E_1(\varepsilon), \ldots, E_m(\varepsilon)$ (stationary points of (4.13), similarly of (1.23)), the associated free gradient $G_0(\varepsilon), G_1(\varepsilon), \ldots, G_m(\varepsilon)$ and let

$$\varphi(\varepsilon) = F_\varepsilon\left(E_1(\varepsilon), E_2(\varepsilon), \ldots, E_m(\varepsilon)\right).$$

(ii) **Outer iteration:** We compute the smallest positive value $\varepsilon_\star$ with

$$\varphi(\varepsilon_\star) = 0,$$

where we use the costless formula

$$\frac{d}{d\varepsilon}\varphi(\varepsilon) = -\kappa(\varepsilon) \left\| \begin{pmatrix} w_0 G_0(\varepsilon) \\ w_1 G_1(\varepsilon) \\ \vdots \\ w_m G_m(\varepsilon) \end{pmatrix} \right\|$$

which can be used by a Newton-based iterative method.

V.4.5 Illustrative example: a linear system of DDEs

Consider the linear system of delay equations

$$\begin{array}{rcl} A_0\dot{x}(t) & = & A_1 x(t) + A_2 x\,(t-1)\,, \qquad t>0 \\ x(t) & = & g(t), \qquad t \in [-\tau, 0] \end{array} \tag{4.15}$$

with A_0, A_1, A_2 given $n \times n$ matrices and $\tau = 1$ constant delay.

It is possible to prove that if A_0 is nonsingular (4.15) fulfils assumptions (i) and (ii).

The associated nonlinear eigenvalue problem is given by

$$\det\left(\lambda A_0 - A_1 - A_2 \mathrm{e}^{-\lambda}\right) = 0, \tag{4.16}$$ eq:nlep

so that - in the considered setting -

$$f_0(\lambda) = \lambda, \qquad f_1(\lambda) = -1, \qquad f_2(\lambda) = -\mathrm{e}^{-\lambda}$$

We set $A_0 = I$ and $w_0 = \infty$ implying this way the identity matrix A_0 is not perturbed. This gives us a system of the form (4.13) (equivalently (1.23)) with matrices E_1 and E_2 (Y_1 and Y_2). Note that E_0 (Y_0) is missing because $w_0 = +\infty$..

α	(w_1, w_2)	ε	α_ε	#steps
-3.312133337e-1	$(1/2, 1/2)$	1.e-3	-3.297978515e-1	2
		1.e-2	-3.170982221e-1	3
		1.e-1	-1.937166436e-1	4
		1.e0	8.647127140e-1	8
	(1/4,∞)	1.e-3	-3.300292687e-1	2
		1.e-2	-3.193270916e-1	3
		1.e-1	-2.075848221e-1	4
		1+00	+1.641134830e00	10
	(∞,1/4)	1.e-3	-3.295667765e-1	2
		1.e-2	-3.149018637e-1	3
		1.e-1	-1.816328080e-1	4
		1+00	+5.599912563e-1	7

Table 4.1. Pseudospectral abscissa $\alpha_\varepsilon = -\varphi(\varepsilon)$ for the delayed PDE problem described in **?**. tabelpde

Fast convergence of Newton method can be observed in Figure V.4.5.

A PDE with a delay

Consider the problem (Jarlebring et al., 2010)

$$\frac{\partial v(x,t)}{\partial t} = \frac{\partial^2 v(x,t)}{\partial x^2} + a_0(x)v(x,t) + a_1(x)v(\pi - x, t-1), \tag{4.17}$$ eq:pdde

where $a_0(x) = -2\sin(x)$, $a_1(x) = 2\sin(x)$, $v_x(0,t) = v_x(\pi,t) = 0$.

Space derivatives approximated by central differences. This gives a delay eigenvalue problem of the considered form with one delay and sparse matrices A_0 and A_1. The dimension is $n = 5000$.

fig:newt

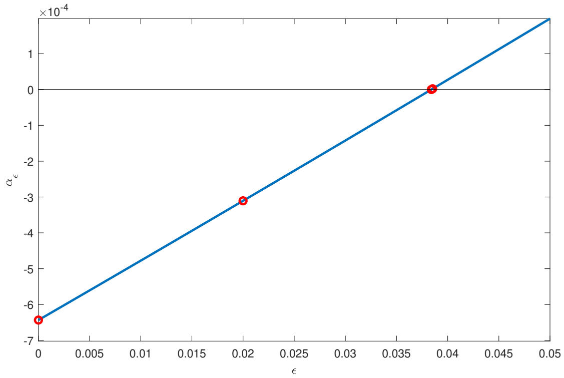

Fig. 4.5. The ε-pseudospectral abscissa $\alpha_\varepsilon = -\varphi(\varepsilon)$ and the Newton iterates for the space-discretized delayed PDE (4.17).

V.5 ε-stability radii

sec:eps-rad

Pseudospectra come with different uses:

- The structured and unstructured pseudospectra provide information as to how the spectrum changes under (possibly structured) perturbations;
- The complex unstructured pseudosectrum provides information on the transient behaviour of linear differential equations, for which the pseudospectral abscissa $\alpha_\varepsilon(A)$ and the stability radius $\varepsilon_\star$ with $\alpha_{\varepsilon_\star}(A) = 0$ are key quantities.

The two items are combined in the following matrix nearness problem, where $\mathcal{S} \subset \mathbb{C}^{n,n}$ is again a structure space, e.g., the space of real matrices or a space of complex or real matrices with a given sparsity pattern.

Problem. *Given $\varepsilon > 0$ and a matrix $A \in \mathbb{C}^{n,n}$ with negative (unstructured) ε-pseudospectral abscissa $\alpha_\varepsilon(A) < 0$, find a structured matrix $\Theta \in \mathcal{S}$ of minimal Frobenius norm such that the ε-pseudospectral abscissa of the perturbed matrix $A + \Theta$ is zero:*

$$\alpha_\varepsilon(A + \Theta) = 0.$$

This equation means that there exists an (unstructured) matrix $\varDelta \in \mathbb{C}^{n,n}$ of Frobenius norm ε such that the rightmost eigenvalues of $A + \varTheta + \varDelta$ are on the imaginary axis.

The Frobenius norm $\theta_\varepsilon = \theta_\varepsilon^{\mathcal{S}}(A)$ of a minimizer $\varTheta_\varepsilon$ is called the $\mathcal{S}$-structured ε-*stability radius* of the matrix A. Note that for $\varepsilon = 0$ this becomes the $\mathcal{S}$-structured stability radius considered previously, and also note that ε is the stability radius of $A + \varTheta_\varepsilon$. The problem can thus be rephrased as asking for a structured perturbation $\varTheta \in \mathcal{S}$ of minimal Frobenius norm such that $A + \varTheta$ has the prescribed stability radius ε.

For the unstructured case $\mathcal{S} = \mathbb{C}^{n,n}$ we have clearly $\theta_\varepsilon = \varepsilon_\star - \varepsilon$, where $\varepsilon_\star$ is the stability radius of A. For structured cases $\mathcal{S} \neq \mathbb{C}^{n,n}$, however, the $\mathcal{S}$-structured ε-stability radius θ_ε can be significantly larger than $\varepsilon_\star - \varepsilon$.

By definition of θ_ε, we have $\alpha_\varepsilon(A + \varTheta) \le 0$ for every $\varTheta \in \mathcal{S}$ with $\|\varTheta\|_F \le \theta_\varepsilon$. This implies that the stability radius of $A + \varTheta$ is at least ε, and so (III.1.10) yields the robust resolvent bound

$$\max_{\operatorname{Re}\lambda \ge 0} \|(A + \varTheta - \lambda I)^{-1}\|_2 \le \frac{1}{\varepsilon} \qquad \text{for every } \varTheta \in \mathcal{S} \text{ with } \|\varTheta\|_F \le \theta_\varepsilon. \tag{5.1}$$

`robust-res-bound`

As in (III.1.13), this implies that solutions to perturbed inhomogeneous linear differential equations $\dot{x}(t) = (A + \varTheta)x(t) + f(t)$ with zero initial value share the bound, for every perturbation $\varTheta \in \mathcal{S}$ with $\|\varTheta\|_F \le \theta_\varepsilon$,

$$\left(\int_0^T \|x(t)\|^2 \, dt \right)^{1/2} \le \frac{1}{\varepsilon} \left(\int_0^T \|f(t)\|^2 \, dt \right)^{1/2}, \qquad 0 \le T \le \infty. \tag{5.2}$$

`robust-L2-bound`

Moreover, we obtain a robust uniform bound of the matrix exponential: for every $\varTheta \in \mathcal{S}$ with $\|\varTheta\|_F \le \theta_\varepsilon$ and every $t > 0$,

$$\left\| \mathrm{e}^{t(A+\varTheta)} \right\|_2 \le \frac{1}{2\pi\varepsilon} \int_\varGamma |\mathrm{e}^{t\lambda}| \, |d\lambda| \le \frac{|\varGamma|}{2\pi\varepsilon}, \tag{5.3}$$

`robust-transient-bound`

where $\varGamma$ is a closed contour in the closed left complex half-plane that is a union of (i) the part in the left half-plane of a contour (or union of several contours) that surrounds the pseudospectrum $\varLambda_{\varepsilon+\theta_\varepsilon}(A)$ and (ii) one or several intervals on the imaginary axis that close the contour; and $|\varGamma|$ is the length of $\varGamma$. The proof of this bound is analogous to the proof of (III.1.11), using (i) $\varLambda_\varepsilon(A+\varTheta) \subset \varLambda_{\varepsilon+\theta_\varepsilon}(A)$, which implies $\|(A+\varTheta-\lambda I)^{-1}\|_2 \le 1/\varepsilon$ for λ in the closure of $\mathbb{C} \setminus \varLambda_{\varepsilon+\theta_\varepsilon}(A)$, and (ii) the bound (5.1) on the imaginary axis.

V.5.1 Two-level iteration

Let a fixed $\varepsilon > 0$ be given. The target eigenvalue $\lambda(M)$ of a matrix M is again chosen as an eigenvalue of M of maximal real part (and among those, the one with maximal imaginary part). For varying $\theta > 0$ we introduce the functional

$$F_\theta(E^{\mathcal{S}}, E) = -\operatorname{Re} \lambda(A + \theta E^{\mathcal{S}} + \varepsilon E) \tag{5.4}$$

`F-delta`

for $E^{\mathcal{S}} \in \mathcal{S}$ and $E \in \mathbb{C}^{n,n}$, both of unit Frobenius norm. With this functional we follow the two-level approach of Section IV.2:

- **Inner iteration:** For a given $\theta > 0$, we aim to compute matrices $E^{\mathcal{S}}(\theta) \in \mathcal{S}$ and $E(\theta) \in \mathbb{C}^{n,n}$, both of unit Frobenius norm, that minimize F_θ:

$$(E^{\mathcal{S}}(\theta), E(\theta)) = \arg \min_{\substack{E^{\mathcal{S}} \in \mathcal{S}, E \in \mathbb{C}^{n,n} \\ \|E^{\mathcal{S}}\|_F = \|E\|_F = 1}} F_\theta(E^{\mathcal{S}}, E). \tag{5.5}$$

E-delta

- **Outer iteration:** We compute the smallest positive value θ_ε with

$$\phi(\theta_\varepsilon) = 0, \tag{5.6}$$

zero-delta

where $\phi(\theta) = F_\theta(E^{\mathcal{S}}(\theta), E(\theta)) = \alpha_\varepsilon\big(A + \theta E^{\mathcal{S}}(\theta)\big)$.

Provided that these computations succeed, we have that $\Theta_\varepsilon = \theta_\varepsilon E^{\mathcal{S}}(\theta_\varepsilon) \in \mathcal{S}$ is a solution to the matrix nearness problem stated above, and θ_ε is the $\mathcal{S}$-structured ε-stability radius of A.

V.5.2 Rank-1 matrix differential equations for the inner iteration

r1-ode-eps-rad

We combine the procedures of Sections II.1 and II.3. As in these sections, we find that along a path $(E^{\mathcal{S}}(t), E(t))$ in $\mathcal{S} \times \mathbb{C}^{n,n}$ we have, assuming simple target eigenvalues $\lambda(A + \theta E^{\mathcal{S}}(t) + \varepsilon E(t))$,

$$\begin{aligned} \frac{1}{\kappa(t)} \frac{d}{dt} F_\theta(E^{\mathcal{S}}(t), E(t)) = \mathrm{Re}\langle \Pi^{\mathcal{S}} G_\theta(E^{\mathcal{S}}(t), E(t)), \theta \dot{E}^{\mathcal{S}}(t)\rangle \\ + \mathrm{Re}\langle G_\theta(E^{\mathcal{S}}(t), E(t)), \varepsilon \dot{E}(t)\rangle \end{aligned}$$

with the rescaled gradient

$$G_\theta(E^{\mathcal{S}}, E) = xy^*,$$

where x and y are left and right eigenvectors, of unit norm and with positive inner product, associated with the simple target eigenvalue λ of $A + \theta E^{\mathcal{S}} + \varepsilon E$. Moreover, $\kappa = 1/(x^* y)$, and $\Pi^{\mathcal{S}}$ again denotes the orthogonal projection onto the structure space $\mathcal{S}$.

Norm-constrained gradient flow. Along solutions of the system of differential equations, with $G = G_\theta(E^{\mathcal{S}}, E)$ for short,

$$\begin{aligned} \theta \dot{E}^{\mathcal{S}} &= -\Pi^{\mathcal{S}} G + \mathrm{Re}\langle G, E^{\mathcal{S}}\rangle E^{\mathcal{S}} \\ \varepsilon \dot{E} &= -G + \mathrm{Re}\langle G, E\rangle E, \end{aligned} \tag{5.7}$$

ode-ES-E

the unit norms of $E^{\mathcal{S}}(t) \in \mathcal{S}$ and $E(t)$ are preserved, and the functional $F_\theta(E^{\mathcal{S}}(t), E(t))$ decreases monotonically. Provided that $\Pi^{\mathcal{S}} G \neq 0$, we have at stationary points that $E^{\mathcal{S}}$ is a real multiple of $\Pi^{\mathcal{S}} G$ and E is a real multiple of G, and we note that G is of rank 1.

Rank-1 matrix differential equations. To make use of the rank-1 structure of optimizers, we follow the approaches of Sections II.1 and II.3 and consider differential equations for rank-1 matrices $E(t)$ and $Y(t)$, where the latter yields $E^{\mathcal{S}}(t) = \Pi^{\mathcal{S}} Y(t)$. These differential equations are obtained from (5.7) by replacing $G = G_\theta(E^{\mathcal{S}}, E)$ by its projection onto the tangent space of the manifold of rank-1 matrices at E and Y, respectively:

$$\begin{aligned} \theta \dot{Y} &= -P_Y G + \mathrm{Re}\langle P_Y G, E^{\mathcal{S}} \rangle Y \quad \text{ with } \; E^{\mathcal{S}} = \Pi^{\mathcal{S}} Y, \\ \varepsilon \dot{E} &= -P_E G + \mathrm{Re}\langle G, E \rangle E. \end{aligned} \tag{5.8}$$

ode-ES-E-1

These differential equations yield rank-1 matrices $Y(t)$ and $E(t)$ and preserve the unit Frobenius norm of $E^{\mathcal{S}}(t)$ and $E(t)$. As in Sections II.1 and II.3 it is shown that under a nondegeneracy condition, the stationary points (Y, E) of (5.8) correspond bijectively to the stationary points $(E^{\mathcal{S}}, E)$ of (5.7) via $E^{\mathcal{S}} = \Pi^{\mathcal{S}} Y$ and with the same E. The differential equations are integrated numerically into a stationary point $(E^{\mathcal{S}}, E)$ in the way described in Sections II.1 and II.3, working with the vectors that define the rank-1 matrices Y and E and advancing them in time using a tailor-made splitting method.

V.5.3 Outer iteration, updating θ

For the solution of the scalar equation $\phi(\theta) = 0$ we use a combined Newton / bisection method as in Section IV.2. The derivative of ϕ for the Newton iteration is obtained with the arguments of the proof of Theorem IV.2.2 (under analogous assumptions), yielding essentially the same formula,

$$\phi'(\theta) = -\kappa(\theta)\, \| \Pi^{\mathcal{S}} G_\theta(E^{\mathcal{S}}(\theta), E(\theta)) \|_F = -\kappa(\theta)\, \| \Pi^{\mathcal{S}} (x(\theta) y(\theta)^*) \|_F,$$

where $x(\theta)$ and $y(\theta)$ are left and right normalized eigenvectors associated with the rightmost eigenvalue of $A + \theta E^{\mathcal{S}}(\theta) + \varepsilon E(\theta)$, and $\kappa(\theta) = 1/(x(\theta)^* y(\theta)) > 0$.

V.6 Notes

V.6.1 Hamiltonian matrix nearness problems

Hamiltonian eigenvalue perturbation problems. Such problems were studied in detail by Mehrmann & Xu (2008) and Alam, Bora, Karow, Mehrmann & Moro (2011), motivated by the passivation of linear control systems and the stabilization of gyroscopic mechanical systems, where eigenvalues of Hamiltonian matrices need to be moved to or away from the imaginary axis. In this context, a solution to Problem A considered here yields a lower bound on the distance to non-passivity of a passive system and a solution to Problem B yields a lower bound of the distance to passivity of a non-passive system. The stabilization of a gyroscopic system requires to move all eigenvalues of a Hamiltonian matrix onto the imaginary axis. Those applications have an additional structure of admissible perturbations, which have not been taken into account in this chapter where we consider general real Hamiltonian perturbations. Understanding this general case is, however, basic to addressing the more specific demands of the applications to control systems or mechanical systems. This will become clear in Section VI.3, where we describe algorithms for finding the nearest passive or non-passive system, which are conceptually close to our treatment of Problems B and A, respectively.

Two-level approach to Hamiltonian matrix nearness problems. Guglielmi, Kressner & Lubich (2015) studied a two-level approach to Problems A and B in the matrix 2-norm instead of the Frobenius norm as considered here. This equally leads to rank-4 differential equations along which the real part of the target eigenvalue decreases (or increases) monotonically. Contrary to the Frobenius norm case, those differential equations cannot be interpreted as constrained gradient systems.

Theorem 1.7 on the square root behaviour of the real parts of eigenvalues near a defective coalescence on the imaginary axis was first stated by Guglielmi, Kressner & Lubich (2015). Here we give a corrected proof based on Theorem 1.5 about the eigenvectors at a defective coalescence, which was first proved by Fazzi, Guglielmi & Lubich (2021).

Complex Hamiltonian matrices. An analogous approach to this chapter can be given for complex Hamiltonian matrices (i.e. matrices A for which JA is hermitian). This case is slightly simpler, as it leads to rank-2 differential equations. Guglielmi, Kressner & Lubich (2015) studied the two-level approach to matrix nearness problems in the complex Hamiltonian case for the matrix 2-norm.

Hamiltonian eigenvalue solver. Structure-preserving eigenvalue solvers as implemented in the SLICOT library (http://slicot.org/), see Benner, Mehrmann, Sima, Van Huffel & Varga (1999), are distinctly favourable over using a standard general eigenvalue solver, especially for eigenvalues on and close to the imaginary axis as are of interest here; see, e.g., Benner, Losse, Mehrmann & Voigt (2015) and references therein.

Chapter VI.
Control systems

chap:lti

In this chapter we reconsider basic problems of robust control of linear time-invariant systems, which we rephrase as eigenvalue optimization problems and matrix (or operator) nearness problems. The problems considered from this perspective include the following:

- computing the $\mathcal{H}_\infty$-norm of the matrix transfer function, which is the L^2-norm of the input-output map;
- computing the $\mathcal{H}_\infty$-distance to uncontrollability of a controllable system;
- passivity enforcement by a perturbation to a system matrix that minimizes the $\mathcal{H}_\infty$-norm of the perturbation to the matrix transfer function; and
- computing the distance to loss of contractivity under structured perturbations to the state matrix.

The distances from systems with undesired properties are important robustness measures of a given control system, whereas algorithms for finding a nearby control system with prescribed desired properties are important design tools. The algorithmic approach to eigenvalue optimization via low-rank matrix differential equations and the two-level approach to matrix nearness problems of previous chapters is extended to a variety of problems from the area of robust control and is shown to yield versatile and efficient algorithms.

We consider the *continuous-time linear time-invariant dynamical system* with inputs $u(t) \in \mathbb{C}^p$, outputs $y(t) \in \mathbb{C}^m$ and states $z(t) \in \mathbb{C}^n$ related by

$$\begin{aligned} \dot{z}(t) &= Az(t) + Bu(t) \\ y(t) &= Cz(t) + Du(t) \end{aligned} \tag{0.1}$$

with the real system matrices $A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,p}$, $C \in \mathbb{R}^{m,n}$ and $D \in \mathbb{R}^{m,p}$, and with the initial state $z(0) = 0$. In this chapter we always assume that all eigenvalues of A have negative real part.

In the final section of this chapter we consider *descriptor systems*, where $\dot{z}(t)$ in (0.1) appears multiplied with a singular matrix $E \in \mathbb{R}^{n,n}$, which yields a differential-algebraic equation instead of the differential equation in (0.1). Descriptor systems play an essential role in modeling and composing networks of systems. We present an algorithm for computing the $\mathcal{H}_\infty$-norm of the associated matrix transfer function, which now needs to be appropriately weighted if the descriptor system has an index higher than 1.

VI.1 $\mathcal{H}_\infty$-norm of the matrix transfer function

sec:Hinf

The matrix-valued *transfer function* associated with the system (0.1) is

$$H(\lambda) = C(\lambda I - A)^{-1}B + D \quad \text{for } \lambda \in \mathbb{C}\backslash\Lambda(A) \tag{1.1}$$

tfmatdef

where $\Lambda(A)$ denotes the spectrum of A. If all eigenvalues of A have negative real part, as we assumed, then H is a matrix-valued holomorphic function on a domain that includes the closed right half-plane $\mathrm{Re}\,\lambda \geq 0$.

The input–output map $u \mapsto y$ given by the variation-of-constants formula,

$$y(t) = \int_0^t Ce^{(t-\tau)A}Bu(\tau)\,d\tau + Du(t), \quad t \geq 0,$$

is the convolution with the inverse Laplace transform of H:

$$y = H(\partial_t)u := (\mathcal{L}^{-1}H) * u.$$

Taking Laplace transforms, we get the formula that explains the name "transfer function",

$$\mathcal{L}y(\lambda) = H(\lambda)\,\mathcal{L}u(\lambda), \quad \mathrm{Re}\,\lambda \geq 0. \tag{1.2}$$

Hu

The Plancherel formula for the Fourier transform $(\mathcal{F}y)(\omega) = (\mathcal{L}y)(\mathrm{i}\omega)$ for $\omega \in \mathbb{R}$ (where y is extended by zero to $t < 0$) then yields that the operator norm of the input–output map $H(\partial_t) : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^m)$ equals $\sup_{\omega\in\mathbb{R}} \|H(\mathrm{i}\omega)\|_2$, since

$$\begin{aligned}\int_0^\infty \|y(t)\|^2\,dt &= \int_\mathbb{R} \|(\mathcal{L}y)(\mathrm{i}\omega)\|^2\,d\omega = \int_\mathbb{R} \|H(\mathrm{i}\omega)(\mathcal{L}u)(\mathrm{i}\omega)\|^2\,d\omega \\ &\leq \sup_{\omega\in\mathbb{R}} \|H(\mathrm{i}\omega)\|_2^2 \int_\mathbb{R} \|(\mathcal{L}u)(\mathrm{i}\omega)\|^2\,d\omega = \sup_{\omega\in\mathbb{R}} \|H(\mathrm{i}\omega)\|_2^2 \int_0^\infty \|u(t)\|^2\,dt\end{aligned}$$

and an approximate δ-function argument shows that $\sup_{\omega\in\mathbb{R}} \|H(\mathrm{i}\omega)\|_2^2$ is the smallest such bound that holds for all square-integrable functions u.

ef:hinfnormcont

Definition 1.1. The $\mathcal{H}_\infty$*-norm* of the matrix transfer function H is

$$\|H\|_\infty := \sup_{\mathrm{Re}\,\lambda\geq 0} \|H(\lambda)\|_2 = \sup_{\omega\in\mathbb{R}} \|H(\mathrm{i}\omega)\|_2. \tag{1.3}$$

hinfnormequivcont

The supremum is a maximum if $\|D\|_2 = \|H(\infty)\|_2$ is strictly smaller than $\|H\|_\infty$, as will be assumed from now on. The second equation in (1.3) is a consequence of the maximum principle, which can be applied on noting that $\|H(\lambda)v\|^2$ is a subharmonic function of λ for every $v \in \mathbb{C}^n$.

As the L^2 operator norm of the input-output map $H(\partial_t)$, the $\mathcal{H}_\infty$-norm of H is a fundamental stability measure. Moreover, using the causality property that $y(t)$ depends only on $u(\tau)$ for $\tau \leq t$, we can rewrite the above bound as

$$\left(\int_0^T \|y(t)\|^2\,dt\right)^{1/2} \leq \|H\|_\infty \left(\int_0^T \|u(t)\|^2\,dt\right)^{1/2}, \qquad 0 \leq T \leq \infty, \tag{1.4}$$

yu-bound

and $\|H\|_\infty$ is the smallest such bound.

Problem. *Compute the $\mathcal{H}_\infty$-norm of the matrix transfer function H.*

Making use of the theory of spectral value sets, which are suitable extensions of pseudospectra presented e.g. in the monograph by Hinrichsen and Pritchard (2005), we will show that the $\mathcal{H}_\infty$-norm equals the reciprocal of the stability radius, which here is the largest value of ε such that the associated ε-spectral value set is contained in the left half-plane.

This characterization allows us to extend the algorithmic approach of previous sections, using rank-1 constrained gradient systems, from pseudospectra to spectral value sets, and then use a scalar Newton-bisection method to approximate the $\mathcal{H}_\infty$-norm.

VI.1.1 Matrix transfer function and perturbed state matrices

subsec:tf-psm

We start with discussing the relationship between the singular vectors of the transfer matrix $H(\lambda)$ and the eigenvectors of a corresponding set of matrices.

Given A, B, C, D defining the linear dynamical system (0.1), consider the *perturbed state matrix*, for perturbations $\varDelta \in \mathbb{C}^{p,m}$ such that $I - D\varDelta$ is invertible,

$$M(\varDelta) = A + B\varDelta(I - D\varDelta)^{-1}C \in \mathbb{C}^{n,n} \tag{1.5}$$

AEdef

and the associated transfer matrix (0.1). The next theorem relates the 2-norm of the transfer matrix, which is its largest singular value, to eigenvalues of perturbed state matrices. This result extends the basic characterization of complex pseudospectra given in Theorem III.1.2 together with (III.1.8), to which it reduces for $B = C = I$ and $D = 0$.

thm:basicequiv

Theorem 1.2 (Singular values and eigenvalues). *Let $\varepsilon > 0$ and $\varepsilon\|D\|_2 < 1$. Then, for $\lambda \notin \varLambda(A)$ the following two statements are equivalent:*

(i) $\|H(\lambda)\|_2 \ge \varepsilon^{-1}$
twoequiv *(ii) λ is an eigenvalue of $M(\varDelta)$ for some $\varDelta \in \mathbb{C}^{p,m}$ with $\|\varDelta\|_2 \le \varepsilon$.*

Moreover, $\varDelta$ can be chosen to have rank 1, and the two inequalities can be replaced by equalities in the equivalence.

Proof. We first observe that under the condition $\varepsilon\|D\|_2 < 1$ we have that $I - D\varDelta$ is invertible when $\|\varDelta\|_2 \le \varepsilon$, and hence $M(\varDelta)$ is then well-defined.

Suppose (i) holds true, with $\rho = \|H(\lambda)\|_2^{-1} \le \varepsilon$. Let u and v be right and left singular vectors of $H(\lambda)$, respectively, corresponding to the largest singular value ρ^{-1}, so that

$$\rho H(\lambda)u = v, \quad \rho v^* H(\lambda) = u^*, \quad \text{and} \quad \|u\| = \|v\| = 1.$$

Define $\varDelta = \rho uv^*$ so that $\|\varDelta\|_2 = \rho \le \varepsilon$. We have $H(\lambda)\varDelta = vv^*$, so

$$(C(\lambda I - A)^{-1}B + D)\varDelta v = v. \tag{1.6}$$

tfmatEv

Next define $Y = (I - D\varDelta)^{-1}C$ and $Z = (\lambda I - A)^{-1}B\varDelta$, so we have $YZv = v$. It follows that $Zv \ne 0$ and $ZYy = y$, with $y := Zv = \rho(\lambda I - A)^{-1}Bu$ an eigenvector of ZY. Multiplying through by $\lambda I - A$, we have

$$B\Delta(I - D\Delta)^{-1}Cy = (\lambda I - A)y, \tag{1.7}$$ xeigvec

which is equivalent to $M(\Delta)y = \lambda y$. This proves the second statement in (1.2).

Suppose (i) holds true, with $\rho = \|H(\lambda)\|_2^{-1} \leq \varepsilon$. Let u and v be right and left singular vectors of $H(\lambda)$, respectively, corresponding to the largest singular value ρ^{-1}, so that

$$\rho H(\lambda)u = v, \quad \rho v^* H(\lambda) = u^*, \quad \text{and} \quad \|u\| = \|v\| = 1.$$

Define $\Delta = \rho uv^*$ so that $\|\Delta\|_2 = \rho \leq \varepsilon$. We have $H(\lambda)\Delta = vv^*$, so

$$H(\lambda)\Delta v = v. \tag{1.8}$$ tfmatEv

With $Z = (\lambda E - A)^{-1}B\Delta$ we thus have $CZv = v$. It follows that $Zv \neq 0$ and $ZCy = y$, with $y := Zv = \rho(\lambda E - A)^{-1}Bu$ an eigenvector of ZC. Multiplying through by $\lambda I - A$, we have

$$B\Delta(I - D\Delta)^{-1}Cy = (\lambda I - A)y, \tag{1.9}$$ xeigvec

which is equivalent to $M(\Delta)y = \lambda y$. This proves the second statement in (1.2).

Conversely, suppose that (ii) holds true. Then there exists $y \neq 0$ such that (1.9) holds. We have $ZYy = y$, so y is an eigenvector of ZY corresponding to the eigenvalue 1. Consequently, $YZw = w$ where $w = Yy \neq 0$ is an eigenvector of YZ. Multiplying by $I - D\Delta$ and rearranging we have

$$(C(\lambda I - A)^{-1}B + D)\Delta w = w, \quad \text{i.e.,} \quad H(\lambda)\Delta w = w.$$

This implies

$$\varepsilon\|H(\lambda)\|_2 \geq \|H(\lambda)\Delta\|_2 \geq 1,$$

which proves the first statement in (1.2).

The equivalence (1.2) also holds if we restrict Δ in the second statement to have rank one. The proof remains unchanged. □

We reformulate the remarkable relationship between eigenvectors of $M(\Delta)$ and singular vectors of $H(\lambda)$ revealed by the previous proof in a separate corollary.

thm:evecssvecs **Corollary 1.3 (Singular vectors and eigenvectors).** *Let $\varepsilon > 0$ and $\varepsilon\|D\|_2 < 1$, and let $u \in \mathbb{C}^p$ and $v \in \mathbb{C}^m$ with $\|u\| = \|v\| = 1$ be right and left singular vectors of $H(\lambda)$, respectively, corresponding to a singular value ε^{-1}. Then, the nonzero vectors*

$$\widetilde{y} = (\lambda I - A)^{-1}Bu \quad \text{and} \quad \widetilde{x} = (\lambda I - A)^{-*}C^*v, \tag{1.10}$$ xyformulas

are (non-normalized) right and left eigenvectors associated with the eigenvalue λ of $M(\Delta)$ for $\Delta = \varepsilon uv^*$.

Proof. In the proof of Theorem 1.2 we showed that $\widetilde{y}$ is a right eigenvector of $M(\Delta)$ for $\Delta = \varepsilon uv^*$ to the eigenvalue λ. The proof for the left eigenvector $\widetilde{x}$ is analogous. □

VI.1.2 Spectral value sets

sec:specvalsets

We define spectral value sets, which generalize the notion of pseudospectrum of a matrix A to linear control systems with the matrices (A, B, C, D), and we show their relationship with the 2-norm of the matrix transfer function.

Definition 1.4. Let $\varepsilon \geq 0$ and $\varepsilon\|D\|_2 < 1$, and define the *spectral value set*

$$\Lambda_\varepsilon(A,B,C,D) = \bigcup \{\Lambda(M(\Delta)) : \Delta \in \mathbb{C}^{p,m}, \|\Delta\|_2 \leq \varepsilon\}\,.$$

Note that $\Lambda_\varepsilon(A,B,C,D) \supset \Lambda_0(A,B,C,D) = \Lambda(A)$, and note further that $\Lambda_\varepsilon(A,I,I,0)$ equals the ε-pseudospectrum $\Lambda_\varepsilon(A)$. The following corollary of Theorem 1.2 is immediate.

cor:SVSequiv

Corollary 1.5 (Characterization of the spectral value set). *Let $\varepsilon > 0$ and $\varepsilon\|D\|_2 < 1$. Then,*

$$\begin{aligned}\Lambda_\varepsilon(A,B,C,D)\backslash\Lambda(A) &= \bigcup \{\Lambda(M(\Delta)) : \Delta \in \mathbb{C}^{p,m}, \|\Delta\|_2 \leq \varepsilon, \mathrm{rank}(\Delta) = 1\} \\ &= \bigcup \left\{\lambda \in \mathbb{C}\backslash\Lambda(A) : \|H(\lambda)\|_2 \geq \varepsilon^{-1}\right\}.\end{aligned}$$

VI.1.3 $\mathcal{H}_\infty$-norm and stability radius

sec:Hinfcont

For $\varepsilon \geq 0$ with $\varepsilon\|D\|_2 < 1$, the *spectral value set abscissa* is

$$\alpha_\varepsilon(A,B,C,D) = \max\{\mathrm{Re}\,\lambda : \lambda \in \Lambda_\varepsilon(A,B,C,D)\} \qquad (1.11)$$

alepsdef

with $\alpha_0(A,B,C,D) = \alpha(A)$, the spectral abscissa of A. This definition extends the notion of the pseudospectral abscissa $\alpha_\varepsilon(A)$.

The $\mathcal{H}_\infty$-norm can be characterized as the reciprocal of the *stability radius*, which is the largest ε such that $\Lambda_\varepsilon(A,B,C,D)$ is contained in the left half-plane. The following theorem states this remarkable equality on which our algorithmic approach to computing the $\mathcal{H}_\infty$-norm will be based. It extends (III.1.10), to which it reduces for $B = C = I$ and $D = 0$.

thm:Hinf-oeps

Theorem 1.6 ($\mathcal{H}_\infty$-norm via the stability radius). *Assume that all eigenvalues of A have negative real part. Let the stability radius of the system (A,B,C,D) be*

$$\varepsilon_\star := \inf\{\varepsilon > 0 \textit{ with } \varepsilon\|D\|_2 < 1 \,:\, \alpha_\varepsilon(A,B,C,D) = 0\},$$

where $\alpha_\varepsilon(A,B,C,D)$ is the spectral value set abscissa defined in (1.11). *Then,*

$$\|H\|_\infty = \frac{1}{\varepsilon_\star}. \qquad (1.12)$$

hinfnormdefcont

Proof. We first consider the case where $\Lambda_\varepsilon(A,B,C,D)$ does not intersect the imaginary axis for any $\varepsilon > 0$ with $\varepsilon\|D\|_2 < 1$. Then we take the infimum in (1.12) to be $1/\|D\|_2$. By Corollary 1.5, $\|H(\mathrm{i}\omega)\| < \varepsilon^{-1}$ for all $\omega \in \mathbb{R}$ and all ε with $\varepsilon\|D\|_2 < 1$, and hence the supremum in (1.3) is at least $\|D\|_2$, and therefore equal to $\|D\|_2$ as is seen by letting $\omega \to \pm\infty$. So we have equality in (1.12) in this degenerate case.

Otherwise, there exists a smallest $\varepsilon_\star$ with $\varepsilon_\star\|D\|_2 < 1$ such that $\alpha_{\varepsilon_\star}(A,B,C,D) = 0$. So there exists $\omega_\star \in \mathbb{R}$ such that $\mathrm{i}\omega_\star \in \Lambda_{\varepsilon_\star}(A,B,C,D)$. By Corollary 1.5, this implies $\|H(\mathrm{i}\omega_\star)\|_2 \geq 1/\varepsilon_\star$. Here we have actually equality, because $\|H(\mathrm{i}\omega_\star)\|_2 = 1/\varepsilon$ with $\varepsilon < \varepsilon_\star$ would imply, again by Corollary 1.5, that $\mathrm{i}\omega_\star \in \Lambda_\varepsilon(A,B,C,D)$ and hence $\alpha_\varepsilon(A,B,C,D) = 0$, which contradicts the minimality of $\varepsilon_\star$. □

It follows from Corollary 1.5 that, for $\varepsilon > 0$ with $\varepsilon\|D\|_2 < 1$, the spectral value set abscissa in (1.11) equals

$$\alpha_\varepsilon(A,B,C,D) = \max\left\{\mathrm{Re}\,\lambda : \lambda \in \Lambda(A) \text{ or } \|H(\lambda)\|_2 \geq \varepsilon^{-1}\right\}. \tag{1.13}$$ alepsdef2

The set of admissible λ must include $\Lambda(A)$ because of the possibility that the spectral value set $\Lambda_\varepsilon(A,B,C,D)$ has isolated points. Excluding such points, we obtain local optimality conditions for (1.13).

In order to proceed we make the following generic assumptions.

assumptcont **Assumption 1.7.** Let $\varepsilon > 0$ with $\varepsilon\|D\|_2 < 1$, and let $\lambda \notin \Lambda(A)$ be a locally rightmost point of $\Lambda_\varepsilon(A,B,C,D)$. We assume:

1. The largest singular value ε^{-1} of $H(\lambda)$ is simple.
2. Letting u and v be corresponding right and left singular vectors and setting $\Delta = \varepsilon uv^*$, the eigenvalue λ of $M(\Delta)$ is simple.

Here we note that ε^{-1} equals the largest singular value of $H(\lambda)$, i.e. $\|H(\lambda)\|_2$, by Corollary 1.5 and the minimality argument at the end of the proof of Theorem 1.6, and λ is an eigenvalue of $M(\Delta)$ by Theorem 1.2.

firstordercont **Lemma 1.8 (Eigenvectors with positive inner product).** *Let $\varepsilon > 0$ with $\varepsilon\|D\|_2 < 1$, and let $\lambda \notin \Lambda(A)$ be a locally rightmost point of $\Lambda_\varepsilon(A,B,C,D)$. Under Assumption 1.7, we then have that*

$$\widetilde{x}^*\widetilde{y} \ \textit{ is real and positive}, \tag{1.14}$$ firstordercont

where $\widetilde{y}$ and $\widetilde{x}$ are the (non-normalized) right and left eigenvectors to the eigenvalue λ of $M(\Delta)$ with $\Delta = \varepsilon uv^*$ that, via (1.10)*, correspond to the right and left singular vectors u and v associated with the largest singular value ε^{-1} of $H(\lambda)$.*

Proof. The standard first-order necessary condition for $\widehat{\zeta} \in \mathbb{R}^2$ to be a local maximizer of an optimization problem $\max\{f(\zeta) : g(\zeta) \leq 0,\ \zeta \in \mathbb{R}^2\}$, when f, g are continuously differentiable and $g(\widehat{\zeta}) = 0$, $\nabla g(\widehat{\zeta}) \neq 0$, is the existence of a Lagrange multiplier $\mu \geq 0$ such that $\nabla f(\widehat{\zeta}) = \mu \nabla g(\widehat{\zeta})$. In our case, identifying $\lambda \in \mathbb{C}$ with $\zeta \in \mathbb{R}^2$, the gradient of the maximization objective is $(1,0)^T$, while the constraint function

$$\frac{1}{\varepsilon} - \|C\,(\lambda I - A)^{-1}\,B + D\|_2$$

is differentiable with respect to λ because of the first part of Assumption 1.7, and it has the gradient

$$\begin{pmatrix} \mathrm{Re}(v^* C(\lambda I - A)^{-2} Bu) \\ \mathrm{Im}(v^* C(\lambda I - A)^{-2} Bu) \end{pmatrix}$$

using standard perturbation theory for singular values. Defining $\varDelta = \varepsilon u v^*$ and applying Theorem 1.3 we know that $\widetilde{x}$ and $\widetilde{y}$ as defined in (1.10) are left and right eigenvectors of $M(\varDelta)$, with inner product

$$\widetilde{x}^* \widetilde{y} = v^* C(\lambda I - A)^{-2} Bu. \tag{1.15}$$

evecip

By the second part of Assumption 1.7, λ is a simple eigenvalue of $M(\varDelta)$ and so $\widetilde{x}^*\widetilde{y} \neq 0$. Therefore, the constraint gradient is nonzero implying that the Lagrange multiplier $\mu > 0$ exists with $v^* C(\lambda I - A)^{-2} Bu = 1/\mu > 0$, and by (1.15) we thus find $\widetilde{x}^*\widetilde{y} > 0$. □

VI.1.4 Two-level iteration

Like for the matrix nearness problems in Chapter IV, we approach the computation of the $\mathcal{H}_\infty$-norm by a two-level method:

- **Inner iteration:** Given $\varepsilon > 0$, we aim to compute a matrix $E(\varepsilon)$ of rank 1 and of unit Frobenius norm, such that the functional

  $$F_\varepsilon(E) = -\frac{\lambda + \overline{\lambda}}{2} = -\,\mathrm{Re}(\lambda), \qquad \text{for } \lambda = \lambda(M(\varepsilon E)),$$

  where $\lambda(M)$ is the rightmost eigenvalue of a matrix M, is minimized in the manifold of rank-1 matrices of unit norm, i.e.

  $$E(\varepsilon) = \arg \min_{E \in \mathcal{M}_1, \|E\|_F = 1} F_\varepsilon(E). \tag{1.16}$$

  E-epsl

  The obtained optimizer is denoted by $E(\varepsilon)$ to emphasize its dependence on ε, and the rightmost eigenvalue of $M\,(\varepsilon E(\varepsilon))$ is denoted by $\lambda(\varepsilon)$. (Note that the Frobenius norm and the matrix 2-norm are the same for a rank-1 matrix.)
- **Outer iteration:** We compute the smallest positive value $\varepsilon_\star$ with

  $$\phi(\varepsilon_\star) = 0, \tag{1.17}$$

  eq:zero

  where $\phi(\varepsilon) = F_\varepsilon\,(E(\varepsilon)) = -\mathrm{Re}\,\lambda(\varepsilon) = -\alpha_\varepsilon(A, B, C, D)$ is minus the spectral value set abscissa.

If the numerical result computed by such a two-level iteration were exact, it would yield the $\mathcal{H}_\infty$-norm in view of Theorem 1.6,

$$\|H\|_\infty = \frac{1}{\varepsilon_\star}.$$

VI.1.5 Norm-constrained gradient flow

specvalsetabsc

In this and the next subsection we show how to deal with the inner iteration, following a programme that directly extends the programme of Section II.1.

As in (1.5), consider the perturbed matrix, for $\varDelta \in \mathbb{C}^{p,m}$,

$$M\left(\varDelta\right) = A + B\varDelta\left(I - D\varDelta\right)^{-1} C.$$

We consider the eigenvalue optimization problem (1.16), but for the moment with respect to all complex perturbations $E \in \mathbb{C}^{p,m}$ of unit Frobenius norm, although later we will restrict to rank-1 matrices E. So we look for

$$\arg \min_{E \in \mathbb{C}^{p,m},\, \|E\|_F = 1} F_\varepsilon(E). \tag{1.18}$$

eq:optim-sys

To treat this eigenvalue optimization problem, we will closely follow the course of Section II.1 and adapt it to the present situation.

Free gradient. To derive the gradient of the functional F_ε, we first state a simple auxiliary result.

Lemma 1.9 (Derivative of the perturbed matrix). *Given a smooth matrix valued function $\varDelta(t)$ with $\|\varDelta(t)\|_2 \|D\|_2 < 1$, we have*

$$\frac{d}{dt} M(\varDelta(t)) = B\left(I - \varDelta(t)D\right)^{-1} \dot{\varDelta}(t)\left(I - D\varDelta(t)\right)^{-1} C. \tag{1.19}$$

lem:derE

Proof. For conciseness, we omit the dependence on t, differentiate and regroup terms as

$$\begin{aligned} \frac{d}{dt}\left(\varDelta\left(I - D\varDelta\right)^{-1}\right) &= \dot{\varDelta}\left(I - D\varDelta\right)^{-1} + \varDelta \frac{d}{dt}\left(I - D\varDelta\right)^{-1} \\ &= \dot{\varDelta}\left(I - D\varDelta\right)^{-1} + \varDelta\left(I - D\varDelta\right)^{-1} D\dot{\varDelta}\left(I - D\varDelta\right)^{-1} \\ &= \left(I + \varDelta\left(I - D\varDelta\right)^{-1} D\right)\dot{\varDelta}\left(I - D\varDelta\right)^{-1}. \end{aligned} \tag{1.20}$$

eq:matrix-Ft-deriv

We then observe that

$$I + \varDelta\left(I - D\varDelta\right)^{-1} D = I + \varDelta\left(\sum_{k=0}^{\infty}(D\varDelta)^k\right) D = I + \sum_{k=1}^{\infty}(\varDelta D)^k = \left(I - \varDelta D\right)^{-1}. \tag{1.21}$$

eq:matrix-inf-series

Combining (1.20) and (1.21) yields

$$\frac{d}{dt}\left(\varDelta(t)\left(I - D\varDelta(t)\right)^{-1}\right) = \left(I - \varDelta(t)D\right)^{-1} \dot{\varDelta}(t)\left(I - D\varDelta(t)\right)^{-1}, \tag{1.22}$$

which implies the result. □

We will from now on use a normalization of the eigenvectors of $M(\Delta)$ of (1.5), which we previously considered in this book:

$$\|x\| = \|y\| = 1 \quad \text{and} \quad x^*y \text{ is real and positive.} \tag{1.23}$$

eq:scaling-sys

In the following we work with the normalized left and right eigenvectors x and y instead of the non-normalized $\widetilde{x}$ and $\widetilde{y}$ of Corollary 1.3.

ambdaderiv-sys

Lemma 1.10 (Derivative of a simple eigenvalue). *Let $\Delta(t)$ be a smooth matrix valued function with $\|\Delta(t)\|_2\|D\|_2 \le 1$. Suppose that $\lambda(t)$ is a simple eigenvalue of $M(\Delta(t))$ depending continuously on t, with associated eigenvectors $x(t)$ and $y(t)$ normalized according to* (1.23)*, and let $\kappa(t) = 1/(x(t)^*y(t)) > 0$. Then, $\lambda(t)$ is differentiable with*

$$\dot\lambda(t) = \kappa(t)\, r(t)^* \dot\Delta(t) s(t)$$

with

$$r(t) = \left(I - \Delta(t)D\right)^{-*} B^* x(t), \quad s(t) = \left(I - D\Delta(t)\right)^{-1} C y(t). \tag{1.24}$$

eq:rsdef

Proof. Applying Theorem VIII.1.1 we get

$$\dot\lambda = \frac{x^*\dot M y}{x^*y}$$

with

$$\dot M = B\left(I - \Delta D\right)^{-1} \dot\Delta \left(I - D\Delta\right)^{-1} C \tag{1.25}$$

eq:optprob

where we omitted the dependence on t for brevity. The result is then immediate. □

A direct consequence of Lemma 1.10, for $\Delta(t) = \varepsilon E(t)$, is that

$$\frac{1}{\varepsilon\kappa(t)}\,\frac{d}{dt} F_\varepsilon(E(t)) = \mathrm{Re}\langle G_\varepsilon(E(t)), \dot E(t)\rangle \tag{1.26}$$

F-der-sys

with the (rescaled) gradient given by the rank-1 matrix

$$G_\varepsilon(E) = -rs^*. \tag{1.27}$$

grad-sys

Let

$$\psi_\varepsilon = \frac{\varepsilon}{1 - \varepsilon v^* D u}. \tag{1.28}$$

eq:psieps

For $E = uv^*$, using the formulas

$$\left(I - \varepsilon uv^* D\right)^{-1} = I + \psi_\varepsilon uv^* D, \qquad \left(I - \varepsilon Duv^*\right)^{-1} = I + \psi_\varepsilon Duv^*$$

we get (with $\beta = u^*b, \gamma = v^*c$)

$$\begin{aligned} r &= \left(I + \overline{\psi_\varepsilon} D^* vu^*\right) b = b + \overline{\psi_\varepsilon}\beta D^* v \\ s &= \left(I + \psi_\varepsilon Duv^*\right) c = c + \psi_\varepsilon \gamma D u. \end{aligned}$$

Norm-constrained gradient flow. In the same way as in Section II.1, this suggests to consider the following constrained gradient flow on the manifold of $p \times m$ complex matrices of unit Frobenius norm:

$$\dot{E} = -G_\varepsilon(E) + \mathrm{Re}\,\langle G_\varepsilon(E), E\rangle\, E, \tag{1.29}$$

ode-E-sys

where $(\lambda(t), x(t), y(t))$ is a rightmost eigentriple for the matrix $M(\varepsilon E(t))$. Assume the initial condition $E(0) = E_0$, a given matrix with unit Frobenius norm, chosen so that $M(\varepsilon E_0)$ has a unique rightmost eigenvalue $\lambda(0)$, which is simple.

We can now closely follow the programme of Section II.1 with straightforward minor adaptations.

Monotonicity. Assuming simple eigenvalues along the trajectory of (1.29), we again have the monotonicity property of Theorem II.1.4,

$$\frac{d}{dt} F_\varepsilon(E(t)) \le 0. \tag{1.30}$$

monotone-sys

Stationary points. Also the characterization of stationary points as given in Theorem II.1.5 extends with the same proof: Let $E \in \mathbb{C}^{p,m}$ with $\|E\|_F = 1$ be such that the rightmost eigenvalue λ of $M(\varepsilon E)$ is simple and $r, s \neq 0$. Then,

E is a stationary point of the differential equation (1.29) stat-sys
if and only if E is a real multiple of rs^*. (1.31)

stat-sys

Since local minima of F_ε are necessarily stationary points of the constrained gradient flow (1.29), this immediately implies the following.

cor:rank-1-sys

Corollary 1.11 (Rank of optimizers). *If E is an optimizer of problem* (1.18) *and we have $r, s \neq 0$, then E is of rank* 1.

As in Section II.1, Corollary 1.11 motivates us to project the differential equation (1.29) onto the manifold $\mathcal{M}_1$ of rank-1 matrices, which is computationally favourable.

VI.1.6 Rank-1 constrained gradient flow

sec:rank-1-sys

Since stationary points of (1.29) have rank 1, we consider - as we have done in Chapter II - the differential equation (1.29) projected onto the tangent space $T_E\mathcal{M}_1$ at E of the rank-1 manifold.

We recall that the orthogonal projection with respect to Frobenius inner product $\langle \cdot, \cdot \rangle$ from $\mathbb{C}^{p,m}$ onto the tangent space $T_E\mathcal{M}_1$ at $E = uv^* \in \mathcal{M}_1$ (with u and v of Euclidean norm 1) is given by Lemma II.1.11 as

$$P_E(Z) = Z - (I - uu^*)Z(I - vv^*) \quad \text{for } Z \in \mathbb{C}^{p,m}. \tag{1.32}$$

P-formula-sys

As in Section II.1.7 we consider the projected gradient system

$$\dot{E} = -P_E\Big(G_\varepsilon(E) + \mathrm{Re}\,\langle P_E(G_\varepsilon(E)), E\rangle\, E\Big) \tag{1.33}$$ $\boxed{\texttt{ode-E-1-sys}}$

and find the following properties, again by the arguments of Section II.1.7.

Conservation of unit norm. Solutions $E(t)$ of (1.33) have Frobenius norm 1 for all t, provided that the initial value $E(0)$ has Frobenius norm 1.

Differential equations for the two vectors. For an initial value $E(0) = u(0)v(0)^*$ with $u(0)$ and $v(0)$ of unit norm, the solution of (1.33) is given as $E(t) = u(t)v(t)^*$, where u and v solve the system of differential equations (for $G = G_\varepsilon(E) = -rs^*$)

$$\begin{array}{rcl} \dot{u} &=& -\frac{\mathrm{i}}{2}\,\mathrm{Im}(u^*Gv)u - (I - uu^*)Gv \\ \dot{v} &=& -\frac{\mathrm{i}}{2}\,\mathrm{Im}(v^*G^*u)v - (I - vv^*)G^*u, \end{array} \tag{1.34}$$ $\boxed{\texttt{ode-uv-sys}}$

which preserves $\|u(t)\| = \|v(t)\| = 1$ for all t. For its numerical integration we can again use the splitting method of Section II.1.8.

Monotonicity. Assuming simple eigenvalues almost everywhere along the trajectory of (1.33), we again have the monotonicity property of Theorem II.1.14,

$$\frac{d}{dt}F_\varepsilon(E(t)) \le 0. \tag{1.35}$$ $\boxed{\texttt{mon-sys}}$

Stationary points. Let $E \in \mathcal{M}_1$ be of unit Frobenius norm and assume that $P_E(rs^*) \neq 0$. If E is a stationary point of the projected differential equation (1.33), then E is already a stationary point of the differential equation (1.29).

VI.1.7 Approximating the $\mathcal{H}_\infty$-norm (outer iteration)

$\boxed{\texttt{bsec:hinf-cont}}$

We wish to compute $\|H\|_\infty$, using the characterization (1.12). We start by observing that since the spectral value set abscissa $\alpha_\varepsilon(A,B,C,D)$ is a monotonically increasing function of ε, we simply need to solve the equation

$$\alpha_\varepsilon(A,B,C,D) = 0 \tag{1.36}$$ $\boxed{\texttt{eq:fdef-sys}}$

for $\varepsilon > 0$. The first step is to characterize how α_ε depends on ε.

$\boxed{\texttt{hm:lambdaprime}}$ **Theorem 1.12.** *Let $\lambda(\varepsilon)$ denote the rightmost point of $\Lambda_\varepsilon(A,B,C,D)$ for $\varepsilon > 0$, $\varepsilon\|D\| < 1$, and assume that Assumption 1.7 holds for all such ε. Define $u(\varepsilon)$ and $v(\varepsilon)$ as right and left singular vectors with unit norm corresponding to ε^{-1}, the largest singular value of $H(\lambda(\varepsilon))$, and applying Theorem 1.3 with $\Delta(\varepsilon) = \varepsilon E(\varepsilon) = \varepsilon u(\varepsilon)v(\varepsilon)^*$, define $\widetilde{x}(\varepsilon)$ and $\widetilde{y}(\varepsilon)$ by* (1.10)*. Furthermore, assume that at ε, the rightmost point $\lambda(\varepsilon)$ is simple and unique. Then λ is continuously differentiable at ε and its derivative is real, with*

$$\frac{d}{d\varepsilon}\alpha_\varepsilon(A,B,C,D) = \frac{d}{d\varepsilon}\lambda(\varepsilon) = \frac{1}{\widetilde{x}(\varepsilon)^*\widetilde{y}(\varepsilon)} > 0. \tag{1.37}$$ $\boxed{\texttt{eq:deralpha}}$

Proof. For the purposes of differentiation, we identify $\lambda \in \mathbb{C}$ with $\zeta \in \mathbb{R}^2$ as in the proof of Lemma 1.8. The first part of Assumption 1.7 ensures that the largest singular value of $H(\lambda)$ is differentiable with respect to λ and that the singular vectors $v(\varepsilon)$ and $u(\varepsilon)$ are well defined up to multiplication of both by a unimodular scalar, and that $E(\varepsilon)$ is not only well defined but differentiable with respect to ε. The second part ensures that $\widetilde{x}(\varepsilon)^*\widetilde{y}(\varepsilon)$ is nonzero, using standard eigenvalue perturbation theory. As in the proof of Lemma 1.8, observe that

$$\frac{1}{\varepsilon} - \|C\left(\lambda(\varepsilon)I - A\right)^{-1}B + D\|_2 = 0$$

so differentiating this with respect to ε and using the chain rule yields

$$\frac{d\lambda(\varepsilon)}{d\varepsilon} = \frac{1}{v^*C(\lambda(\varepsilon)I - A)^{-2}Bu}.$$

Furthermore, (1.15) follows (for $\lambda = \lambda(\varepsilon)$) from (1.10). Combining these with the first-order optimality conditions for (1.13) in (1.14) gives the result. □

Corollary 1.13. *Make the same assumptions as in Theorem* 1.12*, except normalize* $x(\varepsilon)$ *and* $y(\varepsilon)$ *so that they fulfil* (1.5)*. This can be seen to be equivalent to scaling* $\widetilde{x}(\varepsilon)$ *and* $\widetilde{y}(\varepsilon)$ *by* $1/\beta(\varepsilon)$ *and* $1/\gamma(\varepsilon)$ *respectively where*

$$\beta(\varepsilon) = \frac{1 - \varepsilon u(\varepsilon)^* D^* v(\varepsilon)}{u(\varepsilon)^* b(\varepsilon)}, \qquad \gamma(\varepsilon) = \frac{1 - \varepsilon v(\varepsilon)^* D u(\varepsilon)}{v(\varepsilon)^* c(\varepsilon)}. \tag{1.38}$$

beta-gamma-eps

Hence

$$\frac{d}{d\varepsilon}\alpha_\varepsilon(A,B,C,D) = \frac{d}{d\varepsilon}\lambda(\varepsilon) = \frac{1}{\beta(\varepsilon)\gamma(\varepsilon)\big(x(\varepsilon)^*y(\varepsilon)\big)} \in \mathbb{R}^+. \tag{1.39}$$

eq:deralphasc

If A,B,C,D are all real, then $\Lambda_\varepsilon(A,B,C,D)$ is symmetric with respect to the real axis and hence its rightmost points must either be real or part of a conjugate pair. In the latter case, the assumption that $\lambda(\varepsilon)$ is unique does not hold but the result still holds if there is no third rightmost point.

The derivative formula (1.39) naturally leads to a formulation of Newton's method for computing $\|H\|_\infty$, similar to the one previously considered to compute stability radii in Section IV.2.3.

VI.1.8 Algorithm

We present the algorithm in concise form

Since Newton's method may not converge, it is standard practice to combine it with a bisection method to enforce convergence, that maintains an interval known to contain the root, bisecting when the Newton step is either outside the interval or does not yield a sufficient decrease in the absolute function value (in this case $|\alpha_{\varepsilon^j}(A,B,C,D)| = |\mathrm{Re}\ \lambda^j|$).

Algorithm 11: Basic algorithm to compute the H-infinity norm $\|H\|_\infty$

```
Data: Matrices $A, B, C, D$, initial vectors $u_0, v_0$, $\varepsilon_0 > 0$, tol a given positive tolerance
Result: $\varepsilon^J \approx \|H\|_\infty$
begin
    for $j = 0, \dots, j_{\max}$ do
        Approximate numerically $\alpha_{\varepsilon^j}(A, B, C, D)$ by integrating numerically (1.34) into
          a stationary point
        return rightmost point $\lambda^j$, and the associated left and right eigenvectors $x^j$, $y^j$ and
          corresponding scalars $\beta^j$, $\gamma^j$ defined as in (1.38), where $b = B^* x^j$ and $c = C y^j$
        if |Re $\lambda^j$| $< \mathrm{tol}$ then
            Set $J = j$
            Return $\|H\|_\infty \approx 1/\varepsilon^J$
        else
            Set
                $\varepsilon^{j+1} = \varepsilon^j - \big(\text{Re } \lambda^j\big)\beta^j\gamma^j\big((x^j)^* y^j\big).$
```

alg_Hinf

We emphasize, however, that this is still an idealized algorithm because there is no guarantee that the computed stationary point of the ODE will return the correct value of $\alpha_{\varepsilon^j}(A, B, C, D)$.

The algorithm is much faster than the standard Boyd-Balakrishnan-Bruinsma-Steinbuch algorithm to compute the $\mathcal{H}_\infty$-norm when $n \gg \max(m, p)$ and the matrix A is sparse.

VI.1.9 Numerical examples

VI.2 $\mathcal{H}_\infty$-distance to uncontrollability

sec:uncon

In this section we consider the operator nearness problem of finding the distance of a given controllable system to the nearest uncontrollable system, where the distance is taken as the $\mathcal{H}_\infty$-norm of the perturbation to the transfer function. This is a natural metric that measures the change in the input-output behaviour due to the perturbation of the system matrices. We propose and study a two-level algorithm that extends the algorithm of the preceding section.

VI.2.1 Uncontrollable system with nearest output

The linear time-invariant system (0.1) is *controllable* if and only if the $n \times (n+p)$ matrix

$$(A - \lambda I, B) \;\text{ has full row rank for all } \lambda \in \mathbb{C}. \qquad (2.1)$$

controllable

For this, it obviously suffices that the condition holds for the eigenvalues of A. Hence, a system is uncontrollable if and only if there exists an eigenvalue λ_A of A with corresponding left eigenvector x_A such that

$$x_A^* B = 0.$$

We consider the problem of finding the distance of a given controllable system (0.1) to the set of uncontrollable systems with the same matrices A, C and D, but with perturbed input-state matrix $\widetilde{B} = B + \Delta B$.

Here an important question arises: which distance? We might minimize the Frobenius norm of ΔB under the condition that the perturbed system becomes uncontrollable (which would yield $\Delta B = -x_A x_A^* B$ for a left eigenvector x_A to one of the eigenvalues of A), but this choice of a distance is not invariant under similarity transformations of A,

$$A \to V^{-1}AV, \quad B \to V^{-1}B, \quad C \to CV, \tag{2.2}$$

A-sim

which leave the matrix transfer function $H(\lambda) = C(\lambda I - A)^{-1}B + D$ invariant and hence also the input-output map $y = H(\partial_t)u$. We therefore measure the distance by the L^2 operator norm of the difference of the perturbed and unperturbed input-output operators, $\Delta H(\partial_t) = \widetilde{H}(\partial_t) - H(\partial_t)$. This L^2 operator norm is the $\mathcal{H}_\infty$-norm of the perturbation ΔH to the matrix transfer function. Considering *real* system matrices (A, B, C, D) and *real* perturbations ΔB, we arrive at the following.

Problem. *Given a controllable system* (0.1)*, find a perturbation $\Delta B \in \mathbb{R}^{n,p}$ such that the perturbed input-state matrix $B + \Delta B$ yields an uncontrollable system and such that the perturbation to the transfer function, $\Delta H(\lambda) = C(\lambda I - A)^{-1}\Delta B$ for* $\operatorname{Re}\lambda \geq 0$*, is of minimal $\mathcal{H}_\infty$-norm.*

The algorithm proposed below extends the two-level iterative method for the computation of the $\mathcal{H}_\infty$-norm given in the previous section.

VI.2.2 Two-level iteration

We use the following notation. For $\varepsilon > 0$ and for matrices $E \in \mathbb{C}^{p,m}$ and $R \in \mathbb{R}^{n,p}$, we define the perturbed state matrix

$$M_\varepsilon(E, R) = A + \varepsilon REC \in \mathbb{C}^{n,n}. \tag{2.3}$$

MER

This corresponds to (1.5) for $\Delta = \varepsilon E$ and for $R = \Delta B$ in the role of B and $D = 0$, which are the system matrices in $\Delta H(\lambda) = C(\lambda I - A)^{-1}R$.

We consider, for $\varepsilon > 0$ and $E \in \mathbb{C}^{p,m}$ with $\|E\|_F = 1$ and $R \in \mathbb{R}^{n,p}$ (we write R instead of ΔB from now on), the functional

$$F_\varepsilon(E, R) = -\operatorname{Re}\lambda(M_\varepsilon(E, R)), \tag{2.4}$$

FER

where $\lambda(M_\varepsilon(E, R))$ is a rightmost eigenvalue of $M_\varepsilon(E, R)$.

Let λ_A be an eigenvalue of A with normalized left eigenvector x_A. We require the uncontrollability condition $x_A^*(B + R) = 0$, i.e.,

$$x_A^* R = -x_A^* B. \tag{2.5}$$

R-cond

- **Inner iteration:** For a given $\varepsilon > 0$, solve the eigenvalue optimization problem, over $(E,R) \in \mathbb{C}^{p,m} \times \mathbb{R}^{n,p}$ with $\|E\|_F = 1$,

$$(E(\varepsilon), R(\varepsilon)) = \arg\max_R \min_E \, F_\varepsilon(E,R) \quad \text{subject to (2.5).} \tag{2.6}$$

eig-opt-uncon eig-opt-uncon

This constrained saddle point problem is different from the eigenvalue optimization problems considered so far. We will nevertheless use a gradient-based method with rank-1 matrices E that is very similar to the gradient flows of previous chapters.
With the spectral value set abscissa

$$\alpha_\varepsilon(R) := \alpha_\varepsilon(A,R,C,0) = \max_E \, \mathrm{Re}\, \lambda(M_\varepsilon(E,R)) = -\min_E F_\varepsilon(E,R),$$

(2.6) amounts to

$$R(\varepsilon) = \arg\min_R \alpha_\varepsilon(R).$$

We note that

$$F_\varepsilon(E(\varepsilon),R(\varepsilon)) = \max_R \min_E F_\varepsilon(E,R) = \min_E F_\varepsilon(E,R(\varepsilon)). \tag{2.7}$$

minmax-id

- **Outer iteration:** We compute $\varepsilon_\star$ as the smallest $\varepsilon > 0$ such that

$$\phi(\varepsilon) := \mathrm{Re}\, \lambda(M_\varepsilon(E(\varepsilon),R(\varepsilon))) = 0. \tag{2.8}$$

phi-uncon

We use a combined Newton-bisection method for this scalar equation.

thm:opt-uncon

Theorem 2.1 (**$\mathcal{H}_\infty$-distance).** *Let $\varepsilon_\star > 0$ be the exact solution of the problem* (2.5)–(2.8)*. Then, $B + R(\varepsilon_\star)$ is the perturbed input-state matrix that yields an uncontrollable system with minimal $\mathcal{H}_\infty$-distance between the transfer functions of the perturbed and unperturbed systems. This $\mathcal{H}_\infty$-distance equals $1/\varepsilon_\star$.*

Proof. Let $R_\star = R(\varepsilon_\star)$ and $(\Delta H)_\star(\lambda) = C(\lambda I - A)^{-1} R_\star$. By Theorem 1.6,

$$\|(\Delta H)_\star\|_\infty = \frac{1}{\varepsilon_\star}.$$

For an arbitrary R with (2.5), let $\varepsilon_R > 0$ be minimal with the property that $\alpha_{\varepsilon_R}(R) = 0$. By the definition of $\varepsilon_\star$ and by (2.7) we have

$$0 = \mathrm{Re}\, \lambda(M_{\varepsilon_\star}(E(\varepsilon_\star), R(\varepsilon_\star))) = \alpha_{\varepsilon_\star}(R_\star) \le \alpha_{\varepsilon_\star}(R),$$

and so we conclude that $\varepsilon_\star \ge \varepsilon_R$. Let $\Delta H(\lambda) = C(\lambda I - A)^{-1} R$. By Theorem 1.6,

$$\|(\Delta H)_\star\|_\infty = \frac{1}{\varepsilon_\star} \le \frac{1}{\varepsilon_R} = \|\Delta H\|_\infty,$$

which yields the result. □

VI.2.3 Constrained gradient flow

As before, we start with the free gradient of the functional.

`gradient-uncon` **Lemma 2.2 (Free gradient).** *Let $E(t) \in \mathbb{C}^{p,m}$ and $R(t) \in \mathbb{R}^{n,p}$, for real t near t_0, be continuously differentiable paths of matrices, with the derivatives denoted by $\dot{E}(t)$ and $\dot{R}(t)$. Assume that $\lambda(t)$ is a simple eigenvalue of $M_\varepsilon(E(t), R(t))$ depending continuously on t, with associated left and right eigenvectors $x(t)$ and $y(t)$ normalized by* (1.23)*, and let $\kappa(t) = 1/(x(t)^* y(t)) > 0$. Then, the derivative of $F_\varepsilon(E(t), R(t))$ is given by*

$$\frac{1}{\varepsilon\kappa(t)}\,\frac{d}{dt} F_\varepsilon(E(t), R(t)) = \operatorname{Re}\big\langle G_E(E(t), R(t)), \dot{E}(t)\big\rangle + \big\langle G_R(E(t), R(t)), \dot{R}(t)\big\rangle \tag{2.9}$$

`eq:deriv-pass`

with the (rescaled) gradient

$$\begin{aligned} G_E(E,R) &= -(R^T x)(Cy)^* \in \mathbb{C}^{p,m}, \\ G_R(E,R) &= -\operatorname{Re}\big(x(ECy)^*\big) \in \mathbb{R}^{n,p}. \end{aligned} \tag{2.10}$$

`eq:grad-uncon`

Proof. The proof is analogous to the proof of Lemma II.1.1. □

Choosing $\dot{E}$ such that it points in the direction of steepest admissible descent and $\dot{R}$ in the direction of steepest admissible ascent, we arrive, with appropriate signs, at the projected gradient system in which we take the constraints $\operatorname{Re}\langle E, \dot{E}\rangle = 0$ and $x_A^* \dot{R} = 0$ into account:

$$\begin{aligned} \dot{E} &= -G_E(E,R) + \operatorname{Re}\big\langle G_E(E,R), E\big\rangle E, \\ \dot{R} &= +(I - P_{x_A})\, G_R(E,R), \end{aligned} \tag{2.11}$$

`grad-sys-uncon`

where P_{x_A} is the orthogonal projection onto the range of $(\operatorname{Re} x_A, \operatorname{Im} x_A)$. With initial values (E_0, R_0) satisfying the constraints $\|E_0\|_F = 1$ and $x_A^*(B + R_0) = 0$, the constraints are then conserved for all t.

Remark 2.3 (Invariance). The differential equations (2.11) are invariant under transformations (2.2), as is checked by a straightforward lengthy calculation.

We cannot guarantee that all solutions of (2.11) converge to an optimum of (2.6) as $t \to \infty$. However, it can be shown by standard perturbation arguments that we get local convergence near a stationary point at which the partial Hessians of F_ε with respect to E and R are positive definite and negative definite, respectively.

At stationary points, we find again that $G_E(E,R)$ is a multiple of E. This gives us once again the rank-1 property of optimizers, as in Corollary 1.11.

`r:rank-1-uncon` **Corollary 2.4 (Rank of optimizers).** *If (E,R) is an optimizer of problem* (2.6) *and if $R^T x \neq 0$ and $Cy \neq 0$, then E is of rank* 1*.*

As in (1.33), we therefore search for E of rank 1, projecting the differential equation for E to the tangent space at E of the manifold of rank-1 matrices. Numerically we treat the resulting system of differential equations for the factors of $E = uv^*$ in the same way as in Section II.1.8.

VI.2.4 Outer iteration

For the outer iteration we again use a Newton method, which is justified under additional conditions, and we fall back to bisection when the Newton iteration does not work satisfactorily; cf. Section IV.2.3. Under the following assumption, the real function ϕ of (2.8) is differentiable and we will give a simple formula for its derivative.

ass:E-eps-uncon **Assumption 2.5.** For ε close to $\varepsilon_\star$ and $\varepsilon < \varepsilon_\star$, we assume the following for the optimizer $(E(\varepsilon), R(\varepsilon))$ of (2.6):

- The eigenvalue $\lambda(\varepsilon) = \lambda(M_\varepsilon(E(\varepsilon), R(\varepsilon)))$ is a simple eigenvalue.
- The map $\varepsilon \mapsto (E(\varepsilon), R(\varepsilon))$ is continuously differentiable.
- The partial gradient $G_E(\varepsilon) = G_E(E(\varepsilon), R(\varepsilon))$ is nonzero.

Theorem IV.2.2 then extends to the present situation. We again denote the eigenvalue condition number by

$$\kappa(\varepsilon) = \frac{1}{x(\varepsilon)^* y(\varepsilon)} > 0$$

with the corresponding left and right eigenvectors $x(\varepsilon), y(\varepsilon)$.

erivative-uncon **Theorem 2.6 (Derivative for the Newton iteration).** *Under Assumption 2.5, the function ϕ is continuously differentiable in a left neighbourhood of $\varepsilon_\star$ and its derivative is given as*

$$\phi'(\varepsilon) = -\kappa(\varepsilon)\, \|G_E(\varepsilon)\|_F < 0. \tag{2.12}$$ eq:dereps-uncon

Proof. By Lemma 4.1 we obtain, indicating by $'$ differentiation w.r.t. ε,

$$\frac{1}{\kappa(\varepsilon)} \frac{d}{d\varepsilon} F_\varepsilon(E(\varepsilon)) = \mathrm{Re}\big\langle G_E(\varepsilon), E(\varepsilon) + \varepsilon E'(\varepsilon)\big\rangle + \big\langle G_R(\varepsilon), R'(\varepsilon))\big\rangle.$$

As in the proof of Theorem 2.2 we find

$$\mathrm{Re}\big\langle G_E(\varepsilon), E(\varepsilon) + \varepsilon E'(\varepsilon)\big\rangle = -\|G_E(\varepsilon)\|_F.$$

In the stationary point $(E(\varepsilon), R(\varepsilon))$ of (2.11) we have $(I - P_{x_A})G_R(\varepsilon) = 0$, that is, $G_R(\varepsilon) = P_{x_A} G_R(\varepsilon)$. On the other hand, because of (2.5) we have $P_{x_A} R'(\varepsilon) = 0$. Hence,

$$\big\langle G_R(\varepsilon), R'(\varepsilon))\big\rangle = \big\langle P_{x_A} G_R(\varepsilon), R'(\varepsilon))\big\rangle = \big\langle G_R(\varepsilon), P_{x_A} R'(\varepsilon))\big\rangle = 0.$$

This yields the stated result. □

VI.2.5 Numerical example

VI.3 Nearest passive system

sec:pass

Consider the linear time-invariant system (0.1), which we here assume quadratic ($p = m$). The system is called *passive* if every input $u \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ and its corresponding output y satisfy the relation

$$\int_0^T y(t)^\top u(t)\,dt \geq 0 \qquad \text{for all } T > 0. \tag{3.1}$$

passive

Passivity is a fundamental property in control theory. When a given system is not passive, it is often required to enforce passivity by modifying it such that it becomes passive, yet remains 'near' the given system. One approach, as discussed by Grivet-Talocia & Gustavsen (**?**) and to be adopted here, is to perturb only the state-output matrix $C \in \mathbb{R}^{m,n}$ to $C + \varDelta C$ and to minimize a suitable norm of the perturbation such that the perturbed system is passive.

A favoured choice in the literature is to minimize the Frobenius norm of $\varDelta C\, L$, where $L \in \mathbb{R}^{n,n}$ is a Cholesky factor of the controllability Gramian $G_c = LL^T$, which is the unique solution of the Lyapunov equation $AG_c + G_cA^\top = BB^\top$. This Gramian is symmetric positive definite for a controllable system, as will be assumed in the following. The squared norm $\langle \varDelta C\, G_c, \varDelta C\rangle = \|\varDelta C\, L\|_F^2$ is invariant under transformations (2.2) of the system matrices, as is readily checked. Based on the characterization of passivity via Hamiltonian matrices, we present a two-level algorithm for this minimization problem in the Frobenius norm.

In a different direction that appears not to have been addressed in the literature before, we present in Subsection VI.3.5 an algorithm for the problem of enforcing passivity by perturbing C in such a way that the $\mathcal{H}_\infty$-norm of the difference between the transfer functions of the passive perturbed system and of the original system is minimized, similar to the approach in Section VI.2.

VI.3.1 Hamiltonian matrix related to passivity

By the Plancherel formula and the relation (1.2) of the matrix transfer function $H(\lambda)$, we have (with u extended by 0 outside [0,T])

$$\begin{aligned}\int_0^T y(t)^\top u(t)\,dt &= \mathrm{Re}\int_{\mathbb{R}} \mathcal{L}y(\mathrm{i}\omega)^* \mathcal{L}u(\mathrm{i}\omega)\,d\omega = \mathrm{Re}\int_{\mathbb{R}} \mathcal{L}\big(H(\mathrm{i}\omega)u(\mathrm{i}\omega)\big)^* \mathcal{L}u(\mathrm{i}\omega)\,d\omega \\ &= \frac{1}{2}\int_{\mathbb{R}} u(\mathrm{i}\omega)^* \big(H(\mathrm{i}\omega) + H(\mathrm{i}\omega)^*\big) u(\mathrm{i}\omega)\,d\omega,\end{aligned}$$

and so we find that passivity (3.1) is equivalent to the property of *positive realness* of the transfer function:

$$H(\mathrm{i}\omega) + H(\mathrm{i}\omega)^* \text{ is positive semi-definite for all } \omega \in \mathbb{R}. \tag{3.2}$$

pr

Since $H(\mathrm{i}\omega) \to D$ as $\omega \to \infty$, a necessary condition for (3.2) is that $D + D^\top$ is positive semi-definite.

Strict positive realness is defined in the same way, with ‘positive definite’ instead of ‘positive semi-definite’ for all $\omega \in \mathbb{R} \cup \{\infty\}$. A necessary condition is now that $D + D^\top$ is positive definite.

The following remarkable result can be found in Chapter 9 of the book by Grivet-Talocia & Gustavsen (**?**). It goes back to Boyd, Balakrishnan & Kabamba (**?**) and in its conceptual origins further back to Byers (1988). It characterizes strict positive realness, which is a condition on the transfer function on the whole imaginary axis, by the location of the eigenvalues of a single Hamiltonian matrix built from the system matrices.

hm:passive-ham

Theorem 3.1 (Passivity and eigenvalues of a Hamiltonian matrix). *The matrix transfer function $H(\cdot)$ is strictly positive real if and only if $D + D^\top$ is positive definite and the Hamiltonian matrix*

$$K = \begin{pmatrix} A & 0 \\ 0 & -A^T \end{pmatrix} - \begin{pmatrix} B \\ -C^\top \end{pmatrix} (D + D^\top)^{-1} \begin{pmatrix} C^\top \\ B \end{pmatrix}^\top \in \mathbb{R}^{2n \times 2n} \tag{3.3}$$

K-pass

has no eigenvalues on the imaginary axis.

In the following we will consider the Hamiltonian matrix also for systems with perturbed state-output matrices $C + \varDelta C$. To indicate the dependence on C, we write $K(C)$ for K of (3.3).

VI.3.2 Two-level iteration

Theorem 3.1 leads us to the following matrix nearness problem, which is reminiscent of Problem B in Section V.1. Here, the matrix $L \in \mathbb{R}^{n,n}$ is usually chosen as a Cholesky factor of the controllability Gramian, as mentioned above.

Problem (F). *Given a system* (0.1) *with positive definite matrix $D + D^\top$ for which the Hamiltonian matrix $K(C)$ has some purely imaginary eigenvalues, and given $\delta > 0$, compute a perturbed state-output matrix $C + \varDelta C$ with minimal $\|\varDelta C\, L\|_F$ such that all eigenvalues of $K(C + \varDelta C)$ have a real part of absolute value at least δ.*

The proposed algorithm is a two-level iterative method similar to the second method of Section V.1.5.

- **Inner iteration:** For a given $\varepsilon > 0$, we use a (low-rank) gradient system to solve the eigenvalue optimization problem, over $E \in \mathbb{R}^{p \times n}$ with $\|E\|_F = 1$,

$$E(\varepsilon) = \arg \max_{\|E\|_F = 1} \operatorname{Re} \lambda\big(K(C + \varepsilon E L^{-1})\big), \tag{3.4}$$

eig-opt-pass

  where $\lambda(K)$ is an eigenvalue of minimal nonnegative real part (chosen with the largest nonnegative imaginary part) of a real Hamiltonian matrix K.

- **Outer iteration:** We compute the smallest ε such that

$$\phi(\varepsilon) := \operatorname{Re} \lambda\bigl(K(C + \varepsilon E(\varepsilon)L^{-1})\bigr) = \delta \tag{3.5}$$

phi-pass

for the given small threshold $\delta > 0$. This uses a mixed Newton/bisection method and, for very small δ, the asymptotic square-root behavior $\phi(\varepsilon) \sim \sqrt{\varepsilon - \varepsilon_\star}$ as the eigenvalue tends to the imaginary axis at perturbation size $\varepsilon_\star$; see Theorem V.1.7.

VI.3.3 Norm- and rank-constrained gradient flows

dientflows-pass

In this subsection we show how to deal with the inner iteration, once again following and adapting the by now well-trodden path of Chapter II, here in the real version of Section II.2.

In the resulting algorithm we do not move eigenvalues of Hamiltonian matrices on the imaginary axis. Instead, like in the second algorithm of Section V.1.5, we work with Hamiltonian matrices all whose eigenvalues are off the imaginary axis, corresponding to perturbed matrices C that yield a passive system. We move eigenvalues with smallest positive real part toward the imaginary axis, starting from a non-optimal passive perturbation of the original, non-passive system. This starting perturbation can come from a computationally inexpensive but non-optimal passivity enforcement algorithm.

Free gradient. The following lemma will allow us to compute the steepest descent direction of the functional $F_\varepsilon(E) = -\operatorname{Re} \lambda\bigl(K(C + \varepsilon E L^{-1})\bigr)$.

n:gradient-pass

Lemma 3.2 (Real gradient). *Let $E(t) \in \mathbb{R}^{m,n}$, for real t near t_0, be a continuously differentiable path of matrices, with the derivative denoted by $\dot E(t)$. Assume that $\lambda(t)$ is a simple eigenvalue of $K(C + \varepsilon E(t)L^{-1})\bigr)$ depending continuously on t, with associated eigenvectors $x(t)$ and $y(t)$ normalized by* (1.23)*, and let $\kappa(t) = 1/(x(t)^* y(t)) > 0$. Then, the derivative of $F_\varepsilon(E(t)) = \operatorname{Re} \lambda\bigl(K(C + \varepsilon E(t)L^{-1})\bigr)$ is given by*

$$\frac{1}{\varepsilon\kappa(t)} \frac{d}{dt} F_\varepsilon(E(t)) = \bigl\langle G_\varepsilon(E(t)), \dot E(t) \bigr\rangle \tag{3.6}$$

eq:deriv-pass

with the (rescaled) real gradient

$$G_\varepsilon(E) = K'(C + \varepsilon E L^{-1})^* [\operatorname{Re}(xy^*)] L^{-\top} \in \mathbb{R}^{m,n}, \tag{3.7}$$

eq:grad-pass

where $K'(C)^* : \mathbb{R}^{2n,2n} \to \mathbb{R}^{m,n}$ is the adjoint of the derivative $K'(C) : \mathbb{R}^{m,n} \to \mathbb{R}^{2n,2n}$ defined by $\langle K'(C)^*[W], Z\rangle = \langle W, K'(C)[Z]\rangle$ for all $W \in \mathbb{R}^{2n,2n}$ and $Z \in \mathbb{R}^{m,n}$.

Proof. By the derivative formula for simple eigenvalues and the chain rule, we have (omitting the ubiquitous argument t after the first equality sign)

$$\begin{aligned}&\frac{d}{dt} \operatorname{Re} \lambda\bigl(K(C + \varepsilon E(t)L^{-1})\bigr) = \kappa \operatorname{Re}\Bigl(x^* K'(C + \varepsilon E L^{-1})[\varepsilon \dot E L^{-1}] y\Bigr) \\ &= \kappa \bigl\langle \operatorname{Re}(xy^*), K'(C + \varepsilon E L^{-1})[\varepsilon \dot E L^{-1}] \bigr\rangle = \kappa\varepsilon \bigl\langle K'(C + \varepsilon E L^{-1})^* [\operatorname{Re}(xy^*)] L^{-\top}, \dot E \bigr\rangle,\end{aligned}$$

which yields the stated result. □

An explicit formula for the adjoint of the derivative of K is given next.

lem:K-der-adj **Lemma 3.3 (Adjoint of the derivative).** *For* $W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$ *partitioned according to the* $n \times n$ *blocks of* $K(C)$*, we have (with* $T = D + D^\top$ *for short)*

$$K'(C)^*[W] = -T^{-1}B^\top (W_{11} - W_{22}^\top) + T^{-1}C\,(W_{21} + W_{21}^\top). \tag{3.8}$$ K-der-adj

rem:rank-pass **Remark 3.4.** A noteworthy consequence of (3.8) is that $G_\varepsilon(E)$ of (3.7) has rank at most 8. (The rank is at most 4 for real eigenvalues.)

Proof. For any path $C(t)$ we have $\dot K(t) = \frac{d}{dt}K(C(t)) = K'(C(t))[\dot C(t)]$ and hence

$$\langle K'(C)^*[W], \dot C\rangle = \langle W, K'(C)[\dot C]\rangle = \langle W, \dot K\rangle.$$

Differentiation of (3.3) yields

$$\dot K = -\begin{pmatrix} 0 \\ -\dot C^\top \end{pmatrix} T^{-1} \begin{pmatrix} C^\top \\ B \end{pmatrix}^\top - \begin{pmatrix} B \\ -C^\top \end{pmatrix} T^{-1} \begin{pmatrix} \dot C^\top \\ 0 \end{pmatrix}^\top .$$

We note that

$$\begin{aligned} &\left\langle W, \begin{pmatrix} 0 \\ -\dot C^\top \end{pmatrix} T^{-1} \begin{pmatrix} C^\top \\ B \end{pmatrix}^\top \right\rangle = \left\langle W \begin{pmatrix} C^\top \\ B \end{pmatrix} T^{-1}, \begin{pmatrix} 0 \\ -\dot C^\top \end{pmatrix} \right\rangle \\ &\quad = -\big\langle (W_{21}C^\top + W_{22}B)T^{-1}, \dot C^\top \big\rangle = -\big\langle ((W_{21}C^\top + W_{22}B)T^{-1})^\top, \dot C \big\rangle \\ &\left\langle W, \begin{pmatrix} B \\ -C^\top \end{pmatrix} T^{-1} \begin{pmatrix} \dot C^\top \\ 0 \end{pmatrix}^\top \right\rangle = \langle T^{-1}(B^\top, -C)W, (\dot C, 0)\rangle \\ &\quad = \big\langle T^{-1}(B^\top W_{11} - CW_{21}), \dot C \big\rangle \end{aligned}$$

so that finally

$$\langle K'(C)^*[W], \dot C\rangle = \langle W, \dot K\rangle = \langle -T^{-1}B^\top (W_{11} - W_{22}^\top) + T^{-1}C\,2\,\mathrm{Sym}(W_{21}), \dot C\rangle$$

for all $\dot C \in \mathbb{R}^{m,n}$. This yields $K'(C)^*[W]$ as stated. □

Norm-constrained gradient flow. As in Section II.2, we consider the projected gradient flow on the manifold of $m \times n$ real matrices of unit Frobenius norm:

$$\dot E = -G_\varepsilon(E) + \langle G_\varepsilon(E), E\rangle\, E, \tag{3.9}$$ ode-E-pass

where $G_\varepsilon(E)$ is defined by (3.7) via an eigentriple (λ, x, y) of the Hamiltonian matrix $K(C + \varepsilon E L^{-1})$ with λ the target eigenvalue of minimal nonnegative real part (and among these, the one with largest nonnegative imaginary part). The Frobenius norm 1 is conserved along trajectories.

We now follow closely the programme of Section II.2: We again have the monotonic decay of $F_\varepsilon(E(t))$ as in (II.2.7), and the characterization of stationary points as given in Theorem II.1.5 also extends: Let $E \in \mathbb{C}^{m,n}$ with $\|E\|_F = 1$ be such that the target eigenvalue λ of $K(C + \varepsilon E L^{-1})$ is simple and $G_\varepsilon(E) \neq 0$. Then, E is a stationary point of the differential equation (3.9) if and only if E is a real multiple of $G_\varepsilon(E)$. Together with Remark 3.4, this implies the following.

cor:rank-8-pass

Corollary 3.5 (Rank of optimizers). *If E is an optimizer of the eigenvalue optimization problem* (3.11) *and if $G_\varepsilon(E) \neq 0$, then E is of rank at most* 8.

As in Section II.2, this motivates us to constrain the differential equation (3.9) to a manifold of real low-rank matrices, which turns out to be computationally favourable for large systems.

Rank-8 constrained gradient flow. In the same way as in Section II.2.4, this time with rank $r = 8$, we orthogonally project the right-hand side of (3.9) onto the tangent space at E of the manifold $\mathcal{M}_r \subset \mathbb{R}^{m,n}$ of real rank-r matrices, so that solutions starting with rank r retain the rank r:

$$\dot{E} = P_E\Big(-G_\varepsilon(E) + \langle G_\varepsilon(E), E\rangle E\Big). \qquad (3.10)$$

ode-ErF-pass

Then also the Frobenius norm 1 is conserved (see (II.2.15)), and $F_\varepsilon(E(t))$ decays monotonically (see (II.2.17)).

Using the SVD-like factorization $E = USV^\top$, where $U \in \mathbb{R}^{m,r}$ and $V \in \mathbb{R}^{n,r}$ have orthonormal columns and $S \in \mathbb{R}^{r,r}$, the seemingly abstract differential equation (3.10) is solved numerically for the factors U, V, S as described in Section II.2.5.

Using the low-rank structure in the eigenvalue computation. For the computation of the gradient matrix $G_\varepsilon(E)$, one needs to compute the eigenvalue of smallest positive real part and the associated left and right eigenvectors of the Hamiltonian matrix $K(C + \varepsilon E L^{-1})$. Except in the very first step of the algorithm, one can make use of the eigenvalue of smallest real part of the previous step in an inverse iteration (and possibly of the eigenvalues of second and third smallest real part etc. to account for a possible exchange of the leading eigenvalue).

Moreover, we get from a perturbation $\varepsilon E L^{-1}$ with $E = U\varSigma V^\top$ of rank 8 that C is perturbed by $\varDelta C = \varepsilon E L^{-1} = \varepsilon (U\varSigma)(L^{-\top} V)^\top$ of the same rank 8, which yields the perturbed Hamiltonian matrix

$$K(C + \varDelta C) = K(C) + \varDelta K,$$

where the perturbation $\varDelta K$ is still of moderate rank in view of (3.3). This fact can be used in the computation of the required eigenvalues in the case of a high-dimensional system, using the Sherman–Morrison–Woodbury formula in an inverse iteration.

If $p \ll n$, then $K(C)$ can be viewed as a low-rank perturbation to the matrix

$$\begin{pmatrix} A & 0 \\ 0 & -A^\top \end{pmatrix}.$$

With the Sherman–Morrison–Woodbury formula, this can yield an efficient inverse iteration when A is a large and sparse matrix for which shifted linear systems can be solved efficiently.

VI.3.4 Outer iteration

For the solution of the scalar nonlinear equation (3.5) we use a mixed Newton-bisection method as in Section IV.2.3 or, for small δ, the square root model and bisection as in Section V.1.4.

Numerical results obtained with the above method for passivity enforcement are given by Fazzi, Guglielmi & Lubich (**?**).

VI.3.5 $\mathcal{H}_\infty$-nearest passive system

ec:pass-Hinfty

We extend the approach of Section VI.2 to the problem of $\mathcal{H}_\infty$-optimal passivity enforcement.

Problem ($\mathcal{H}_\infty$). *Given a system* (0.1) *with positive definite matrix $D + D^\top$ for which the Hamiltonian matrix $K(C)$ has some purely imaginary eigenvalues, and given $\delta > 0$, compute a perturbed state-output matrix $C + \Delta C$ with minimal $\mathcal{H}_\infty$-norm of the perturbation to the transfer function, $\Delta H(\lambda) = \Delta C(\lambda I - A)^{-1}B$, such that all eigenvalues of $K(C + \Delta C)$ have a real part of absolute value at least δ.*

We write S for ΔC and use the functional

$$F^K(S) = -\mathrm{Re}\,\lambda(K(C+S)),$$

where, as before, $\lambda(K)$ is an eigenvalue of minimal nonnegative real part (chosen with the largest nonnegative imaginary part) of a Hamiltonian matrix $K \in \mathbb{R}^{2n,2n}$.

For $\varepsilon > 0$ and for matrices $E \in \mathbb{C}^{m,m}$ with $\|E\|_F = 1$ and $S \in \mathbb{C}^{m,n}$, we define the perturbed state matrix, cf. (2.3),

$$M_\varepsilon(E,S) = A + \varepsilon BES \in \mathbb{C}^{n,n}.$$

We use the functional

$$F_\varepsilon^M(E,S) = -\mathrm{Re}\,\lambda(M_\varepsilon(E,S)),$$

where now (with slight abuse of notation) $\lambda(M)$ is a rightmost eigenvalue of a matrix $M \in \mathbb{C}^{n,n}$. With this functional we propose a two-level approach similar to Section VI.2.

- **Inner iteration:** For given $\delta > 0$ and $\varepsilon > 0$, we use a constrained gradient system (with E of rank 1) to solve the constrained eigenvalue optimization problem, over $(E,S) \in \mathbb{C}^{m,m} \times \mathbb{R}^{m,n}$ with $\|E\|_F = 1$,

$$(E(\varepsilon), S(\varepsilon)) = \arg\max_S \min_E F_\varepsilon^M(E,S) \qquad \text{subject to} \quad -F^K(S) \geq \delta. \qquad (3.11)$$

eig-opt-pass

– **Outer iteration:** We compute $\varepsilon_\star$ as the smallest $\varepsilon > 0$ such that

$$\phi(\varepsilon) := \mathrm{Re}\,\lambda(M_\varepsilon(E(\varepsilon), S(\varepsilon))) = 0. \tag{3.12}$$ phi-pass-2

We again use a combined Newton-bisection method for this scalar equation.

We then have the following analogue of Theorem 2.1, which is proved by the same argument based on Theorem 1.6.

thm:opt-pass **Theorem 3.6** **($\mathcal{H}_\infty$-distance).** *Let $\varepsilon_\star > 0$ be the exact solution of the problem* (3.11)–(3.12)*. Then, the perturbed state-output matrix $C + S(\varepsilon_\star)$ yields a passive system having $\mathrm{Re}\,\lambda(K(C+S)) \geq \delta$ with minimal $\mathcal{H}_\infty$-distance between the transfer functions of the perturbed and unperturbed systems. This $\mathcal{H}_\infty$-distance equals $1/\varepsilon_\star$.*

We formulate a gradient system for solving the constrained max-min eigenvalue optimization problem (3.11). Let $G_E^M(E,S)$, $G_S^M(E,S)$ and $G^K(S)$ be the gradients of F_ε^M and F^K obtained as in Lemmas 4.1 and 3.2, respectively. Choosing $\dot{E}$ and $\dot{S}$ in the directions of steepest admissible descent and ascent, respectively, yields the projected gradient system

$$\begin{aligned}\dot{E} &= -G_E^M(E,S) + \mathrm{Re}\langle G_E^M(E,S), E\rangle E,\\ \dot{S} &= +G_S^M(E,S) - \mu\, G_S^K(S),\end{aligned}$$

where the Lagrange multiplier μ equals 0 if $-F^K(S) > \delta$ or if $-F^K(S) = \delta$ and $-\langle G^K(S), G_S^M(E,S)\rangle \geq 0$, and else $\mu > 0$ is determined from the condition $\langle G^K(S), \dot{S}\rangle = 0$, i.e., $\mu = \langle G^K(S), G_S^M(E,S)\rangle \,/\, \|G_S^K\|_F^2$. As previously, E can be further constrained to be of rank 1.

VI.4 Structured contractivity radius

A linear time-invariant system (0.1) is called *contractive* if its transfer function H is bounded by

$$\|H\|_\infty \leq 1,$$

and it is called *strictly contractive* if the above inequality is strict. Contractive systems play an important role as subsystems in large networks, because their composition remains contractive and thus yields well-controlled input-output relations. A strictly contractive system may be susceptible to perturbations (or uncertainties) in the entries of its matrices (A,B,C,D), and it is then of interest to know which size of perturbations still guarantees contractivity. Here, we consider perturbations only in the state matrix A and allow for structured perturbations $\Theta \in \mathcal{S}$, where the structure space $\mathcal{S} \subset \mathbb{C}^{n,n}$ is a given complex- or real-linear subspace, e.g. real matrices with a prescribed sparsity pattern. We study the following problem in this section.

Let (A, B, C, D) be the system matrices of a strictly contractive linear time-invariant system. In particular, this implies $\|D\|_2 < 1$. We consider structured perturbations $A \to A + \Theta$ with $\Theta \in \mathcal{S}$, which yield perturbed transfer functions

$$H_\Theta(\lambda) = C(\lambda I - A - \Theta)^{-1} B + D.$$

Problem. *Find the largest possible perturbation size $\theta_\star > 0$ such that*

$$\|H_\Theta\|_\infty \le 1 \quad \textit{for all } \Theta \in \mathcal{S} \textit{ with } \|\Theta\|_F \le \theta_\star .$$

The number $\theta_\star > 0$ measures the robustness of contractivity of a system and is called the $\mathcal{S}$-structured *contractivity radius*. We present and discuss a two-level algorithm that is closely related to that of Section V.5 with $\varepsilon = 1$, to which it reduces for the special case $B = C = I$ and $D = 0$.

VI.4.1 Two-level iteration

We consider the matrix $M(\Delta)$ of (1.5) that corresponds to $A + \Theta$ instead of A. So we let

$$M(\Theta, \Delta) = A + \Theta + B\Delta(I - D\Delta)^{-1} C.$$

By Theorem 1.2, $\|H_\Theta\|_\infty = 1$ if and only if there exists $\Delta \in \mathbb{C}^{p,m}$ with $\|\Delta\|_2 = 1$ such that $M(\Theta, \Delta)$ has an eigenvalue with nonnegative real part. Moreover, the optimizing matrix Δ is of rank 1, and hence its Frobenius and 2-norms are the same. We define the functional $F_\theta(E^{\mathcal{S}}, E)$ (for $E^{\mathcal{S}} \in \mathcal{S}$ and $E \in \mathbb{C}^{p,m}$, both of unit Frobenius norm) by

$$F_\theta(E^{\mathcal{S}}, E) = -\mathrm{Re}\,\lambda\big(M(\theta E^{\mathcal{S}}, E)\big), \qquad (4.1) \qquad \boxed{\texttt{F-eps-con}}$$

where $\lambda(M)$ is the eigenvalue of M of largest real part (and among those, the one with largest imaginary part). With this functional we follow the two-level approach of Section IV.2:

- **Inner iteration:** For a given $\theta > 0$, we aim to compute matrices $E^{\mathcal{S}}(\theta) \in \mathcal{S}$ and $E(\theta) \in \mathbb{C}^{p,m}$, both of unit Frobenius norm, that minimize F_θ:

$$(E^{\mathcal{S}}(\theta), E(\theta)) = \arg \min_{\substack{E^{\mathcal{S}} \in \mathcal{S}, E \in \mathbb{C}^{p,m} \\ \|\Delta\|_F = \|E\|_F = 1}} F_\theta(E^{\mathcal{S}}, E). \qquad (4.2) \qquad \boxed{\texttt{E-theta}}$$

- **Outer iteration:** We compute the smallest positive value $\theta_\star$ with

$$\phi(\theta_\star) = 0, \qquad (4.3) \qquad \boxed{\texttt{zero-delta}}$$

  where $\phi(\theta) = F_\theta(E^{\mathcal{S}}(\theta), E(\theta)) = \alpha_\varepsilon(A + \theta E^{\mathcal{S}}(\theta), B, C, D)$ for $\varepsilon = 1$.

Provided that these computations succeed, we have from Theorem 1.6 that $\Theta_\star = \theta_\star E^{\mathcal{S}}(\theta_\star) \in \mathcal{S}$ is a perturbation matrix with $\|H_{\Theta_\star}\|_\infty = 1$, and $\theta_\star$ is the $\mathcal{S}$-structured contractivity radius.

VI.4.2 Rank-1 matrix differential equations for the inner iteration

The following lemma is obtained as in the proof of Lemma 1.10.

gradient-uncon

Lemma 4.1 (Structured gradient). *Let $E^{\mathcal{S}}(t) \in \mathcal{S}$ and $E(t) \in \mathbb{C}^{p,m}$, for real t near t_0, be continuously differentiable paths of matrices. Assume that $\lambda(t)$ is a simple eigenvalue of $M(\theta E^{\mathcal{S}}(t), E(t))$ depending continuously on t, with associated left and right eigenvectors $x(t)$ and $y(t)$ normalized by* (1.23)*, and let $\kappa(t) = 1/(x(t)^* y(t)) > 0$. Then, the derivative of $F_\theta(E^{\mathcal{S}}(t), E(t))$ is given by*

$$\frac{1}{\kappa(t)} \frac{d}{dt} F_\theta(E^{\mathcal{S}}(t), E(t)) = \text{Re}\left\langle G^{\mathcal{S}}_{\Theta}(\theta E^{\mathcal{S}}(t), E(t)), \theta \dot{E}^{\mathcal{S}}(t) \right\rangle + \text{Re}\left\langle G_{\Delta}(\theta E^{\mathcal{S}}(t), E(t)), \dot{E}(t) \right\rangle \tag{4.4}$$

eq:deriv-pass

with the (rescaled) gradient

$$\begin{aligned} G^{\mathcal{S}}_{\Theta}(\Theta, \Delta) &= \Pi^{\mathcal{S}}(xy^*) \in \mathcal{S}, \\ G_{\Delta}(\Theta, \Delta) &= rs^* \in \mathbb{C}^{p,m}, \end{aligned} \tag{4.5}$$

eq:grad-uncon

where r, s are obtained from x, y via (1.24)*.*

Norm- and structure-constrained gradient flow. Similar to Section V.5, we consider the projected gradient flow, with $G^{\mathcal{S}}_{\Theta} = G^{\mathcal{S}}_{\Theta}(\theta E^{\mathcal{S}}, E)$ and $G_\Delta = G_\Delta(\theta E^{\mathcal{S}}, E)$ for short,

$$\begin{aligned} \theta \dot{E}^{\mathcal{S}} &= -G^{\mathcal{S}}_{\Theta} + \text{Re}\left\langle G^{\mathcal{S}}_{\Theta}, E^{\mathcal{S}} \right\rangle E^{\mathcal{S}}, \\ \dot{E} &= -G_\Delta + \text{Re}\left\langle G_\Delta, E \right\rangle E. \end{aligned} \tag{4.6}$$

ode-ES-E-contr

The unit Frobenius norm of $E^{\mathcal{S}}$ and E is conserved along trajectories and the functional $F_\theta(E^{\mathcal{S}}(t), E(t))$ decreases monotonically. At a non-degenerate stationary point $(E^{\mathcal{S}}, E)$, where $G^{\mathcal{S}}_{\Theta}$ and G_Δ do not vanish, we find that $E^{\mathcal{S}}$ and E are real multiples of $G^{\mathcal{S}}_{\Theta}$ and G_Δ, respectively. Hence, $E^{\mathcal{S}}$ is the projection onto $\mathcal{S}$ of a rank-1 matrix, and E is a rank-1 matrix.

Rank-1 matrix differential equations. To make use of the rank-1 structure of optimizers, we proceed as in Section V.5.2 and combine the rank-1 approaches of Sections II.1 and II.3. We consider differential equations for rank-1 matrices $Y(t)$ and $E(t)$, where the former yields $E^{\mathcal{S}}(t) = \Pi^{\mathcal{S}} Y(t)$. These differential equations are obtained from (5.7) by replacing $G^{\mathcal{S}}_{\Theta}$ and G_Δ by their projections P_Y and P_E onto the tangent spaces of the manifold of rank-1 matrices at Y and E, respectively:

$$\begin{aligned} \theta\, \dot{Y} &= -P_Y G^{\mathcal{S}}_{\Theta} + \text{Re}\,\langle P_Y G^{\mathcal{S}}_{\Theta}, E^{\mathcal{S}} \rangle Y \quad \text{with } E^{\mathcal{S}} = \Pi^{\mathcal{S}} Y, \\ \dot{E} &= -P_E G_\Delta + \text{Re}\,\langle G_\Delta, E \rangle E. \end{aligned} \tag{4.7}$$

ode-ES-E-1-contr

These differential equations yield rank-1 matrices $Y(t)$ and $E(t)$ and preserve the unit Frobenius norm of $E^{\mathcal{S}}(t)$ and $E(t)$. As in Sections II.1 and II.3 it is shown that under a nondegeneracy condition, the stationary points (Y, E) of (4.7) correspond bijectively to the stationary points $(E^{\mathcal{S}}, E)$ of (4.6) via $E^{\mathcal{S}} = \Pi^{\mathcal{S}} Y$ and with the same E. The differential equations are integrated numerically into a stationary point $(E^{\mathcal{S}}, E)$ as is done in Sections II.1 and II.3, working with the vectors that define the rank-1 matrices Y and E and advancing them in time with a suitable splitting method.

VI.4.3 Outer iteration, updating $\boldsymbol{\theta}$

For the solution of the scalar equation $\phi(\theta) = 0$ we use a combined Newton / bisection method as in Section IV.2. The derivative of ϕ for the Newton iteration is obtained with the arguments of the proof of Theorem IV.2.2 (under analogous assumptions), which yields

$$\phi'(\theta) = -\kappa(\theta)\, \| G_{\Theta}^{\mathcal{S}}(\theta E^{\mathcal{S}}(\theta), E(\theta)) \|_F = -\kappa(\theta)\, \| \Pi^{\mathcal{S}}(x(\theta) y(\theta)^*) \|_F,$$

where $x(\theta)$ and $y(\theta)$ are left and right normalized eigenvectors associated with the rightmost eigenvalue of $M(\theta E^{\mathcal{S}}(\theta), E(\theta))$, and $\kappa(\theta) = 1/(x(\theta)^* y(\theta)) > 0$.

VI.5 Descriptor systems

sec:descriptor

We consider a *descriptor system*, which formally differs from the system (0.1) only in that the derivative of the state vector is multiplied with a singular matrix[1] $E \in \mathbb{R}^{n,n}$: for $t \geq 0$,

$$\begin{aligned} E\dot z(t) &= Az(t) + Bu(t) \\ y(t) &= Cz(t). \end{aligned} \tag{5.1}$$

descriptor

We choose zero initial values and assume that the input function u can be extended by zero to a sufficiently differentiable function on the whole real axis. Since E is singular, the equation for the state vector z is now a differential-algebraic equation instead of a pure differential equation. Descriptor systems arise naturally in systems with state constraints and in modelling and composing networks of such systems.

We assume that all finite eigenvalues of the matrix pencil $A - \lambda E$ have negative real part, i.e.,

$$A - \lambda E \;\text{ is invertible for all } \lambda \in \mathbb{C} \text{ with } \mathrm{Re}\,\lambda \geq 0. \tag{5.2}$$

pencil-ass

In particular, the matrix A is invertible.

The matrix transfer function of the descriptor system is

$$H(\lambda) = C(\lambda E - A)^{-1} B, \qquad \mathrm{Re}\,\lambda \geq 0. \tag{5.3}$$

H-desc

[1] In this section only we adhere to the convention in the control literature to denote E the matrix multiplying the time derivative of the state vector and to work with the matrix pencil (A, E). In the rest of this book E appears as a matrix of unit Frobenius norm when writing a perturbation matrix as $\Delta = \varepsilon E$. In this section we will write instead $\Delta = \varepsilon Z$ with Z of Frobenius norm 1, choosing Z as the letter of last resort.

Remark 5.1. In the equation for the output y in (5.1) we have set the feedthrough matrix $D = 0$ for convenience. The term $Du(t)$ could be added for nonzero D. The required changes in the theory and algorithm of this section can be done by combining the constructions and arguments of Section VI.1 with those given here. As we wish to concentrate on the effects of the singular matrix E, we chose to forego the technical complications resulting from a nonzero feedthrough matrix D, which were already dealt with in Section VI.1.

VI.5.1 Index and asymptotics of the transfer function at infinity

In contrast to (0.1), the transfer function $H(\lambda)$ of (5.3) need not be uniformly bounded for $\operatorname{Re}\lambda \geq 0$. We show that, aside from exceptional choices of B and C, the norm of $H(\lambda)$ grows proportionally to $|\lambda|^{k-1}$ as $\lambda \to \infty$, where $k \geq 1$ is the *index* of the differential-algebraic equation $E\dot{z} = Az + f$. The index can be determined from the Schur normal form of $A^{-1}E$ as follows.

We premultiply (5.1) with A^{-1} and λ^{-1} so that the Laplace-transformed state equation $(\lambda E - A)\mathcal{L}z(\lambda) = B\,\mathcal{L}u(\lambda)$ becomes

$$\Bigl(A^{-1}E - \frac{1}{\lambda}\,I\Bigr)\mathcal{L}z(\lambda) = \frac{1}{\lambda}\,A^{-1}B\,\mathcal{L}u(\lambda).$$

We want to understand the behaviour of the inverse of the matrix in brackets on the left-hand side as $\lambda \to \infty$, which is not obvious as E is singular. To this end we transform to a block Schur normal form

$$A^{-1}E = Q\begin{pmatrix} G & K \\ 0 & N \end{pmatrix} Q^\top \tag{5.4}$$

block-schur-dae

with an orthogonal matrix Q, an invertible matrix G and a nilpotent matrix N. The smallest integer $k \geq 1$ such that

$$N^k = 0$$

is called the *index* of the matrix pencil (A, E) (or of the differential-algebraic equation $E\dot{z} = Az + f$, or of the descriptor system (5.1)).

We have, for $\operatorname{Re}\lambda \geq 0$ and $\zeta = 1/\lambda$,

$$(A^{-1}E - \zeta I)^{-1} = Q\begin{pmatrix} (G-\zeta I)^{-1} & -(G-\zeta I)^{-1}K(N-\zeta I)^{-1} \\ 0 & (N-\zeta I)^{-1} \end{pmatrix} Q^\top.$$

Here we note that

$$-\lambda^{-1}(N - \lambda^{-1}I)^{-1} = (I - \lambda N)^{-1} = I + \lambda N + \ldots + \lambda^{k-1}N^{k-1}.$$

We conclude that the norm of $H(\lambda) = C(A^{-1}E - \lambda^{-1}I)^{-1}\lambda^{-1}A^{-1}B$ is bounded by a constant times $|\lambda|^{k-1}$ as $\lambda \to \infty$, and for generic B and C the asymptotic growth is actually proportional to $|\lambda|^{k-1}$.

VI.5.2 Weighted matrix transfer function and its $\mathcal{H}_\infty$-norm

For a system of index k, we therefore want to bound the *weighted* matrix transfer function

$$H^{[k]}(\lambda) = (1+\lambda)^{-(k-1)} H(\lambda), \qquad \mathrm{Re}\,\lambda \geq 0. \tag{5.5}$$

H-k-desc

(In the scalar factor, λ should be replaced by $\tau\lambda$ with a characteristic time scale $\tau > 0$, which we assume to be 1 for ease of presentation.) Note that the Laplace-transformed input-output relation is

$$\mathcal{L}y(\lambda) = H(\lambda)\,\mathcal{L}u(\lambda) = H^{[k]}(\lambda)\,(1+\lambda)^{k-1}\mathcal{L}u(\lambda), \qquad \mathrm{Re}\,\lambda \geq 0, \tag{5.6}$$

yu-L-Hk

and that

$$(1+\lambda)^{k-1}\mathcal{L}u(\lambda) = \big(\mathcal{L}(1+d/dt)^{k-1}u\big)(\lambda)$$

under our running assumption that u together with its extension by 0 to the negative real half-axis is a sufficiently differentiable function. As in (1.4), since $\|H^{[k]}\|_\infty = \sup_{\mathrm{Re}\,\lambda\geq 0} \|H^{[k]}(\lambda)\|_2$ is finite, these relations imply that the output y is bounded in terms of the input u as

$$\left(\int_0^T \|y(t)\|^2\,dt\right)^{1/2} \leq \|H^{[k]}\|_\infty \left(\int_0^T \|(1+d/dt)^{k-1}u(t)\|^2\,dt\right)^{1/2}, \quad 0 \leq T \leq \infty. \tag{5.7}$$

yu-bound-k

Note that for index $k \geq 2$, the bound depends on derivatives of u up to order $k-1$. This raises the following problem.

Problem. *Compute the $\mathcal{H}_\infty$-norm of the weighted matrix transfer function $H^{[k]}$ of the descriptor system.*

In the following we restrict our attention to the case of principal interest where

$$\|H^{[k]}(\infty)\|_2 < \|H^{[k]}\|_\infty = \sup_{\mathrm{Re}\,\lambda\geq 0} \|H^{[k]}(\lambda)\|_2. \tag{5.8}$$

ass:H-k-inf

Then the supremum is a maximum that is attained at a finite $\lambda = \mathrm{i}\omega$ on the imaginary axis. We will modify the algorithm of Section VI.1 to compute $\|H^{[k]}\|_\infty$.

VI.5.3 $\mathcal{H}_\infty$-norm via a stability radius

In this subsection we give analogues of Theorems 1.2 and 1.6 for the descriptor system (5.1) with the transfer function $H(\lambda)$ of (5.3). We use the notation

$$\Lambda(A,E) = \{\lambda \in \mathbb{C} \,:\, A - \lambda E \text{ is singular}\}.$$

The following result will later be used with $\varphi(\lambda) = (1+\lambda)^{-(k-1)}$, where k is the index of the matrix pencil (A,E).

basicequiv-pen

Theorem 5.2 (Singular values and eigenvalues). *Let $\varepsilon > 0$, $\lambda \in \mathbb{C} \setminus \Lambda(A,E)$, and nonzero $\varphi(\lambda) \in \mathbb{C}$. The following two statements are equivalent:*

(i) $\|\varphi(\lambda)H(\lambda)\|_2 \geq \varepsilon^{-1}$.

(ii) There exists $\Delta \in \mathbb{C}^{p,m}$ with $\|\Delta\|_F \leq \varepsilon$ such that λ is an eigenvalue of the following nonlinear eigenvalue problem: There is an eigenvector $y \in \mathbb{C}^n \setminus \{0\}$ such that

$$(A + \varphi(\lambda)B\Delta C - \lambda E)y = 0. \tag{5.9}$$

twoequiv-pen

Moreover, Δ can be chosen to have rank 1, and the two inequalities can be replaced by equalities in the equivalence.

Proof. For $\varphi(\lambda) = 1$, we can repeat the proof of Theorem 1.2, noting that there the replacement of $A - \lambda I$ by $A - \lambda E$ only leads to obvious changes. For general nonzero $\varphi(\lambda)$, we use the result with $\widetilde{\varepsilon} = |\varphi(\lambda)|\varepsilon$ and $\widetilde{\Delta} = \varphi(\lambda)\Delta$. □

Given a descriptor system such that the matrix pencil (A,E) satisfies (5.2) and is of index k, and choosing $\varphi(\lambda) = (1+\lambda)^{-(k-1)}$ so that $H^{[k]}(\lambda) = \varphi(\lambda)H(\lambda)$, we proceed as in Section VI.1 and define the corresponding spectral value set

$$\begin{aligned}\Lambda_\varepsilon^{[k]} &= \{\lambda \in \mathbb{C} \setminus \Lambda(A,E) \,:\, \|H^{[k]}(\lambda)\|_2 \geq \varepsilon^{-1}\} \\ &= \{\lambda \in \mathbb{C} \setminus \Lambda(A,E) \,:\, \lambda \text{ satisfies (ii) of Theorem 5.2}\}.\end{aligned}$$

The spectral value abscissa

$$\alpha_\varepsilon^{[k]} = \sup\{\operatorname{Re}\lambda \,:\, \lambda \in \Lambda_\varepsilon^{[k]}\}$$

then yields

$$\sup_{\operatorname{Re}\lambda \geq \alpha_\varepsilon^{[k]}} \|H^{[k]}(\lambda)\|_2 = \frac{1}{\varepsilon},$$

and by (5.8), the supremum is a maximum if ε is so small that $\alpha_\varepsilon^{[k]} \leq 0$. With the stability radius

$$\varepsilon_\star^{[k]} = \min\{\varepsilon > 0 \,:\, \alpha_\varepsilon^{[k]} = 0\},$$

we therefore again characterize the $\mathcal{H}_\infty$-norm, which takes the maximum over $\operatorname{Re}\lambda \geq 0$, as the inverse stability radius. Since this result is essential for our numerical approach, we formulate it as a theorem.

hm:Hinf-oeps-k

Theorem 5.3 ($\mathcal{H}_\infty$-norm via the stability radius). *Let the descriptor system be such that the matrix pencil (A,E) has all finite eigenvalues with negative real part and is of index k. Then, the $\mathcal{H}_\infty$-norm of the weighted matrix transfer function $H^{[k]}$ of* (5.5) *and the stability radius $\varepsilon_\star^{[k]}$ are related by*

$$\|H^{[k]}\|_\infty = \frac{1}{\varepsilon_\star^{[k]}}.$$

VI.5.4 Two-level iteration

We use a two-level iteration similar to that of Section VI.1 to compute $\|H^{[k]}\|_\infty$.

- **Inner iteration:** For a given $\varepsilon > 0$, compute the spectral value abscissa $\alpha_\varepsilon^{[k]}$ using the nonlinear eigenvalue problem (5.9) with perturbation matrices Δ of norm ε and of rank 1, which are determined via a rank-1 projected gradient flow that aims to maximize the real part of the rightmost eigenvalue.
- **Outer iteration:** Compute $\varepsilon_\star^{[k]}$ as the smallest $\varepsilon > 0$ such that $\alpha_\varepsilon^{[k]} = 0$, using a mixed Newton / bisection method.

Provided that these computations succeed, we obtain $\|H^{[k]}\|_\infty = 1/\varepsilon_\star^{[k]}$ by Theorem 5.3.

VI.5.5 Inner iteration: constrained gradient flow

We aim to find $\Delta \in \mathbb{C}^{p,m}$ of Frobenius norm ε and of rank 1 such that the rightmost eigenvalue λ yielding a singular matrix

$$M(\Delta, \lambda) := A + \varphi(\lambda) B \Delta C - \lambda E$$

has maximal real part, which equals the ε-spectral value abscissa $\alpha_\varepsilon^{[k]}$. To this end we extend the norm- and rank-constrained gradient flow approach of Section II.1 for doing the inner iteration by a discretized rank-1 gradient flow. Instead of Lemma II.1.1 we now have the following gradient.

n:gradient-desc **Lemma 5.4 (Free gradient).** *Let $\Delta(t) \in \mathbb{C}^{p,m}$, for real t near t_0, be a continuously differentiable path of matrices. Let $\lambda(t)$ be a unique continuously differentiable path of eigenvalues that yield singular matrices $M(\Delta(t), \lambda(t))$ of co-rank 1, with associated left and right eigenvectors $x(t)$ and $y(t)$. Assume that*

$$\eta(t) := x(t)^* \big(E - \varphi'(\lambda(t)) B \Delta(t) C\big) y(t) \neq 0 \quad \textit{and set} \quad \gamma(t) := \frac{1}{\eta(t)}.$$

Then,

$$-\mathrm{Re}\, \dot\lambda(t) = \mathrm{Re} \big\langle G(\Delta(t)), \dot\Delta(t) \big\rangle \tag{5.10}$$

eq:deriv-desc

with the rank-1 matrix (omitting the argument t)

$$G(\Delta) = -\big(\gamma \varphi(\lambda) C y x^* B\big)^* \in \mathbb{C}^{p,m}. \tag{5.11}$$

eq:freegrad-desc

Proof. Differentiating the matrix $A + \varphi(\lambda(t)) B \Delta(t) C - \lambda(t) E$ and multiplying with $x(t)^*$ from the left and $y(t)$ from the right yields (omitting the argument t)

$$-\eta \dot\lambda + \varphi(\lambda) x^* B \dot\Delta C y = 0,$$

which implies

$$\dot\lambda = \gamma \varphi(\lambda) x^* B \dot\Delta C y = \langle x, \gamma \varphi(\lambda) B \dot\Delta C y \rangle = \langle \overline{\gamma \varphi(\lambda)} B^* x y^* C^*, \dot\Delta \rangle$$

and hence yields the stated result. □

With the gradient (5.11), the whole programme of Section II.1 carries through, as we briefly sketch in the following. We write $\varDelta \in \mathbb{C}^{p,m}$ of Frobenius norm ε as

$$\varDelta = \varepsilon Z \quad \text{with} \quad \|Z\|_F = 1$$

(in previous sections we wrote $\varDelta = \varepsilon E$ with $\|E\|_F = 1$, but now E is the singular matrix in the descriptor system) and

$$G_\varepsilon(Z) = G(\varepsilon Z).$$

As in II.4.13, we consider the gradient flow on the Frobenius-norm unit sphere of $\mathbb{C}^{p,m}$,

$$\dot{Z} = -G_\varepsilon(Z) + \mathrm{Re}\langle G_\varepsilon(Z), Z\rangle Z. \tag{5.12}$$

ode-Z-desc

As in Theorem II.1.4 we have that $-\mathrm{Re}\,\lambda(t)$ decreases along solutions of (5.13), where $\lambda(t)$ is a rightmost eigenvalue yielding a singular matrix $M(\lambda(t))$, provided the assumptions of Lemma 5.4 are satisfied.

Stationary points $Z_\star$ of (5.13) are again real multiples of $G_\varepsilon(Z_\star)$ and are therefore of rank 1, as in Corollary II.1.10. As in Section II.1.7 we therefore consider the rank-1-constrained gradient flow

$$\dot{Z} = -P_Z\Big(G_\varepsilon(Z) - \mathrm{Re}\langle G_\varepsilon(Z), Z\rangle Z\Big), \tag{5.13}$$

ode-Z-desc

where P_Z is the orthogonal projection onto the tangent space at Z of the manifold of rank-1 matrices in $\mathbb{C}^{p,m}$. This differential equation has the same properties as Equation (II.1.23) and is discretized in the same way, as described in Section II.1.8.

Computing eigenvalues of the nonlinear eigenvalue problem. What differs from Section II.1 is the computation of eigenvalues λ that yield a singular matrix $M(\lambda)$. If ε is sufficiently small, this can be done efficiently by a fixed-point iteration. Given an iterate λ_n, we set $A_n = A + \varphi(\lambda_n)B\varDelta C$ and compute λ_{n+1} as the rightmost eigenvalue of the matrix pencil $A_n - \lambda E$, for which $\zeta_{n+1} = (\lambda_{n+1}+1)/(\lambda_{n+1}-1)$ is the eigenvalue of largest modulus of the matrix $(A_n - E)^{-1}(A_n + E)$. When ε is not small, one can use algorithms for general nonlinear eigenvalue problems, such as the method of Beyn (**?**) or other methods as reviewed by Güttel & Tisseur (**?**).

VI.5.6 Outer iteration, updating ε

We need to compute the zero of

$$\phi(\varepsilon) = -\alpha_\varepsilon^{[k]} = -\mathrm{Re}\,\lambda_\varepsilon,$$

where λ_ε is a rightmost eigenvalue of the nonlinear eigenvalue problem for the matrix-valued function $A + \varphi(\lambda)B\varepsilon Z_\varepsilon C - \lambda E$ and Z_ε maximizes the real part of the rightmost eigenvalue among all matrices $Z \in \mathbb{C}^{p,m}$ of Frobenius norm 1. Note that Z_ε is to be computed in the inner iteration.

As in Section IV.2 we use a mixed Newton / bisection method, for which we need the derivative $\phi'(\varepsilon)$. As in Theorem IV.2.2 we find, under appropriate regularity assumptions, that

$$\phi'(\varepsilon) = -\|G_\varepsilon(Z_\varepsilon)\|_F.$$

VI.6 Notes

The standard method to compute the $\mathcal{H}_\infty$-norm is the Boyd-Balakrishnan-Bruinsma-Steinbuch algorithm **?**, henceforth called the BBBS algorithm, which generalizes and improves an algorithm of Byers 1988 for computing the distance to instability for A. The method relies on Lemma **??**: for stable A, it needs only to maximize $\|H(\mathrm{i}\omega)\|$ for $\omega \in \mathbb{R}$. The key idea is that, given any $\delta > 0$, it is possible to determine whether or not $\omega \in \mathbb{R}$ exists such that $\|H(\mathrm{i}\omega)\| = \delta$ by computing all eigenvalues of an associated $2n \times 2n$ Hamiltonian matrix and determining whether any are imaginary. The algorithm is quadratically convergent, but the computation of the eigenvalues and the evaluation of the norm of the transfer matrix both require on the order of n^3 operations which is not practical when n is sufficiently large.

For discrete systems

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k + Du_k. \end{aligned} \tag{6.1}$$

an analogous approach can be adopted, with $F_\varepsilon(E) = -|\lambda|^2$ and $\phi(\varepsilon) = F_\varepsilon\left(E(\varepsilon)\right) = \rho_\varepsilon(A,B,C,D)$, where

$$\rho_\varepsilon(A,B,C,D) = \max\{|\lambda| : \lambda \in \Lambda_\varepsilon(A,B,C,D)\}. \tag{6.2}$$

rhoepsdef

is the spectral value set radius. Here one looks for the smallest solution of $f(\varepsilon_\star) = 1$.

Chapter VII.
Graphs

chap:graphs

VII.1 Stability of spectral clustering

VII.2 Constrained graph partitioning

Chapter VIII. Appendix

chap:appendix

VIII.1 Derivatives of eigenvalues and eigenvectors

sect:eig-deriv

In this section we provide a few basic results concerning first order perturbation theory of eigenvalues and eigenvectors. We refer to Greenbaum, Li & Overton (2020) for a recent review, which also traces the rich history of this subject area.

VIII.1.1 First order perturbation theory for simple eigenvalues

We often use the following standard perturbation result for eigenvalues; see e.g. Horn & Johnson (1990), Lemma 6.3.10 and Theorem 6.3.12, and Greenbaum, Li & Overton (2020), Theorem 1.

lem:eigderiv

Theorem 1.1 (Derivative of simple eigenvalues). *Consider a continuously differentiable path of square complex matrices $A(t)$ for t in an open interval I. Let $\lambda(t)$, $t \in I$, be a continuous path of simple eigenvalues of $A(t)$. Let $x(t)$ and $y(t)$ be left and right eigenvectors, respectively, of $A(t)$ to the eigenvalue $\lambda(t)$. Then, $x(t)^* y(t) \neq 0$ for $t \in I$ and λ is continuously differentiable on I with the derivative (denoted by a dot)*

$$\dot{\lambda} = \frac{x^* \dot{A} y}{x^* y}\,. \tag{1.1}$$

Moreover, "continuously differentiable" can be replaced with "analytic" in the assumption and the conclusion.

Since we have $x(t)^* y(t) \neq 0$, we can apply the normalization

$$\|x(t)\| = 1, \;\; \|y(t)\| = 1, \quad x(t)^* y(t) \text{ is real and positive.} \tag{1.2}$$

eq:scalxy

Clearly, a pair of left and right eigenvectors x and y fulfilling (1.2) may be replaced by μy and μx for any complex μ of modulus 1 without changing the property (1.2).

Next we turn to singular values. The following result is obtained from Theorem 1.1 by using the equivalence between singular values of M and eigenvalues of $(0\ M; M^*\ 0)$; see Horn & Johnson (1990), Theorem 7.3.7.

lem:singderiv

Corollary 1.2 (Derivative of singular values). *Consider a continuously differentiable path of matrices $M(t) \in \mathbb{C}^{m,n}$ for t in an open interval I. Let $\sigma(t)$, $t \in I$, be a path of simple singular values of $M(t)$. Let $u(t)$ and $v(t)$ be left and right singular vectors of $M(t)$ to the singular value $\sigma(t)$, that is, $M(t)v(t) = \sigma(t)u(t)$ and $u(t)^*M(t) = \sigma(t)v(t)^*$ with $\|u(t)\| = \|v(t)\| = 1$. Then, σ is differentiable on I with the derivative*

$$\dot{\sigma} = \mathrm{Re}(u^* \dot{M} v).$$

VIII.1.2 First order perturbation theory for eigenvectors

For the derivative of eigenvectors we need the notion of group inverse (or reduced resolvent); see Meyer & Stewart (1988) as well as Kato (1995), Section I.5.3.

def:groupinv

Definition 1.3 (Group inverse). Let $N \in \mathbb{C}^{n,n}$ be a singular matrix with a simple zero eigenvalue. The *group inverse* (or reduced resolvent) of N is the unique matrix Z with

$$NZ = ZN, \qquad ZNZ = Z, \quad \text{and} \quad NZN = N. \tag{1.3}$$

group-inv-cond

It is known from Meyer & Stewart (1988) that if N is normal, then its group inverse Z is equal to the better known Moore–Penrose pseudoinverse $N^\dagger$. In general, the two pseudoinverses are not the same. They are, however, related by the following result, which is a special case of a more general result in Appendix A of Guglielmi, Overton & Stewart (2015) but is also simply verified directly.

thm:Ginv

Theorem 1.4 (Group inverse via Moore–Penrose pseudoinverse). *Suppose that the matrix N has the simple eigenvalue 0 with corresponding left and right eigenvectors x and y of unit norm and such that $x^*y > 0$. Let Z be the group inverse of N, and with $\kappa = 1/(x^*y)$ define the projection $\Pi = I - \kappa y x^*$. Then, the group inverse Z of N is related to the Moore–Penrose pseudoinverse $N^\dagger$ by*

$$Z = \Pi N^\dagger \Pi. \tag{1.4}$$

groupinvformula

The Moore-Penrose pseudo-inverse $N^\dagger$ is obtained from the singular value decomposition $N = U\Sigma V^*$ with unitary matrices U and V and the diagonal matrix $\Sigma = (\Sigma_+ \ 0;\ 0\ 0)$, where Σ_+ is the diagonal matrix of the positive singular values. Then, $N^\dagger = V\Sigma^\dagger U^*$, where $\Sigma^\dagger = ((\Sigma_+)^{-1} \ 0;\ 0\ 0)$. It is then a simple exercise to verify that $\Pi N^\dagger \Pi$ satisfies the conditions (1.3) that define the group inverse Z. In particular, we find $NZ = ZN = \Pi$, which implies the other two conditions.

The group inverse appears in the following result on the derivative of eigenvectors, which is a variant of Theorem 2 of Meyer & Stewart (1988).

thm:eigvecderiv

Theorem 1.5 (Derivative of eigenvectors). *Consider an analytic path of square complex matrices $A(t)$ for t in an open interval I. Let $\lambda(t)$, $t \in I$, be a path of simple eigenvalues of $A(t)$. Then, there exists a continuously differentiable path of associated left and right eigenvectors $x(t)$ and $y(t)$, $t \in I$, which are of unit norm with $x^*(t)y(t) > 0$ and satisfy the differential equations*

$$\begin{aligned}\dot{x}^* &= -x^*\dot{A}Z + \mathrm{Re}(x^*\dot{A}Zx)x^*,\\ \dot{y} &= -Z\dot{A}y \;\; + \mathrm{Re}(y^*Z\dot{A}y)y,\end{aligned} \tag{1.5}$$

deigvec

where $Z(t)$ is the group inverse of $A(t) - \lambda(t)I$.

We note that the last terms on the right-hand sides of (1.5) are in the direction of x^* and y. They serve to ensure that the unit norm of $x(t)$ and $y(t)$ is conserved and that $x(t)^*y(t)$ remains real (and hence positive by Theorem 1.1). This is shown by verifying that $(d/dt)(x^*y)$ is real and that $(d/dt)\|x\|^2 = 2\,\mathrm{Re}\,\dot{x}^*x$ and $(d/dt)\|y\|^2 = 2\,\mathrm{Re}\,y^*\dot{x}$ vanish, using the relations $x^*Z = 0$ and $Zy = 0$, which follow from (1.4).

In Theorem 2 of Meyer & Stewart (1988), the last terms in (1.5) appear without taking the real part. While this preserves the unit norm of the left and right eigenvectors, the positivity of their inner product is then not conserved. Dropping the last terms (which are not analytic) altogether yields an analytic path of non-normalized left and right eigenvectors $x^*(t)$ and $y(t)$ with constant inner product $x^*(t)y(t)$; see Greenbaum, Li & Overton (2020), Theorem 2 and Section 3.4.

Bibliography

AbMS09 P.-A. Absil, R, Mahony and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. *[II.4]*

AhmAB10 S. S. Ahmad, R. Alam, and R. Byers. On pseudospectra, critical points, and multiple eigenvalues of matrix pencils. *SIAM J. Matrix Anal. Appl.*, 31(4): 1915–1933, 2010. *[III.4]*

AlaBKMM11 R. Alam, S. Bora, M. Karow, V. Mehrmann, and J. Moro. Perturbation theory for Hamiltonian matrices and the distance to bounded-realness. *SIAM J. Matrix Anal. Appl.*, 32(2): 484–514, 2011. *[V.6]*

AntMic07 P.J. Antsaklis, and A.N. Michel *A linear systems primer* Birkhäuser Boston Inc, 2007, pages = xvi+517.

BekG01 C. Bekas and E. Gallopoulos. Cobra: parallel path following for computing the matrix pseudospectrum. *Parallel Comput.*, 27(14): 1879–1896, 2001. *[III.4]*

BenLMV15 P. Benner, P. Losse, V. Mehrmann, and M. Voigt. Numerical linear algebra methods for linear differential-algebraic equations. In *Surveys in differential-algebraic equations III*, 117–175. Springer, Cham. 2015. *[V.6]*

BenMSVHV99 P. Benner, V. Mehrmann, V. Sima, S. Van Huffel, and A. Varga. SLICOT–A subroutine library in systems and control theory. In *Applied and computational control, signals, and circuits*, 499–539. Birkhäuser, Boston, MA. 1999. *[V.6]*

BenM19 P. Benner & T. Mitchell. Extended and improved criss-cross algorithms for computing the spectral value set abscissa and radius. *SIAM J. Matrix Anal. Appl.*, 40(4): 1325–1352, 2019. *[III.4]*

BoyB90 S. Boyd and V. Balakrishnan. A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L^∞-norm. *Systems Control Lett.*, 15: 1–7, 1990. *[IV.6]*

BoyBK89 S. Boyd, V. Balakrishnan, and P. Kabamba. A bisection method for computing the $\mathbf{H}_\infty$ norm of a transfer matrix and related problems. *Math. Control Signals Systems*, 2(3): 207–219, 1989. *[III.4]*

Bru96 M. Brühl. A curve tracing algorithm for computing the pseudospectrum. *BIT*, 36(3): 441–454, 1996. *[III.4]*

BruS90 N. A. Bruinsma and M. Steinbuch. A fast algorithm to compute the H^∞-norm of a transfer function matrix, *Systems Control Lett.*, 14: 287–293, 1990. *[IV.6]*

BuLeOv03 J. V. Burke, A. S. Lewis, and M. L. Overton. Robust stability and a criss-cross algorithm for pseudospectra. *IMA J. Numer. Anal.*, 23(3): 359–375, 2003. *[III.2]*

ButGN12 P. Buttà, N. Guglielmi, and S. Noschese, Computing the structured pseudospectrum of a Toeplitz matrix and its extreme points. *SIAM J. Matrix Anal. Appl.*, 33(4): 1300–1319, 2012.

Bye88 R. Byers. A bisection method for measuring the distance of a stable matrix to the unstable matrices. *SIAM J. Sci. Statist. Comput.*, 9(5): 875–881, 1988. *[III.4], [IV.6], [VI.3], [VI.6]*

CeL21 G. Ceruti and Lubich. An unconventional robust integrator for dynamical low-rank approximation. *BIT*, to appear, 2021. *[II.2], [II.4], [V.1]*

ChoGS20 N. Choudhary, N. Gillis and P. Sharma. On approximating the nearest Ω-stable matrix. *Numer. Linear Algebra Appl.*, 27(3): e2282, 13 pp., 2020. *[IV.6]*

CurMO17 F. E. Curtis, T. Mitchell and M. L. Overton. A BFGS-SQP method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles. *Optim. Methods Softw.*, 32(1): 148–181, 2017. *[IV.6]*

Dav07 E. B. Davies. *Linear operators and their spectra*. Cambridge Studies in Advanced Mathematics, vol. 106. Cambridge University Press, Cambridge, 2007. *[III.4]*

Dor97 J. L. M. van Dorsselaer. Pseudospectra for matrix pencils and stability of equilibria. *BIT*, 37(4): 833–845, 1997. *[III.4]*

DorKS93 J. L. M. van Dorsselaer, J. F. B. M. Kraaijevanger, and M. N. Spijker. Linear stability analysis in the numerical solution of initial value problems. *Acta numerica*, 1993, 199–237, Acta Numer., Cambridge Univ. Press, Cambridge, 1993. *[III.4]*

Eis10 T. Eisner. *Stability of operators and operator semigroups*. Operator Theory: Advances and Applications, Vol. 209. Birkhäuser, Basel, 2010. *[III.4]*

FGL21 A. Fazzi, N. Guglielmi and C. Lubich. Finding the nearest passive or nonpassive system via Hamiltonian eigenvalue optimization. *SIAM J. Matrix Anal. Appl.*, 42(4): 1553–1580, 2021. *[IV.6], [V.6]*

FreS11 M. A. Freitag and A. Spence. A Newton-based method for the calculation of the distance to instability. *Linear Algebra Appl.*, 435(12): 3189–3205, 2011. *[IV.6]*

FreS14 M. A. Freitag and A. Spence. A new approach for calculating the real stability radius. *BIT*, 54(2): 381–400, 2014. *[IV.6]*

GilKS19 N. Gillis, M. Karow and P. Sharma. Approximating the nearest stable discrete-time system. *Linear Algebra Appl.*, 573, 37–53, 2019. *[IV.6]*

GilS17 N. Gillis and P. Sharma. On computing the distance to stability for matrices using linear dissipative Hamiltonian systems. *Automatica J. IFAC*, 85: 113–121, 2017. *[IV.6]*

GreLO20 A. Greenbaum, R.-C. Li, and M. L. Overton. First-order perturbation theory for eigenvalues and eigenvectors. *SIAM Rev.*, 62(2): 463–482, 2020. *[VIII.1]*

GriG15 S. Grivet-Talocia and B. Gustavsen. *Passive macromodeling: Theory and applications.* John Wiley & Sons, 2015. *[III.4]*

Gug16 N. Guglielmi. On the method by Rostami for computing the real stability radius of large and sparse matrices. *SIAM J. Sci. Comput.*, 38(3): A1662–A1681, 2016. *[IV.6]*

GugGO13 N. Guglielmi, M. Gürbüzbalaban, and M. L. Overton. Fast approximation of the H_∞ norm via optimization over spectral value sets. *SIAM J. Matrix Anal. Appl.*, 34(2): 709–737, 2013. *[III.4]*

GKL14 N. Guglielmi, D. Kressner, and C. Lubich. Computing extremal points of symplectic pseudospectra and solving symplectic matrix nearness problems. *SIAM J. Matrix Anal. Appl.*, 35(4): 1407–1428, 2014. *[III.4]*

GKL15 N. Guglielmi, D. Kressner, and C. Lubich. Low rank differential equations for Hamiltonian matrix nearness problems. *Numer. Math.*, 129(2): 279–319, 2015. *[II.4], [III.4], [IV.6], [V.6]*

GL11 N. Guglielmi and C. Lubich. Differential equations for roaming pseudospectra: paths to extremal points and boundary tracking. *SIAM J. Numer. Anal.*, 49: 1194–1209, 2011. *[II.4], [III.2], [III.4], [IV.6]*

GL12 N. Guglielmi and C. Lubich. Erratum/addendum: Differential equations for roaming pseudospectra: paths to extremal points and boundary tracking. *SIAM J. Numer. Anal.*, 50: 977–981, 2012. *[II.4], [III.4]*

GL13 N. Guglielmi and C. Lubich. Low-rank dynamics for computing extremal points of real pseudospectra. *SIAM J. Matrix Anal. Appl.*, 34(1): 40–66, 2013. *[II.4], [III.4], [IV.6]*

GL17 N. Guglielmi and C. Lubich. Matrix stabilization using differential equations. *SIAM J. Numer. Anal.*, 55(6): 3097–3119, 2017. *[IV.6]*

GM15 N. Guglielmi and M. Manetta. Approximating real stability radii. *IMA J. Numer. Anal.*, 35(3): 1402–1425, 2015. *[IV.6]*

GO11 N. Guglielmi and M. L. Overton. Fast algorithms for the approximation of the pseudospectral abscissa and pseudospectral radius of a matrix. *SIAM J. Matrix Anal. Appl.*, 32(4): 1166–1192, 2011. *[II.4], [III.2], [III.4], [IV.6]*

GOS15 N. Guglielmi, M. L. Overton, and G.W. Stewart. An efficient algorithm for computing the generalized null space decomposition. *SIAM J. Matrix Anal. Appl.*, 36(1): 38–54, 2015. *[VIII.1]*

GP18 N. Guglielmi and V. Yu. Protasov. On the closest stable/unstable nonnegative matrix and related stability radii. *SIAM J. Matrix Anal. Appl.*, 39(4): 1642–1669, 2018. *[IV.6]*

HeW99 C. He and G.A. Watson. An algorithm for computing the distance to instability. *SIAM J. Matrix Anal. Appl.*, 20(1): 101–116 (1999). *[IV.6]*

Hig89 N. J. Higham. Matrix nearness problems and applications. Applications of matrix theory (Bradford, 1988), 1–27, Inst. Math. Appl. Conf. Ser. New Ser., 22, Oxford Univ. Press, New York, 1989. *[IV.6]*

HinK93 D. Hinrichsen and B. Kelb. Spectral value sets: a graphical tool for robustness analysis. *Systems Control Lett.*, 21(2): 127–136, 1993. *[III.4]*

HinKL89 D. Hinrichsen, B. Kelb and A. Linnemann. An algorithm for the computation of the structured complex stability radius. *Automatica J. IFAC*, 25(5): 771–775, 1989. *[IV.6]*

HinP86a D. Hinrichsen and A. J. Pritchard. Stability radii of linear systems. *Systems Control Lett.*, 7(1): 1–10, 1986. *[III.4], [IV.6]*

HinP86b D. Hinrichsen and A. J. Pritchard. Stability radius for structured perturbations and the algebraic Riccati equation. *Systems Control Lett.*, 8(2): 115–113, 1986. *[III.4], [IV.6]*

HinP90 D. Hinrichsen and A. J. Pritchard. Real and complex stability radii: a survey. Control of uncertain systems (Bremen, 1989), 119–162, Progr. Systems Control Theory, 6, Birkhäuser, Boston, MA, 1990. *[III.4], [IV.6]*

HinP05 D. Hinrichsen and A. J. Pritchard. *Mathematical systems theory I: modelling, state space analysis, stability and robustness*. Springer, Berlin, 2005. *[III.4], [VI.1]*

HJ90 R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original. *[VIII.1]*

KanMMM18 F. Kangal, K. Meerbergen, E. Mengi, and W. Michiels. A subspace method for large-scale eigenvalue optimization. *SIAM J. Matrix Anal. Appl.*, 39(1): 48–82, 2018. *[III.4]*

Kar03 M. Karow. *Geometry of spectral value sets*. Doctoral dissertation, Univ. Bremen, 2003. *[III.4]*

KarKK10 M. Karow, E. Kokiopoulou, and D. Kressner. On the computation of structured singular values and pseudospectra. *Systems Control Lett.*, 59(2): 122–129, 2010. *[III.4]*

Kat95 T. Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition. *[VIII.1]*

KL07 O. Koch and C. Lubich. Dynamical low-rank approximation. *SIAM J. Matrix Anal. Appl.*, 29(2): 434–454, 2007. *[II.4]*

Kr62 H.-O. Kreiss. Über die Stabilitätsdefinition für Differenzengleichungen, die partielle Differentialgleichungen approximieren. *Nordisk Tidskr. Informations-Behandling*, 2: 153–181, 1962. *[III.4]*

Kre06 D. Kressner. Finding the distance to instability of a large sparse matrix. In 2006 IEEE Conference on Computer Aided Control System Design, 2006 IEEE International Conference on Control Applications, 2006 IEEE International Symposium on Intelligent Control. IEEE. pp. 31–35, 2006. *[IV.6]*

KV14 D. Kressner and B. Vandereycken. Subspace methods for computing the pseudospectral abscissa and the stability radius. *SIAM J. Matrix Anal. Appl.*, 35(1): 292–313, 2014. *[III.2]*

KreLV18 D. Kressner, D. Lu, and B. Vandereycken. Subspace acceleration for the Crawford number and related eigenvalue optimization problems. *SIAM J. Matrix Anal. Appl.*, 39(2): 961–982, 2018. *[III.4]*

LeSY98 R. B.Lehoucq, D. C. Sorensen, and C. Yang. ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, SIAM Publications, Philadelphia, 1998.

LeVT84 R. J. LeVeque and L. N. Trefethen. On the resolvent condition in the Kreiss matrix theorem. *BIT*, 24(4): 584–591, 1984. *[III.4]*

LeO96 A. S. Lewis and M. L. Overton. Eigenvalue optimization. *Acta Numerica*, 5:149–190, 1996. *[II.4]*

LuV17 D. Lu and B. Vandereycken. Criss-cross type algorithms for computing the real pseudospectral abscissa. *SIAM J. Matrix Anal. Appl.*, 38(3): 891–923, 2017. *[III.4]*

LubN91 C. Lubich & O. Nevanlinna. On resolvent conditions and stability estimates. *BIT*, 31(2): 293–313, 1991. *[III.4]*

LuO14 C. Lubich and I. V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT*, 54(1): 171–188, 2014. *[II.4]*

MeX08 V. Mehrmann and H. Xu. Perturbation of purely imaginary eigenvalues of Hamiltonian matrices under structured perturbations. *Electron. J. Lin. Alg.*, 17: 234–257, 2008. *[V.1], [V.6]*

MeOv05 E. Mengi and M. L. Overton. Algorithms for the computation of the pseudospectral radius and the numerical radius of a matrix. *IMA J. Numer. Anal.*, 25(4): 648–669, 2005. *[III.4]*

MeR96 K. Meerbergen, D. Roose, Matrix transformations for computing rightmost eigenvalues of large sparse non-symmetric eigenvalue problems. *IMA J. Numer. Anal.* 16: 297–346, 1996.

MeSR94 K. Meerbergen, A. Spence, D. Roose. Shift-invert and Cayley transforms for detection of rightmost eigenvalues of nonsymmetric matrices. *BIT*, 34: 409–423, 1994.

MS88 C. D. Meyer and G. W. Stewart. Derivatives and perturbations of eigenvectors. *SIAM J. Numer. Anal.*, 25(3): 679–691, 1988. *[VIII.1]*

MezP02 D. Mezher and B. Philippe. Parallel computation of pseudospectra of large sparse matrices. *Parallel Computing*, 28(2): 199–221, 2002. *[III.4]*

NesP20 Yu. Nesterov and V. Yu. Protasov Computing closest stable nonnegative matrix. *SIAM J. Matrix Anal. Appl.*, 41(1): 1–28, 2020. *[IV.6]*

NofP21 V. Noferini and F. Poloni. Nearest Ω-stable matrix via Riemannian optimization. *Numer. Math.*, 148(4): 817–851, 2021. *[IV.6]*

OrbNVD13 F.-X. Orbandexivry, Yu. Nesterov and P. Van Dooren. Nearest stable system using successive convex approximations. *Automatica J. IFAC*, 49(5): 1195–1203, 2013. *[IV.6]*

PaiVL81 C. Paige and C. Van Loan. A Schur decomposition for Hamiltonian matrices. *Linear Algebra Appl.*, 41: 11–32, 1981. *[V.1]*

QiuBRDYD95 L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, P. M. Young, J. C. Doyle, A formula for computation of the real stability radius. *Automatica J. IFAC*, 31(6): 879– 890, 1995. *[III.4], [IV.6]*

RedT92 S. C. Reddy and L. N. Trefethen. Stability of the method of lines. *Numerische Mathematik*, 62(1): 235–267, 1992. *[III.4]*

Ros15 M. W. Rostami, New algorithms for computing the real structured pseudospectral abscissa and the real stability radius of large and sparse matrices. *SIAM J. Sci. Comput.*, 37(5): S447–S471, 2015. *[IV.6]*

Spi91 M. N. Spijker. On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem. *BIT*, 31(3): 551–555, 1991. *[III.4]*

SreVDT96 J. Sreedhar, P. Van Dooren and A.L. Tits. A fast algorithm to compute the real structured stability radius. Stability theory (Ascona, 1995), Internat. Ser. Numer. Math. 121: 219–230, 1996. *[IV.6]*

TisH01 F. Tisseur & N. J. Higham. Structured pseudospectra for polynomial eigenvalue problems, with applications. *SIAM J. Matrix Anal. Appl.*, 23(1): 187–208, 2001. *[III.4]*

Tre92 L. N. Trefethen. Pseudospectra of matrices. Numerical analysis 1991 (Dundee, 1991), 234–266, Pitman Res. Notes Math. Ser., 260, Longman Sci. Tech., Harlow, 1992. *[III.4]*

Tre97 L. N. Trefethen. Pseudospectra of linear operators. *SIAM Rev.*, 39: 383–406, 1997. *[III.4]*

Tre99 L. N. Trefethen. Computation of pseudospectra. *Acta numerica*, 8: 247–295, 1999. *[III.4]*

TreE05 L. N. Trefethen and M. Embree. *Spectra and Pseudospectra. The behavior of nonnormal matrices and operators.*. Princeton University Press, Princeton, NJ, 2005. *[III.4]*

VL85 C. Van Loan. How near is a stable matrix to an unstable matrix? In *Linear algebra and its role in systems theory (Brunswick, Maine, 1984)*, volume 47 of *Contemp. Math.*, pages 465–478. Amer. Math. Soc., Providence, RI, 1985. *[III.4], [IV.6]*

Var67 J. M. Varah. The computation of bounds for the invariant subspaces of a general matrix operator. Tech. Rep. CS66, Stanford Univ., 1967. *[III.4]*

Wil65 J. H. Wilkinson. *The algebraic eigenvalue problem.* Clarendon Press, Oxford, 1965. *[III.4], [V.2]*

Wil86 J. H. Wilkinson. Sensitivity of eigenvalues. II. *Utilitas Math.*, 30:243–286, 1986. *[III.4]*

Wri02 T. G. Wright. Eigtool: a graphical tool for nonsymmetric eigenproblems. *Oxford University Computing Laboratory, http://www.comlab.ox.ac.uk/pseudospectra/eigtool/*, 2002. *[III.3], [III.4]*

WriT01 T. G. Wright and L. N. Trefethen. Large-scale computation of pseudospectra using ARPACK and eigs. *SIAM J. Sci. Comput.*, 23(2): 591–605, 2001. *[III.4]*

WriT02 T. G. Wright and L. N. Trefethen. Pseudospectra of rectangular matrices. *IMA J. Numer. Anal.*, 22(4), 501–519, 2002. *[III.4]*

P81 C. Paige. Properties of numerical algorithms related to computing controllability, *IEEE Trans. Automat. Control*, 26:130–138, 1981.

W91 J. C. Willems. Paradigms and puzzles in the theory of dynamical systems, *IEEE Trans. Automat. Control*, 36(3):259–294, 1991.

PW98 J. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory*. New York: Springer-Verlag, 1998.

CH80 J. B. Conway and P. R. Halmos. Finite-dimensional points of continuity of Lat. *Linear Algebra Appl.*, 31, 93–102, 1980. *[V.2]*

VN50 J. von Neumann. *Functional Operators. II. The Geometry of Orthogonal Spaces*. Annals of Mathematics Studies, no. 22. Princeton University Press, 1950.