Functions of Nonnormal Matrices and the Behavior of Dynamical Systems

Mark Embree
Department of Mathematics
Virginia Tech

AIM Workshop on Mathematical Aspects of Physics with Non-Self-Adjoint Operators

San Jose · June 2015

Setting

Given a nonnormal matrix $\mathbf{A} \in \mathbf{C}^{n \times n}$ (or operator in a Hilbert space), we seek to understand how functions of $\mathbf{A}$ and dynamical systems driven by $\mathbf{A}$,

$$f(\mathbf{A}), \qquad \mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t),$$

behave, in contrast to analogous problems for normal/self-adjoint operators.

We will survey a set of related problems motivated by transient dynamics.

- Linear Stability Analysis and Transient Dynamics
- Diagonalization and Davies's Conjecture
- Numerical Range
  – Crouzeix's Conjecture on $\|f(\mathbf{A})\|$
  – Inverse numerical range problems / localization of Ritz values
- Pseudospectra
  – "Do pseudospectra determine behavior?"
  – Pseudospectra of matrix pencils for Differential Algebraic Equations
- Lyapunov Equations
  – Singular values of Lyapunov solutions with nonnormal coefficients

Linear Stability Analysis and Transients

Linear Stability Analysis for Dynamical Systems

Linear Stability Analysis

Consider the autonomous nonlinear system $\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t))$.

- *Find a steady state* $\mathbf{u}_*$, *i.e.,* $\mathbf{f}(\mathbf{u}_*) = \mathbf{0}$.
- *Linearize* $\mathbf{f}$ *about this steady state and analyze small perturbations,* $\mathbf{u}(t) = \mathbf{u}_* + \mathbf{x}(t)$*:*

$$\begin{aligned} \mathbf{x}'(t) = \mathbf{u}'(t) &= \mathbf{f}(\mathbf{u}_* + \mathbf{x}(t)) \\ &= \mathbf{f}(\mathbf{u}_*) + \mathbf{A}\mathbf{x}(t) + O(\|\mathbf{x}(t)\|^2) \\ &= \mathbf{A}\mathbf{x}(t) + O(\|\mathbf{x}(t)\|^2). \end{aligned}$$

- *Ignore higher-order effects, and analyze the linear system* $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$. *The steady state* $\mathbf{u}_*$ *is stable provided* $\mathbf{A}$ *is stable: i.e., all its eigenvalues are in the left half-plane.*

Linear Stability Analysis for Dynamical Systems

Linear Stability Analysis

Consider the autonomous nonlinear system $\mathbf{u}'(t) = \mathbf{f}(\mathbf{u}(t))$.

- *Find a steady state* $\mathbf{u}_*$, *i.e.,* $\mathbf{f}(\mathbf{u}_*) = \mathbf{0}$.
- *Linearize* $\mathbf{f}$ *about this steady state and analyze small perturbations,* $\mathbf{u}(t) = \mathbf{u}_* + \mathbf{x}(t)$:

$$\begin{aligned} \mathbf{x}'(t) = \mathbf{u}'(t) &= \mathbf{f}(\mathbf{u}_* + \mathbf{x}(t)) \\ &= \mathbf{f}(\mathbf{u}_*) + \mathbf{A}\mathbf{x}(t) + O(\|\mathbf{x}(t)\|^2) \\ &= \mathbf{A}\mathbf{x}(t) + O(\|\mathbf{x}(t)\|^2). \end{aligned}$$

- *Ignore higher-order effects, and analyze the linear system* $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$. *The steady state* $\mathbf{u}_*$ *is stable provided* $\mathbf{A}$ *is stable: i.e., all its eigenvalues are in the left half-plane.*

But what if the small perturbation $\mathbf{x}(t)$ grows
by orders of magnitude before eventually decaying?

Example: A nonlinear heat equation

An example from [Zworski; Galkowski, 2012]:

For $x \in [-1, 1]$ and $t \geq 0$ with $u(-1, t) = u(1, t) = 0$, consider

$$u_t(x, t) = \nu u_{xx}(x, t)$$

with $\nu > 0$

Example: A nonlinear heat equation

An example from [Zworski; Galkowski, 2012]:

For $x \in [-1, 1]$ and $t \geq 0$ with $u(-1, t) = u(1, t) = 0$, consider

$$u_t(x, t) = \nu u_{xx}(x, t) + \sqrt{\nu} u_x(x, t)$$

with $\nu > 0$

Example: A nonlinear heat equation

An example from [Zworski; Galkowski, 2012]:

For $x \in [-1, 1]$ and $t \geq 0$ with $u(-1, t) = u(1, t) = 0$, consider

$$u_t(x, t) = \nu u_{xx}(x, t) + \sqrt{\nu} u_x(x, t) + \tfrac{1}{8} u(x, t)$$

with $\nu > 0$

Example: A nonlinear heat equation

An example from [Zworski; Galkowski, 2012]:

For $x \in [-1, 1]$ and $t \geq 0$ with $u(-1, t) = u(1, t) = 0$, consider

$$u_t(x, t) = \nu u_{xx}(x, t) + \sqrt{\nu} u_x(x, t) + \tfrac{1}{8} u(x, t) + u(x, t)^p$$

with $\nu > 0$ and $p > 1$.

Example: A nonlinear heat equation

An example from [Zworski; Galkowski, 2012]:

For $x \in [-1, 1]$ and $t \geq 0$ with $u(-1, t) = u(1, t) = 0$, consider

$$u_t(x, t) = \nu u_{xx}(x, t) + \sqrt{\nu} u_x(x, t) + \tfrac{1}{8} u(x, t) + u(x, t)^p$$

with $\nu > 0$ and $p > 1$.

The linearization L, an advection–diffusion operator,

$$Lu = \nu u_{xx} + \sqrt{\nu} u_x + \tfrac{1}{8} u$$

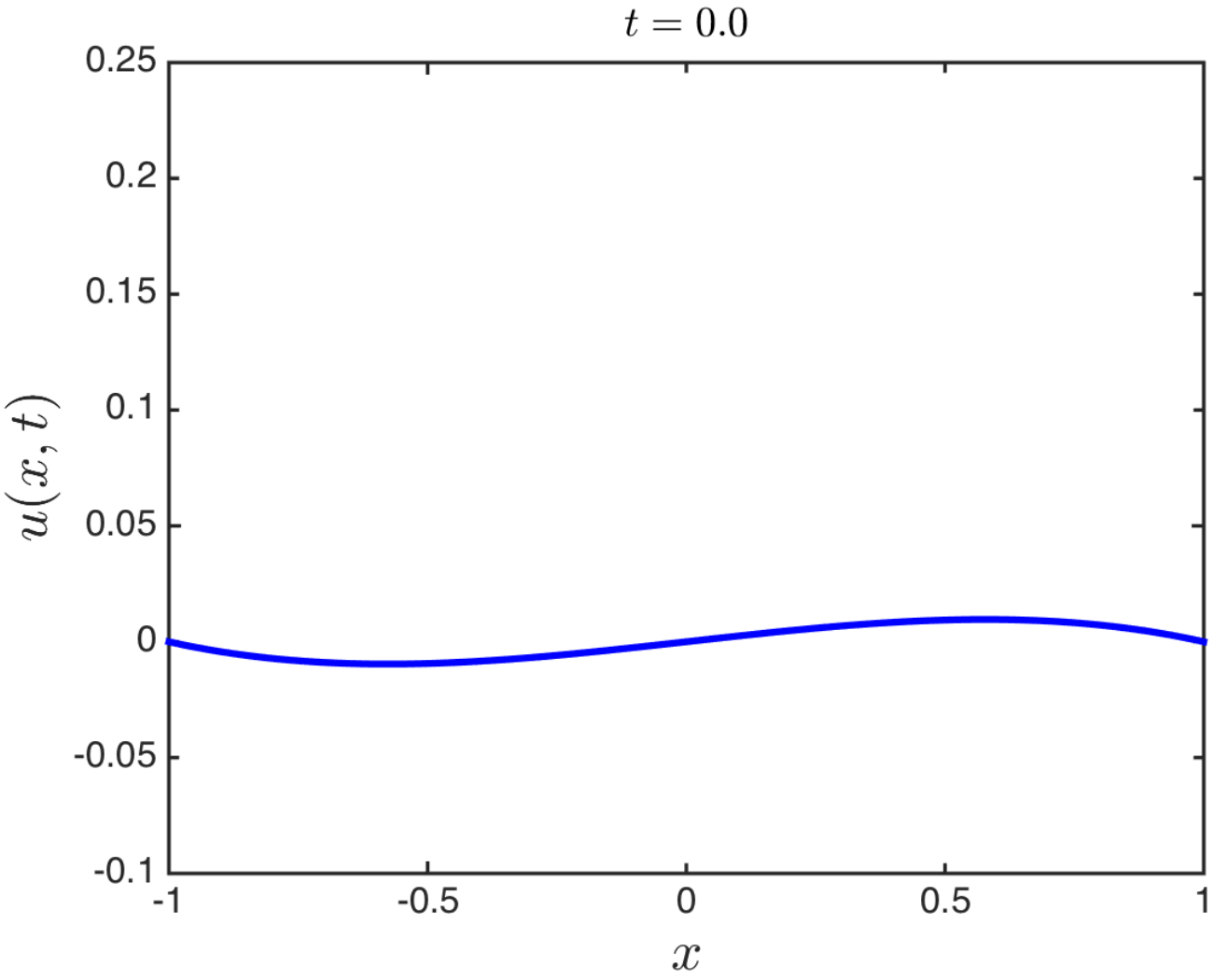
has eigenvalues and eigenfunctions

$$\lambda_n = -\frac{1}{8} - \frac{n^2 \pi^2 \nu}{4}, \qquad u_n(x) = e^{-x/(2\sqrt{\nu})} \sin(n\pi x/2);$$

see, e.g., [Reddy & Trefethen 1994].

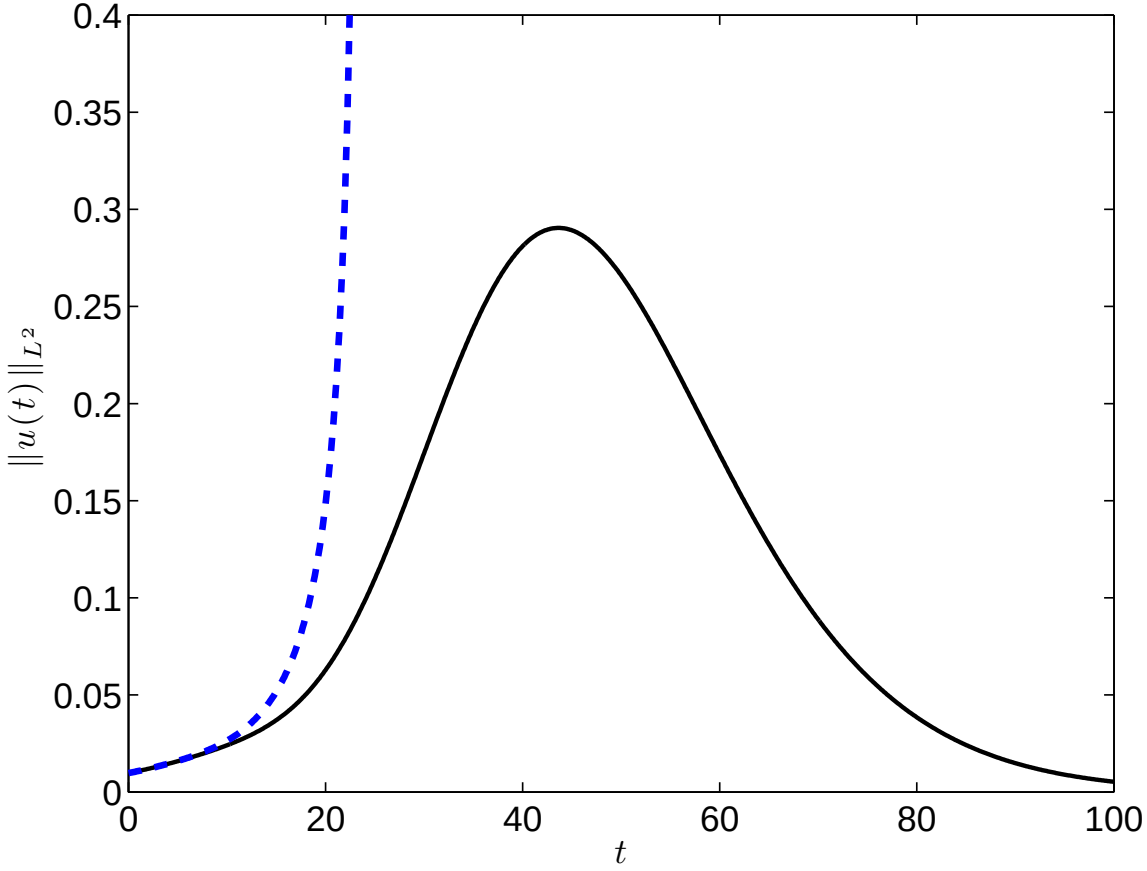
The linearized operator is stable for all $\nu > 0$, but has interesting transients . . . .

Evolution of a small initial condition



Nonlinear model (blue) and linearization (black)

Transient behavior



Linearized system (black) and nonlinear system (dashed blue)

Nonnormal growth feeds the nonlinear instability.

Transient behavior: reduction of the linearized model

The linearization L is stable. So too is any reasonable discretization $\mathbf{L}$.

What happens when we apply model reduction to the discretization, e.g., to create a surrogate in a design problem?

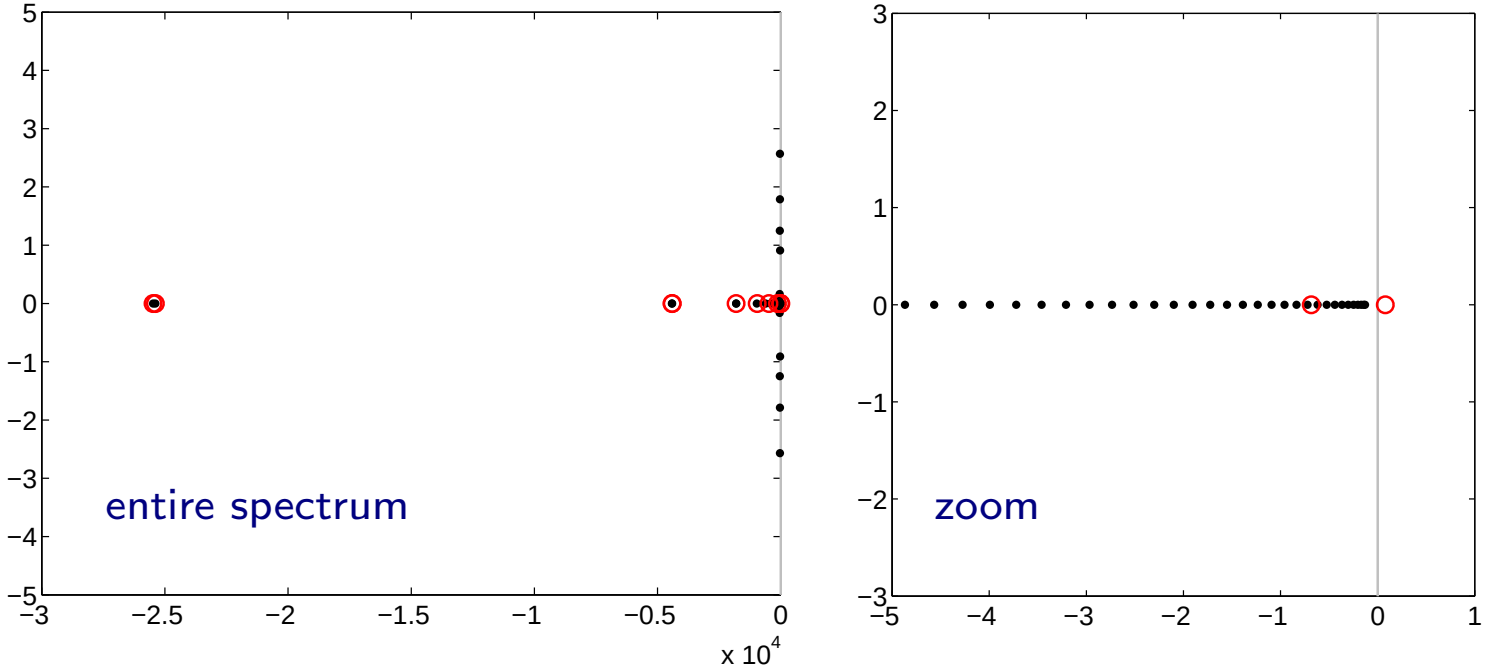
Apply Arnoldi moment-matching model reduction to the discretization $\mathbf{L}$ of order 100 to generate a $k = 10$ dimensional model $\mathbf{L}_{10} = \mathbf{V}_{10}^* \mathbf{L} \mathbf{V}_{10}$.
(This does not guarantee stability, but we will have $W(\mathbf{L}_{10}) \subseteq W(\mathbf{L})$.)

Transient behavior: reduction of the linearized model

The linearization L is stable. So too is any reasonable discretization $\mathbf{L}$.

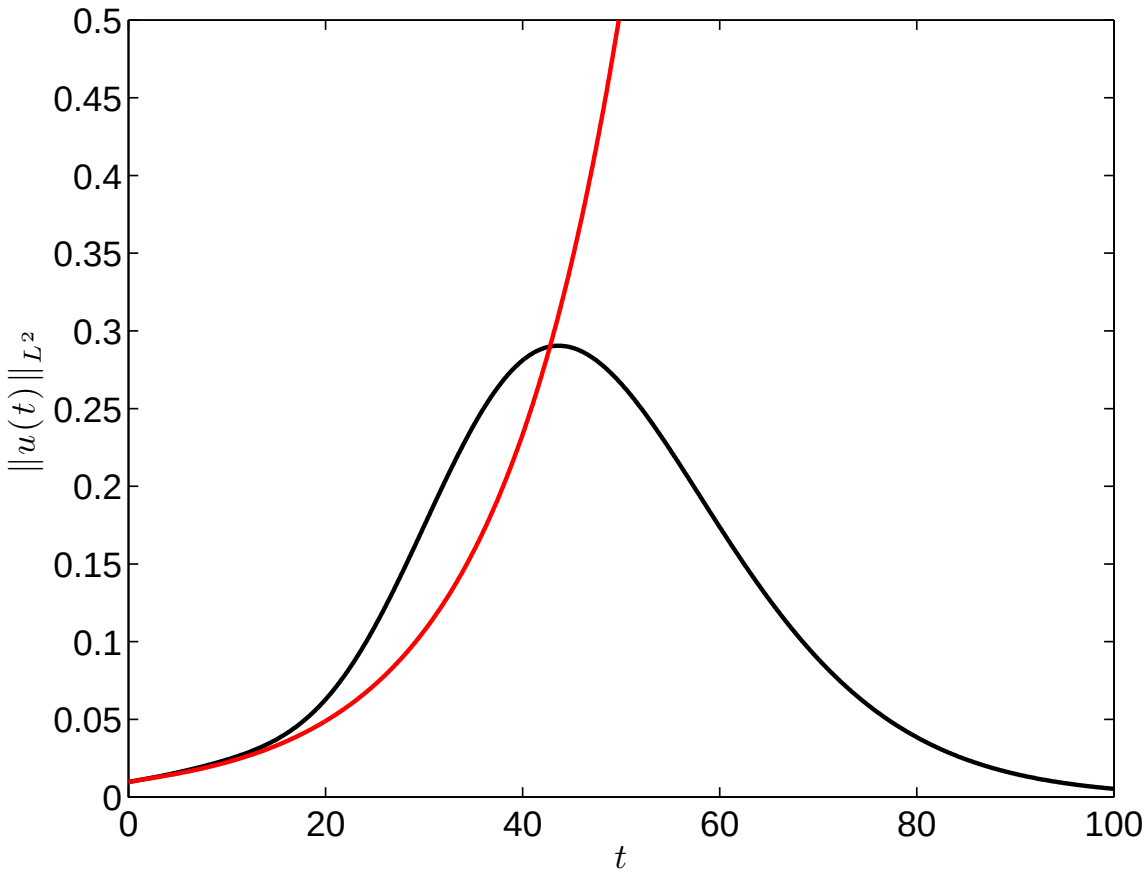
What happens when we apply model reduction to the discretization, e.g., to create a surrogate in a design problem?

Apply Arnoldi moment-matching model reduction to the discretization $\mathbf{L}$ of order 100 to generate a $k = 10$ dimensional model $\mathbf{L}_{10} = \mathbf{V}_{10}^{*}\mathbf{L}\mathbf{V}_{10}$.
(This does not guarantee stability, but we will have $W(\mathbf{L}_{10}) \subseteq W(\mathbf{L})$.)



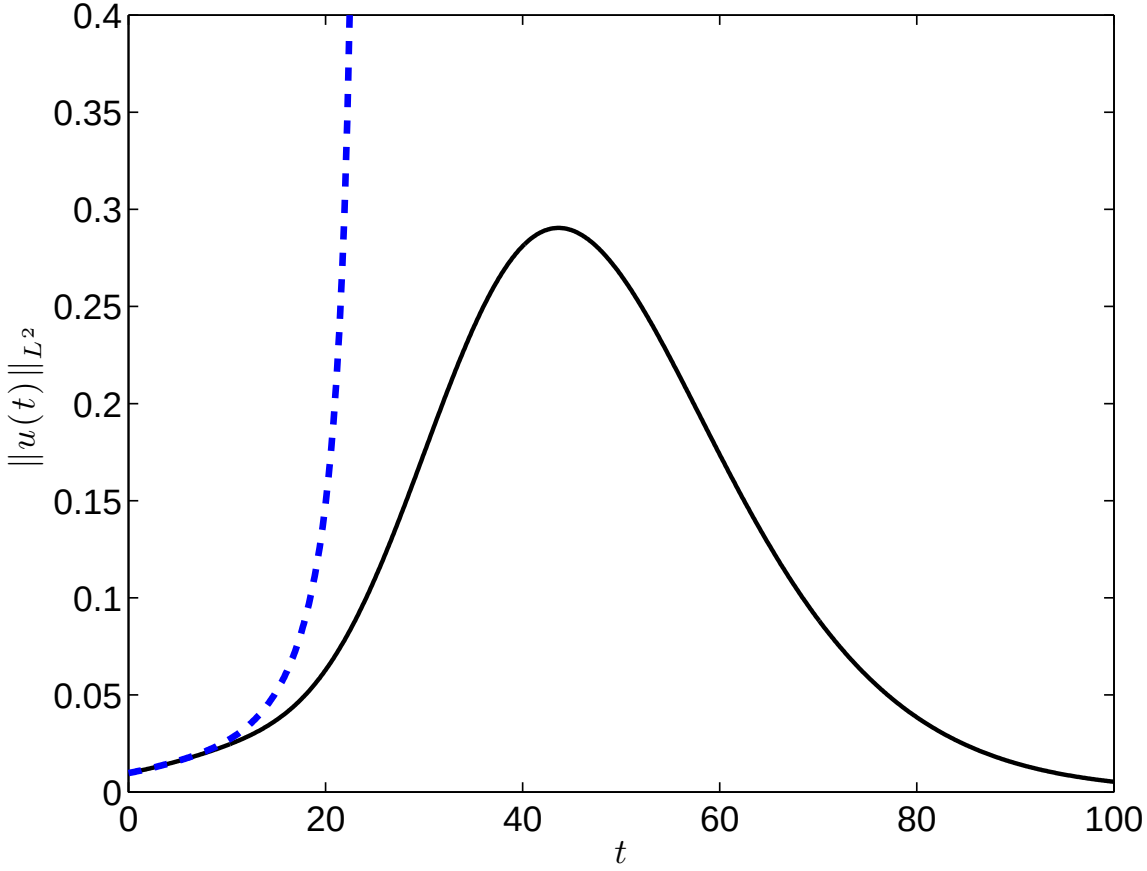
Spectral discretization, $n = 128$ (black) and Arnoldi reduction, $k = 10$ (red).
[Many Ritz values capture *spurious* eigenvalues of the discretization of the left.]

Transient behavior: reduction of the linearized model



Spectral discretization, $n = 128$ (black) and Arnoldi reduction, $k = 10$ (red).

Transient behavior: nonlinear versus linear system



Linearized system (black) and nonlinear system (dashed blue)

Nonnormal growth feeds the nonlinear instability.

Tools for Understanding Transient Growth: Eigenvectors

We need spectral tools to understand the potential for transient dynamics.

If $\mathbf{A}$ is diagonalizable, $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$, then one can bound the transient growth in $e^{t\mathbf{A}}$ using the *condition number of the eigenvector matrix*.

Example (Eigenvalue/Eigenvector Bound for Continuous-Time Systems)

$$\begin{aligned}
\|\mathbf{x}(t)\| = \|e^{t\mathbf{A}}\mathbf{x}(0)\| &\leq \|e^{t\mathbf{A}}\|\|\mathbf{x}(0)\| \\
&\leq \|\mathbf{V}e^{t\boldsymbol{\Lambda}}\mathbf{V}^{-1}\|\|\mathbf{x}(0)\| \\
&\leq \|\mathbf{V}\|\|\mathbf{V}^{-1}\| \max_{\lambda\in\sigma(\mathbf{A})} |e^{t\lambda}|\|\mathbf{x}(0)\|.
\end{aligned}$$

Tools for Understanding Transient Growth: Numerical Range

Definition (Numerical Range, a.k.a. Field of Values)

The *numerical range* of $\mathbf{A}$ is the set

$$W(\mathbf{A}) = \{\mathbf{v}^*\mathbf{A}\mathbf{v} : \|\mathbf{v}\| = 1\}.$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|\mathrm{e}^{t\mathbf{A}}\mathbf{x}_0\|\Big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{x}_0^*\mathrm{e}^{t\mathbf{A}^*}\mathrm{e}^{t\mathbf{A}}\mathbf{x}_0\right)^{1/2}\Big|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{x}_0^*(\mathbf{I}+t\mathbf{A}^*)(\mathbf{I}+t\mathbf{A})\mathbf{x}_0\right)^{1/2}\Big|_{t=0} = \frac{1}{\|\mathbf{x}_0\|}\mathbf{x}_0^*\left(\frac{\mathbf{A}+\mathbf{A}^*}{2}\right)\mathbf{x}_0
\end{aligned}$$

So, the rightmost point in $W(\mathbf{A})$ reveals the maximal slope of $\|\mathrm{e}^{t\mathbf{A}}\|$ at $t = 0$.

Tools for Understanding Transient Growth: Numerical Range

Definition (Numerical Range, a.k.a. Field of Values)

The *numerical range* of $\mathbf{A}$ is the set

$$W(\mathbf{A}) = \{\mathbf{v}^*\mathbf{A}\mathbf{v} : \|\mathbf{v}\| = 1\}.$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|\mathrm{e}^{t\mathbf{A}}\mathbf{x}_0\|\Big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{x}_0^*\mathrm{e}^{t\mathbf{A}^*}\mathrm{e}^{t\mathbf{A}}\mathbf{x}_0\right)^{1/2}\Big|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{x}_0^*(\mathbf{I}+t\mathbf{A}^*)(\mathbf{I}+t\mathbf{A})\mathbf{x}_0\right)^{1/2}\Big|_{t=0} = \frac{1}{\|\mathbf{x}_0\|}\mathbf{x}_0^*\left(\frac{\mathbf{A}+\mathbf{A}^*}{2}\right)\mathbf{x}_0
\end{aligned}$$

So, the rightmost point in $W(\mathbf{A})$ reveals the maximal slope of $\|\mathrm{e}^{t\mathbf{A}}\|$ at $t = 0$.

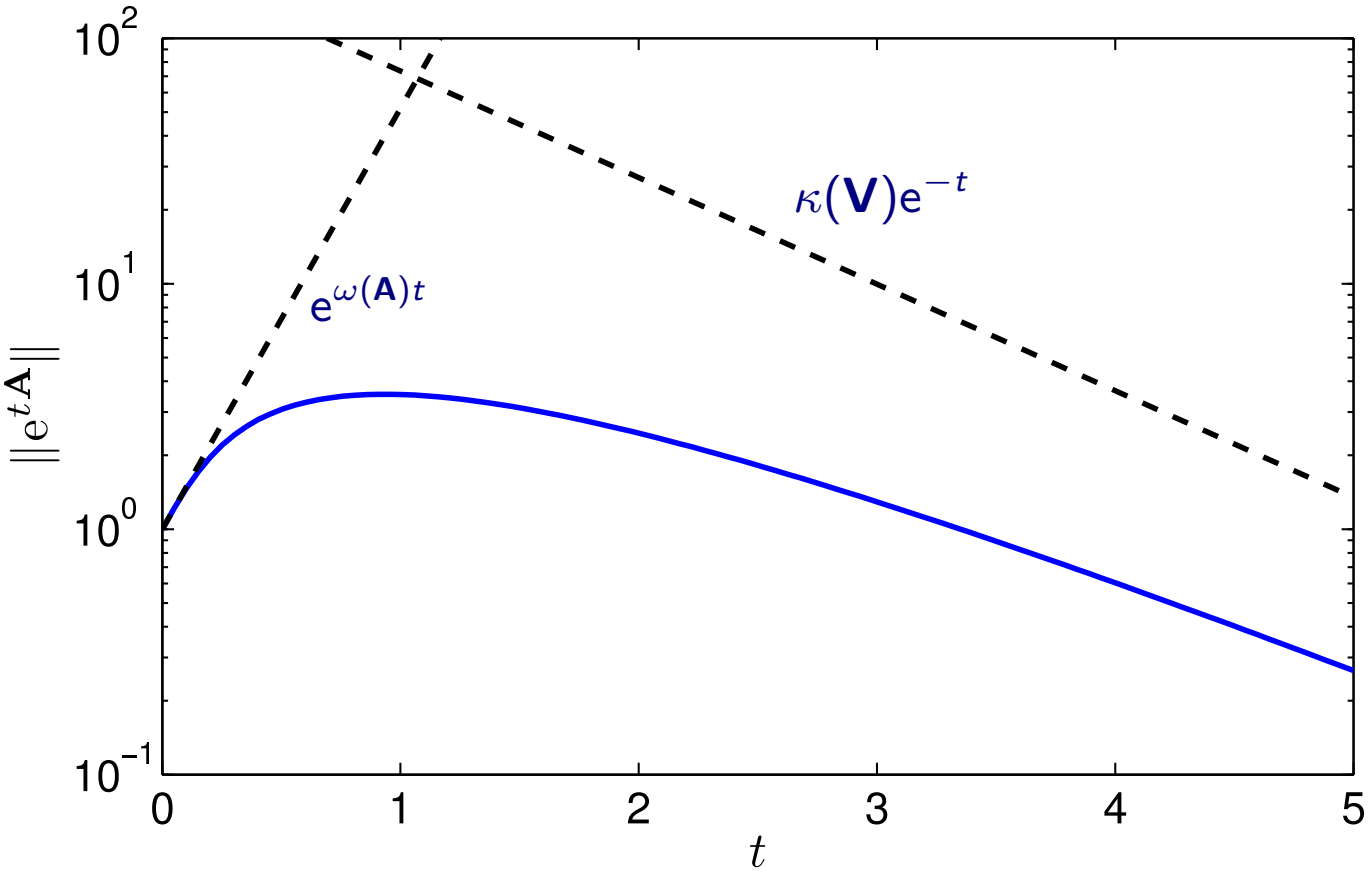
Definition (numerical abscissa)

The *numerical abscissa* is the rightmost in $W(\mathbf{A})$:

$$\omega(\mathbf{A}) := \max_{z\in W(\mathbf{A})} \operatorname{Re} z.$$

Initial Transient Growth via Numerical Abscissa

$$\mathbf{A} = \begin{bmatrix} -1.1 & 10 \\ 0 & -1 \end{bmatrix}.$$



Tools for Understanding Transient Growth: Pseudospectra

[Use the convention that if $\mathbf{A}$ does not have a bounded inverse, $\|\mathbf{A}^{-1}\| = \infty$.]

Theorem

The following three definitions of the ε-pseudospectrum are equivalent:

1. $\sigma_\varepsilon(\mathbf{A}) = \{z \in \boldsymbol{C} : z \in \sigma(\mathbf{A} + \mathbf{E})$ *for some bounded* $\mathbf{E}$ *with* $\|\mathbf{E}\| < \varepsilon\}$;
2. $\sigma_\varepsilon(\mathbf{A}) = \{z \in \boldsymbol{C} : \|(z - \mathbf{A})^{-1}\| > 1/\varepsilon\}$;
3. $\sigma_\varepsilon(\mathbf{A}) = \{z \in \boldsymbol{C} : z \in \sigma(\mathbf{A})$ *or* $\|\mathbf{A}\mathbf{v} - z\mathbf{v}\| < \varepsilon$ *for some unit vector* $\mathbf{v}\}$.

See, e.g., [Trefethen, E. 2005].

Tools for Understanding Transient Growth: Pseudospectra

[Use the convention that if $\mathbf{A}$ does not have a bounded inverse, $\|\mathbf{A}^{-1}\| = \infty$.]

Theorem

The following three definitions of the ε-pseudospectrum are equivalent:

1. $\sigma_\varepsilon(\mathbf{A}) = \{z \in \boldsymbol{C} : z \in \sigma(\mathbf{A} + \mathbf{E})$ *for some bounded* $\mathbf{E}$ *with* $\|\mathbf{E}\| < \varepsilon\}$;
2. $\sigma_\varepsilon(\mathbf{A}) = \{z \in \boldsymbol{C} : \|(z - \mathbf{A})^{-1}\| > 1/\varepsilon\}$;
3. $\sigma_\varepsilon(\mathbf{A}) = \{z \in \boldsymbol{C} : z \in \sigma(\mathbf{A})$ *or* $\|\mathbf{A}\mathbf{v} - z\mathbf{v}\| < \varepsilon$ *for some unit vector* $\mathbf{v}\}$.

See, e.g., [Trefethen, E. 2005].

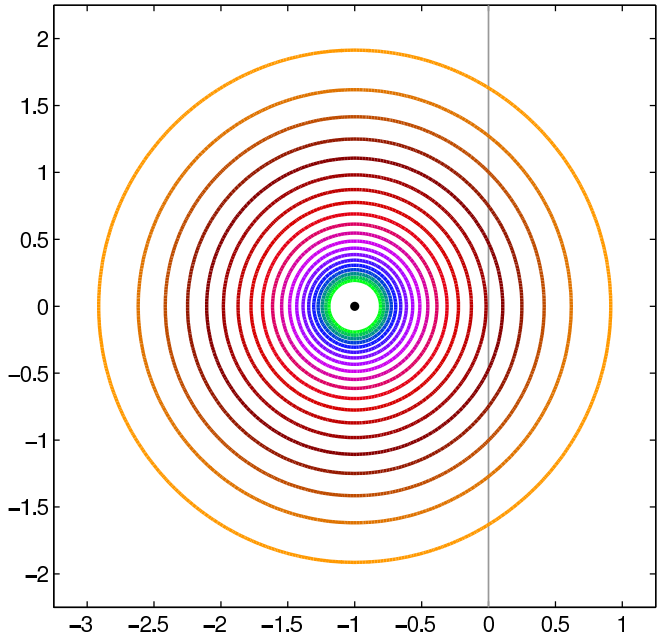
These different definitions are useful in different contexts:

1. interpreting numerically computed eigenvalues;
2. analyzing matrix behavior/functions of matrices;
   computing pseudospectra on a grid in $\mathbf{C}$;
3. proving bounds on a particular $\sigma_\varepsilon(\mathbf{A})$.

Example of Pseudospectra

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & & & \\ & -1 & \ddots & & \\ & & \ddots & 2 & \\ & & & -1 & 2 \\ & & & & -1 \end{bmatrix} \in \mathbf{C}^{20\times 20}.$$

Pseudospectra of Toeplitz matrices have been deeply studied [Böttcher et al.].



$\sigma_\varepsilon(\mathbf{A})$ for $\varepsilon = 10^{-20}, 10^{-19}, \ldots, 10^{-1}$

Pseudospectral Bounds on the Matrix Exponential

We wish to use pseudospectra to bound $\|e^{t\mathbf{A}}\|$ (cf. Hille–Yosida theory).

Definition

The *ε-pseudospectral abscissa* is the supremum of the real parts of $z \in \sigma_\varepsilon(\mathbf{A})$:

$$\alpha_\varepsilon(\mathbf{A}) := \sup_{z \in \sigma_\varepsilon(\mathbf{A})} \operatorname{Re} z.$$

Pseudospectral Bounds on the Matrix Exponential

We wish to use pseudospectra to bound $\|e^{t\mathbf{A}}\|$ (cf. Hille–Yosida theory).

Definition

The *ε-pseudospectral abscissa* is the supremum of the real parts of $z \in \sigma_\varepsilon(\mathbf{A})$:

$$\alpha_\varepsilon(\mathbf{A}) := \sup_{z \in \sigma_\varepsilon(\mathbf{A})} \operatorname{Re} z.$$

Theorem (Upper and Lower Bounds on $\|e^{t\mathbf{A}}\|$)

For any $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\varepsilon > 0$,

$$\|e^{t\mathbf{A}}\| \le \frac{L_\varepsilon}{2\pi\varepsilon} e^{t\alpha_\varepsilon(\mathbf{A})},$$

where L_ε denotes the contour length of the boundary of $\sigma_\varepsilon(\mathbf{A})$.

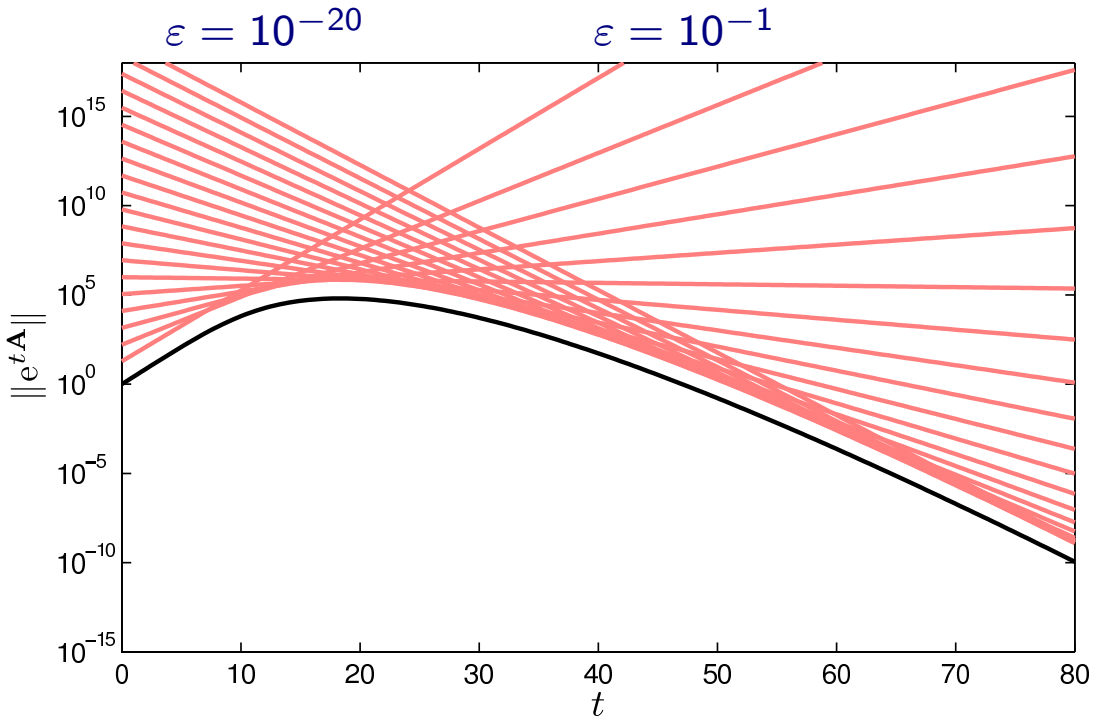
For stable $\mathbf{A}$ and any $\varepsilon > 0$,

$$\sup_{t\ge 0} \|e^{t\mathbf{A}}\| \ge \frac{\alpha_\varepsilon(\mathbf{A})}{\varepsilon}.$$

The supremum of the right hand side over all $\varepsilon > 0$ is the Kreiss constant.

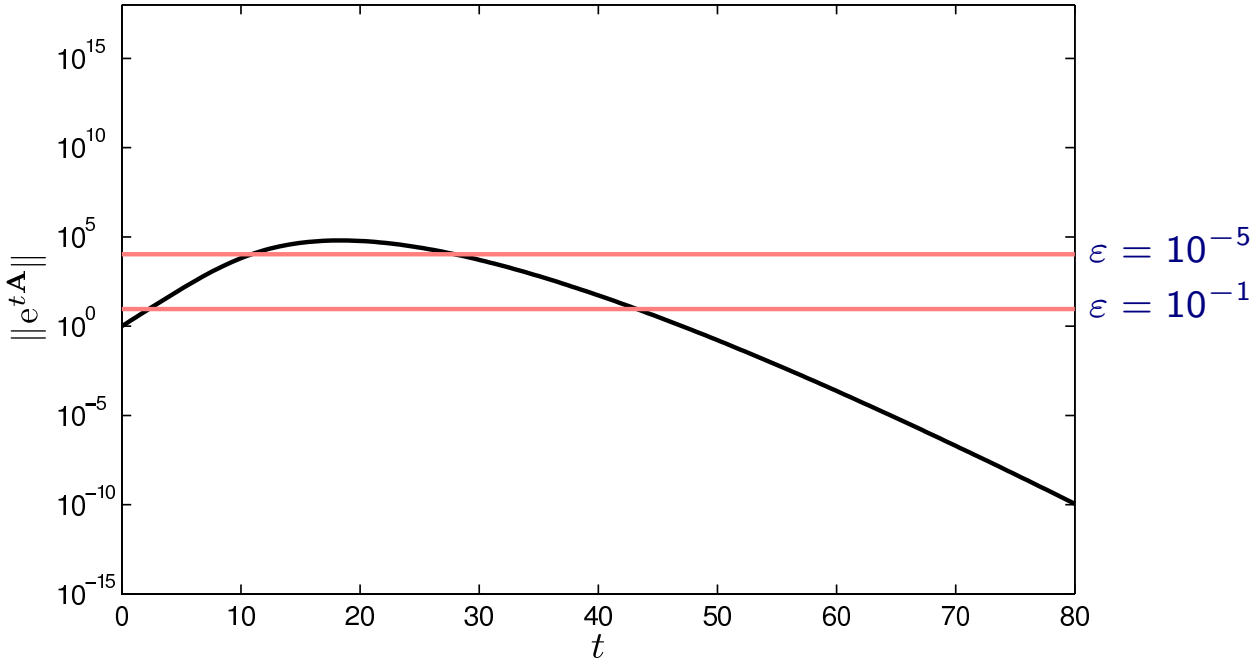
Upper Bound on the Matrix Exponential from Pseudospectra

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & & & \\ & -1 & \ddots & & \\ & & \ddots & 2 & \\ & & & -1 & 2 \\ & & & & -1 \end{bmatrix} \in \mathbf{C}^{20\times 20}.$$

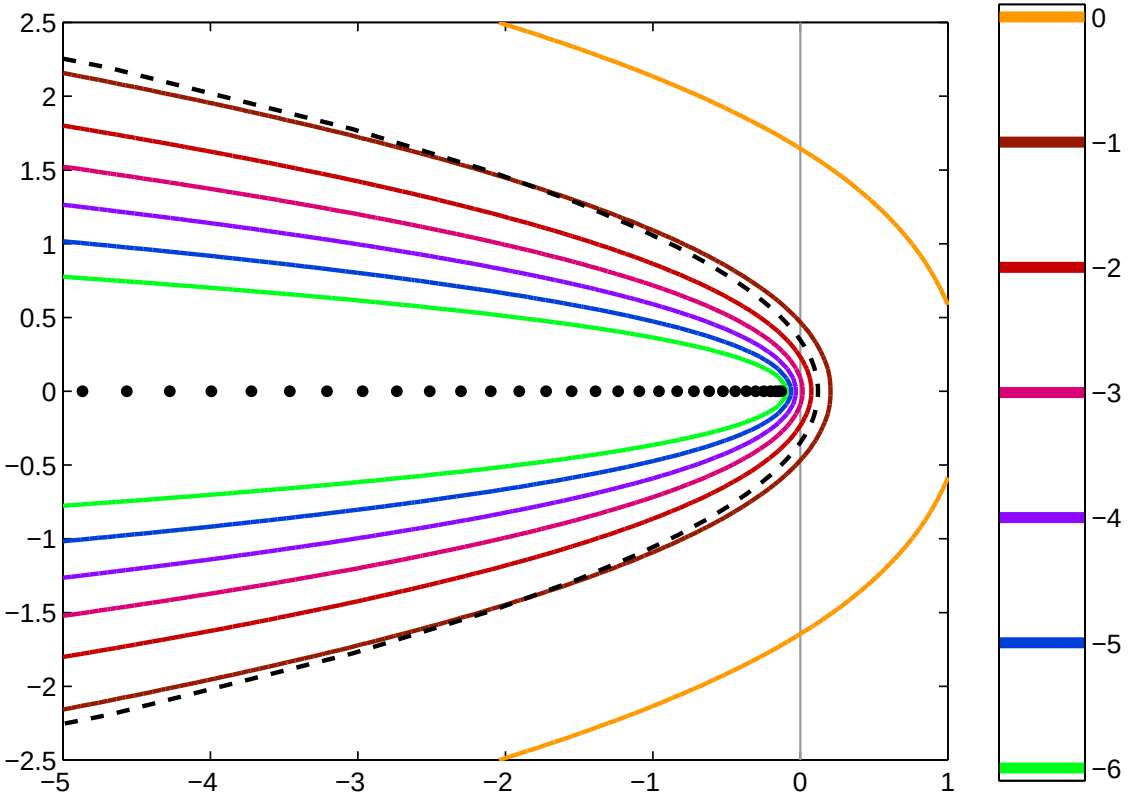


Lower Bound on the Matrix Exponential from Pseudospectra

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & & & \\ & -1 & \ddots & & \\ & & \ddots & 2 & \\ & & & -1 & 2 \\ & & & & -1 \end{bmatrix} \in \mathbf{C}^{20\times 20}.$$



Nonnormality in the Linearized PDE Example



Spectrum, pseudospectra, and numerical range (L^2 norm, $\nu = 0.02$)

Transient growth can feed the nonlinearity; cf. [Trefethen, Trefethen, Reddy, Driscoll 1993], [Baggett, Driscoll, Trefehen 1995]

Eigenvalues and Eigenvectors: Davies's Conjecture

Diagonalization and Matrix Behavior

For a diagonalizable matrix $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$, we can compute

$$f(\mathbf{A}) = \mathbf{V} f(\boldsymbol{\Lambda}) \mathbf{V}^{-1}$$

and bound

$$\begin{aligned} \|f(\mathbf{A})\| &= \|\mathbf{V} f(\boldsymbol{\Lambda}) \mathbf{V}^{-1}\| \\ &\leq \|\mathbf{V}\| \|\mathbf{V}^{-1}\| \max_{\lambda \in \sigma(\mathbf{A})} |f(\lambda)|. \end{aligned}$$

Here $\|\mathbf{V}\| \|\mathbf{V}^{-1}\|$ is called the *condition number of* $\mathbf{V}$.
It reflects the secants of the angles between right and left eigenvectors:

$\mathbf{v}_j = j$th column of $\mathbf{V}$ so that $\mathbf{A}\mathbf{v}_j = \mathbf{v}_j \lambda_j$
$\widehat{\mathbf{v}}_j^* = j$th row of $\mathbf{V}^{-1}$ so that $\widehat{\mathbf{v}}_j^* \mathbf{A} = \lambda_j \widehat{\mathbf{v}}_j^*$

$$\sec(\mathbf{v}_j, \widehat{\mathbf{v}}_j) = \frac{\|\widehat{\mathbf{v}}_j\| \|\mathbf{v}_j\|}{|\widehat{\mathbf{v}}_j^* \mathbf{v}_j|} = \|\widehat{\mathbf{v}}_j\| \|\mathbf{v}_j\|.$$

Diagonalization and Matrix Behavior

For a diagonalizable matrix $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$, we can compute

$$f(\mathbf{A}) = \mathbf{V} f(\boldsymbol{\Lambda}) \mathbf{V}^{-1}$$

and bound

$$\begin{aligned} \|f(\mathbf{A})\| &= \|\mathbf{V} f(\boldsymbol{\Lambda}) \mathbf{V}^{-1}\| \\ &\leq \|\mathbf{V}\| \|\mathbf{V}^{-1}\| \max_{\lambda \in \sigma(\mathbf{A})} |f(\lambda)|. \end{aligned}$$

Here $\|\mathbf{V}\| \|\mathbf{V}^{-1}\|$ is called the *condition number of* $\mathbf{V}$.
It reflects the secants of the angles between right and left eigenvectors:

$\mathbf{v}_j = j$th column of $\mathbf{V}$ so that $\mathbf{A}\mathbf{v}_j = \mathbf{v}_j \lambda_j$
$\widehat{\mathbf{v}}_j^* = j$th row of $\mathbf{V}^{-1}$ so that $\widehat{\mathbf{v}}_j^* \mathbf{A} = \lambda_j \widehat{\mathbf{v}}_j^*$

$$\sec(\mathbf{v}_j, \widehat{\mathbf{v}}_j) = \frac{\|\widehat{\mathbf{v}}_j\| \|\mathbf{v}_j\|}{|\widehat{\mathbf{v}}_j^* \mathbf{v}_j|} = \|\widehat{\mathbf{v}}_j\| \|\mathbf{v}_j\|.$$

For many non-self-adjoint matrices, $\|\mathbf{V}\| \|\mathbf{V}^{-1}\|$ can be so large as to render this method of computing $f(\mathbf{A})$ and the corresponding bound useless.

Diagonalization and Matrix Behavior

Consider the $n \times n$ Toeplitz matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1/2 & & \\ -2 & 0 & \ddots & \\ & \ddots & \ddots & 1/2 \\ & & -2 & 0 \end{bmatrix},$$

which has n distinct eigenvalues (purely imaginary) and hence is always diagonalizable, $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$.

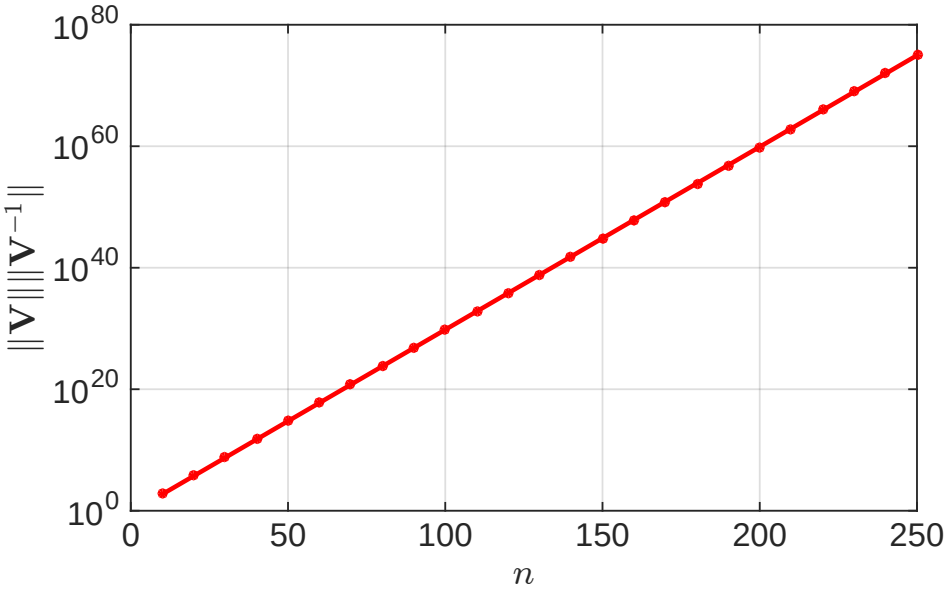
Diagonalization and Matrix Behavior

Consider the $n \times n$ Toeplitz matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1/2 & & \\ -2 & 0 & \ddots & \\ & \ddots & \ddots & 1/2 \\ & & -2 & 0 \end{bmatrix},$$

which has n distinct eigenvalues (purely imaginary) and hence is always diagonalizable, $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$.

$$\|\mathbf{V}\|\,\|\mathbf{V}^{-1}\| = 2^{n-1}$$



$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$ cannot be computed with any numerical accuracy, so cannot be used to compute $f(\mathbf{A})$; cf. Moler & Van Loan "Nineteen Dubious Ways to Compute the Exponential of a Matrix", 1978.

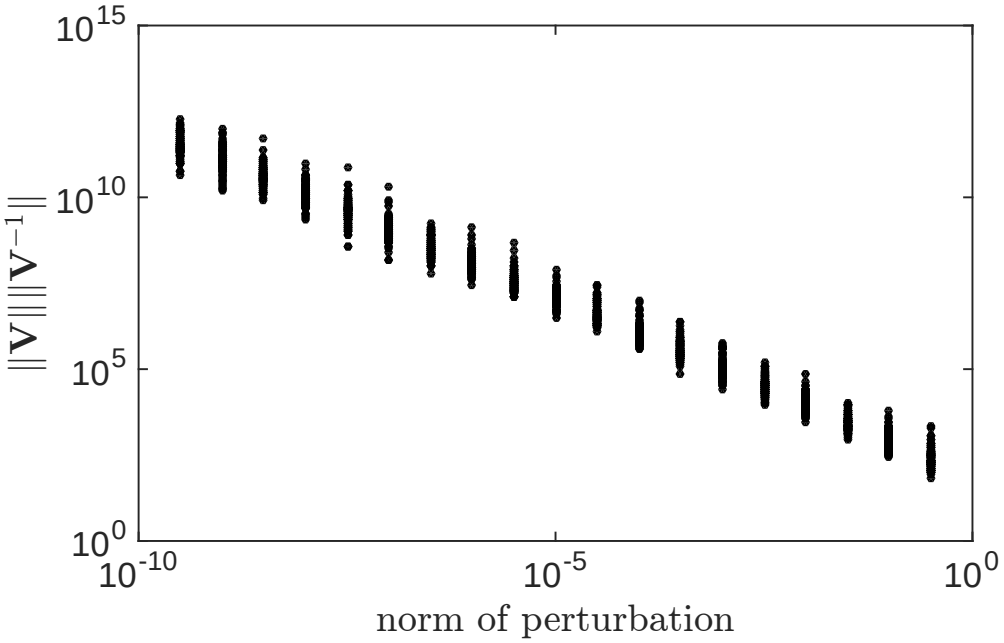
Diagonalization and Matrix Behavior

For the previous example with $n = 100$, the unperturbed problem gave $\|\mathbf{V}\|\|\mathbf{V}^{-1}\| = 2^{99} \approx 6.34 \times 10^{29}$.

However, $\mathbf{A}$ is often *close to a matrix with a much better conditioned eigenvector matrix*.

Diagonalize $\mathbf{A} + \mathbf{\Delta} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ for random perturbations of size $\mathbf{\Delta}$.

Compute 100 random diagonalizations for each perturbation size.



Cf. [Spielman, Teng 2004], [Tao, Vu 2009] on conditioning of randomly perturbed matrices.

Davies's Conjecture about "Approximate Diagonalization"

SIAM J. MATRIX ANAL. APPL.
Vol. 29, No. 4, pp. 1051–1064

APPROXIMATE DIAGONALIZATION*

E. B. DAVIES†

Davies formalized this idea in a 2007 article.

Given any $\mathbf{A} \in \mathbf{C}^{n \times n}$:

- Pick a desired *precision* $\varepsilon > 0$;
- Pick a *perturbation* $\mathbf{\Delta} \in \mathbf{C}^{n \times n}$;
- Diagonalize $\mathbf{A} + \mathbf{\Delta} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$;
- Measure the quality of the *approximate diagonalization* by

$$\|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\|,$$

  which balances the quality of the diagonalization against size of the perturbation.

Because different diagonalizations yield different values of $\|\mathbf{V}\|\|\mathbf{V}^{-1}\|$, define

$$s(\mathbf{A}, \mathbf{\Delta}, \varepsilon) = \inf_{\substack{\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\|$$

Davies's Conjecture about "Approximate Diagonalization"

Given

$$s(\mathbf{A}, \mathbf{\Delta}, \varepsilon) = \inf_{\substack{\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\|,$$

one might naturally ask:

For a fixed precision ε, what perturbation $\mathbf{\Delta}$ minimizes $s(\mathbf{A}, \mathbf{\Delta}, \varepsilon)$?

Davies's Conjecture about "Approximate Diagonalization"

Given

$$s(\mathbf{A}, \mathbf{\Delta}, \varepsilon) = \inf_{\substack{\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\|,$$

one might naturally ask:

For a fixed precision ε, what perturbation $\mathbf{\Delta}$ minimizes $s(\mathbf{A}, \mathbf{\Delta}, \varepsilon)$?

$$\underline{s}(\mathbf{A}, \varepsilon) = \inf_{\mathbf{\Delta}} s(\mathbf{A}, \mathbf{\Delta}, \varepsilon)$$

Davies's Conjecture about "Approximate Diagonalization"

Given

$$s(\mathbf{A}, \mathbf{\Delta}, \varepsilon) = \inf_{\substack{\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\|,$$

one might naturally ask:

For a fixed precision ε, what perturbation $\mathbf{\Delta}$ minimizes $s(\mathbf{A}, \mathbf{\Delta}, \varepsilon)$?

$$\underline{s}(\mathbf{A}, \varepsilon) = \inf_{\mathbf{\Delta}} s(\mathbf{A}, \mathbf{\Delta}, \varepsilon)$$

Conjecture (Davies, 2007)

Given a matrix dimension n, there exists a constant C_n such that

$$\underline{s}(\mathbf{A}, \varepsilon) \leq C_n \sqrt{\varepsilon}$$

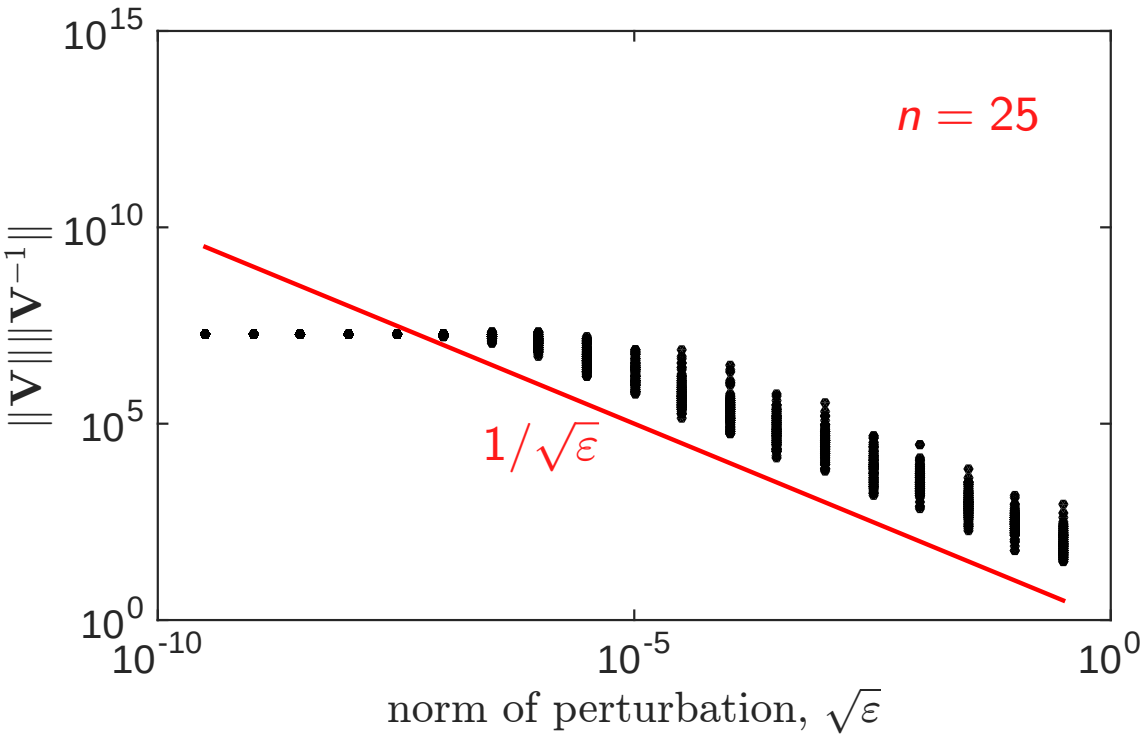
for all $\mathbf{A} \in \mathbb{C}^{n \times n}$ and all $\varepsilon \in (0, 1)$.

Davies's Conjecture about "Approximate Diagonalization"

Conjecture: for any n there exists C_n such that for all $\mathbf{A} \in \mathbf{C}^{n\times n}$ and $\varepsilon \in (0,1)$,

$$\underline{s}(\mathbf{A},\varepsilon) = \inf_{\substack{\mathbf{\Delta},\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\| \le C_n\sqrt{\varepsilon}.$$

For the Toeplitz matrix example, perturbations $\mathbf{\Delta}$ of size $\|\mathbf{\Delta}\| = \sqrt{\varepsilon}$ give diagonalizations $\mathbf{A}+\mathbf{\Delta} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ with $\|\mathbf{V}\|\|\mathbf{V}^{-1}\| \sim 1/\sqrt{\varepsilon}$.

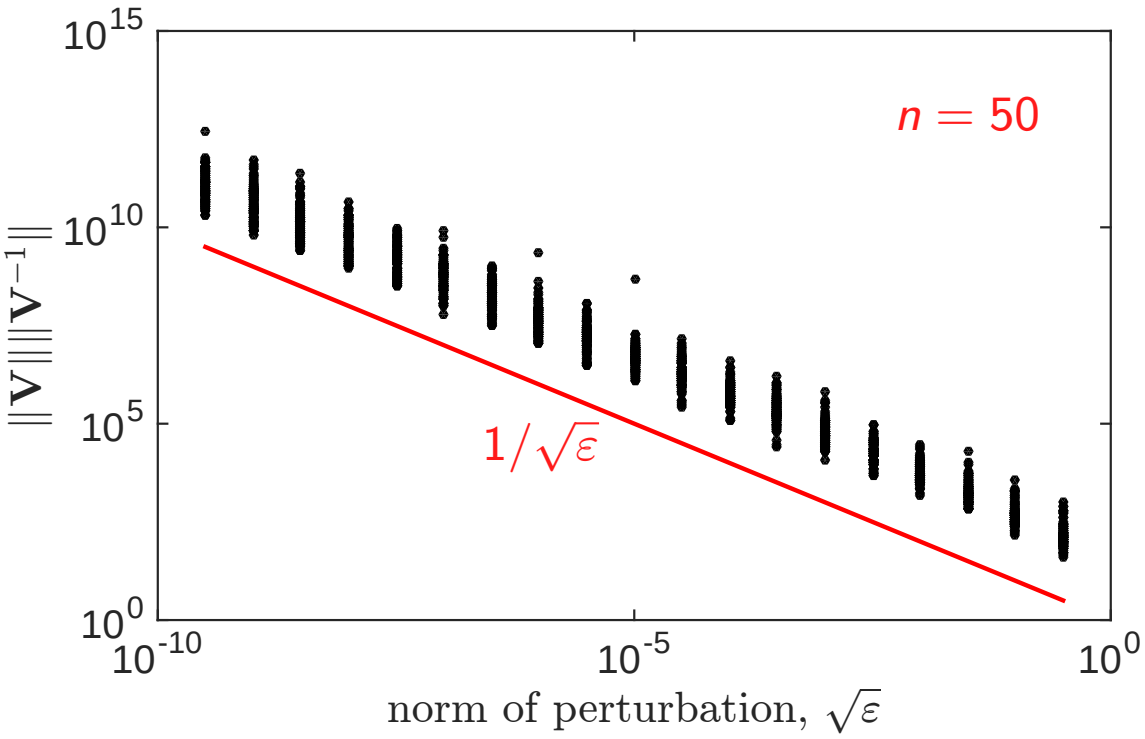


Davies's Conjecture about "Approximate Diagonalization"

Conjecture: for any n there exists C_n such that for all $\mathbf{A} \in \mathbf{C}^{n\times n}$ and $\varepsilon \in (0,1)$,

$$\underline{s}(\mathbf{A},\varepsilon) = \inf_{\substack{\mathbf{\Delta},\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\| \le C_n\sqrt{\varepsilon}.$$

For the Toeplitz matrix example, perturbations $\mathbf{\Delta}$ of size $\|\mathbf{\Delta}\| = \sqrt{\varepsilon}$ give diagonalizations $\mathbf{A}+\mathbf{\Delta} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ with $\|\mathbf{V}\|\|\mathbf{V}^{-1}\| \sim 1/\sqrt{\varepsilon}$.

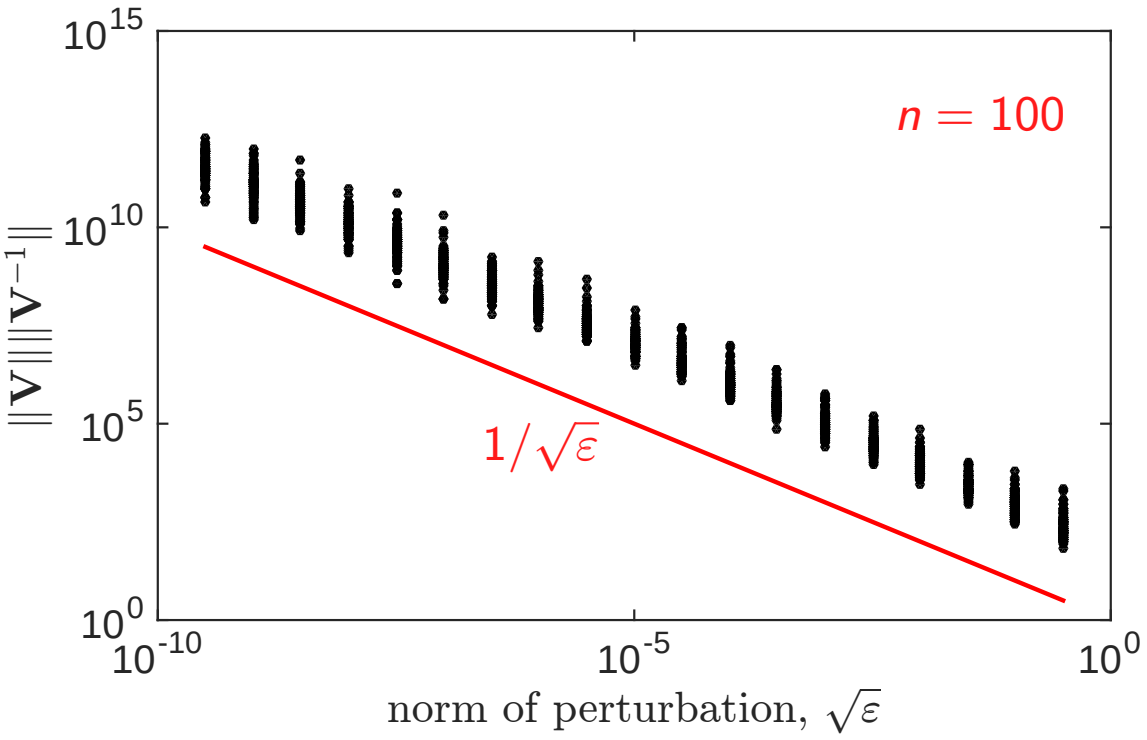


Davies's Conjecture about "Approximate Diagonalization"

Conjecture: for any n there exists C_n such that for all $\mathbf{A} \in \mathbf{C}^{n\times n}$ and $\varepsilon \in (0,1)$,

$$\underline{s}(\mathbf{A},\varepsilon) = \inf_{\substack{\mathbf{\Delta},\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\| \le C_n\sqrt{\varepsilon}.$$

For the Toeplitz matrix example, perturbations $\mathbf{\Delta}$ of size $\|\mathbf{\Delta}\| = \sqrt{\varepsilon}$ give diagonalizations $\mathbf{A}+\mathbf{\Delta} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ with $\|\mathbf{V}\|\|\mathbf{V}^{-1}\| \sim 1/\sqrt{\varepsilon}$.

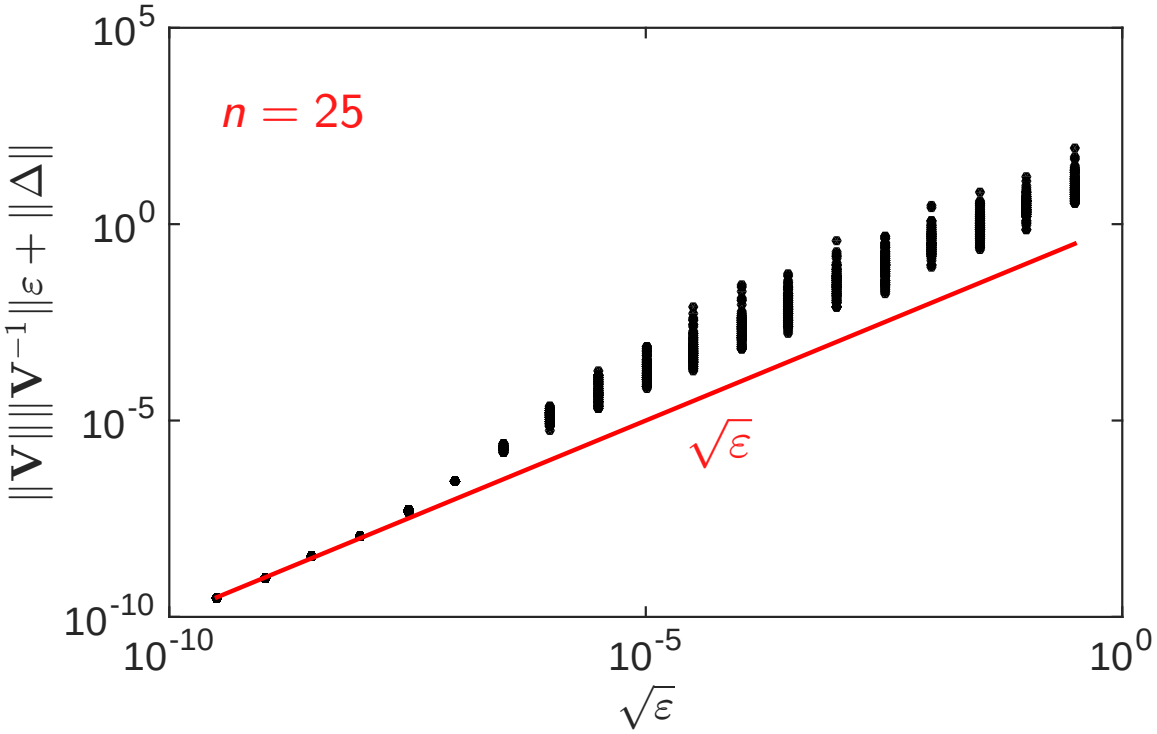


Davies's Conjecture about "Approximate Diagonalization"

Conjecture: for any n there exists C_n such that for all $\mathbf{A} \in \mathbf{C}^{n\times n}$ and $\varepsilon \in (0,1)$,

$$\underline{s}(\mathbf{A},\varepsilon) = \inf_{\substack{\mathbf{\Delta},\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\| \le C_n\sqrt{\varepsilon}.$$

Testing the conjecture with 100 trials per ε, fixing $\|\mathbf{\Delta}\| = \sqrt{\varepsilon}$.

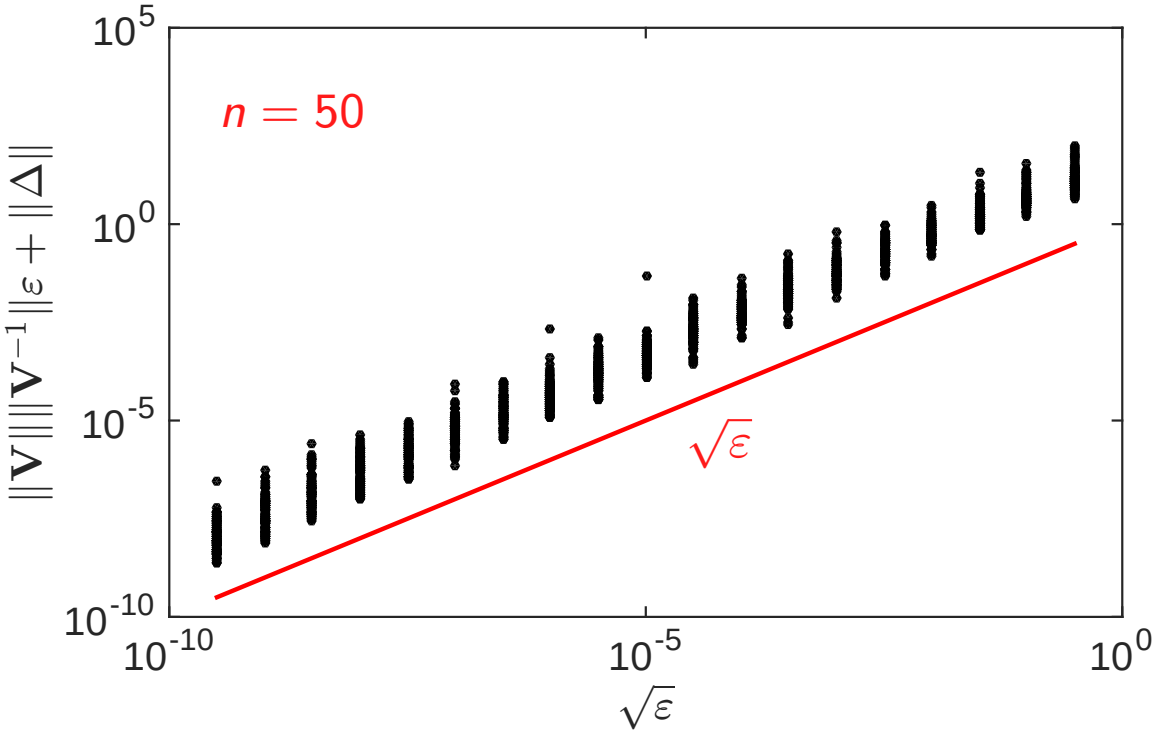


Davies's Conjecture about "Approximate Diagonalization"

Conjecture: for any n there exists C_n such that for all $\mathbf{A} \in \mathbf{C}^{n\times n}$ and $\varepsilon \in (0,1)$,

$$\underline{s}(\mathbf{A},\varepsilon) = \inf_{\substack{\mathbf{\Delta},\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\| \le C_n\sqrt{\varepsilon}.$$

Testing the conjecture with 100 trials per ε, fixing $\|\mathbf{\Delta}\| = \sqrt{\varepsilon}$.

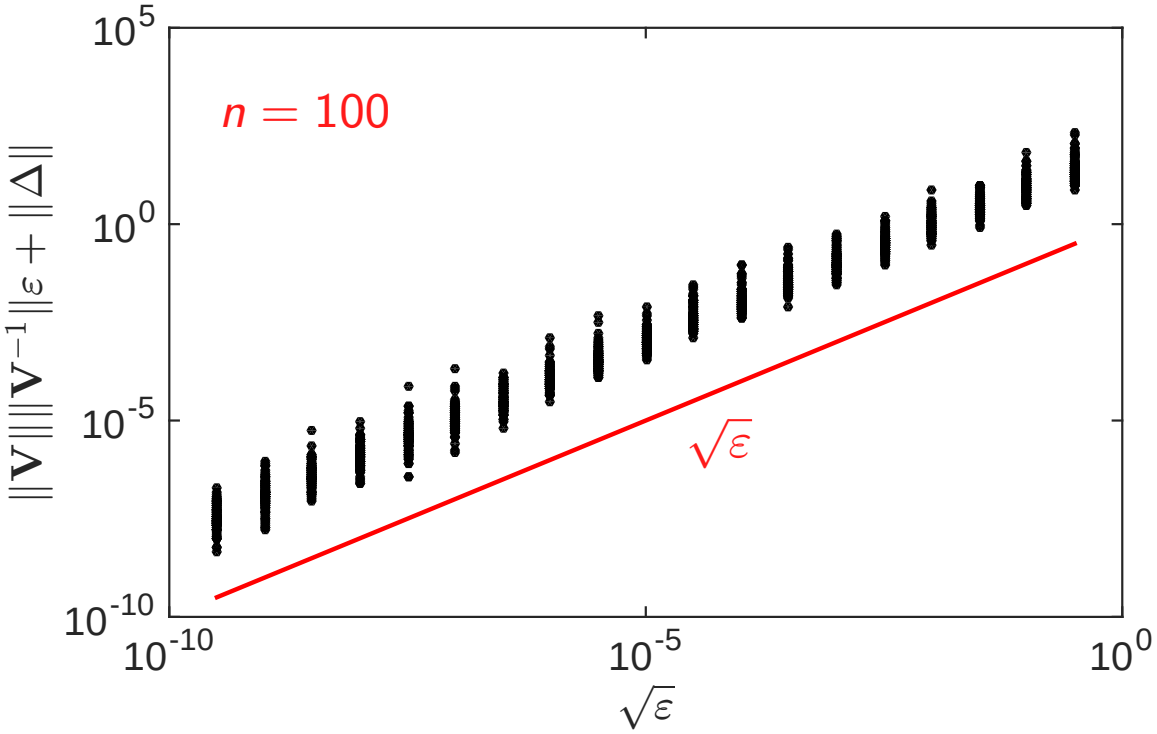


Davies's Conjecture about "Approximate Diagonalization"

Conjecture: for any n there exists C_n such that for all $\mathbf{A} \in \mathbf{C}^{n\times n}$ and $\varepsilon \in (0,1)$,

$$\underline{s}(\mathbf{A},\varepsilon) = \inf_{\substack{\mathbf{\Delta},\mathbf{V} \\ \mathbf{V}^{-1}(\mathbf{A}+\mathbf{\Delta})\mathbf{V} \text{ diag}}} \|\mathbf{V}\|\|\mathbf{V}^{-1}\|\varepsilon + \|\mathbf{\Delta}\| \le C_n\sqrt{\varepsilon}.$$

Testing the conjecture with 100 trials per ε, fixing $\|\mathbf{\Delta}\| = \sqrt{\varepsilon}$.



Davies's Conjecture: partial results and applications

Davies has proved the conjecture for several classes of matrices:

- Jordan blocks (ones on the superdiagonal): $C_n = 2$ suffices.
- 3×3 matrices with $\|\mathbf{A}\| \leq 1$: $C_n = 4$ suffices.
- For any $\mathbf{A}$, there exists $C_{\mathbf{A}}$ such that $\underline{s}(\mathbf{A}, \varepsilon) \leq C_{\mathbf{A}}\sqrt{\varepsilon}$, but it is not known if $C_{\mathbf{A}}$ can be bounded independent of $\mathbf{A}$.

Applications of this result:

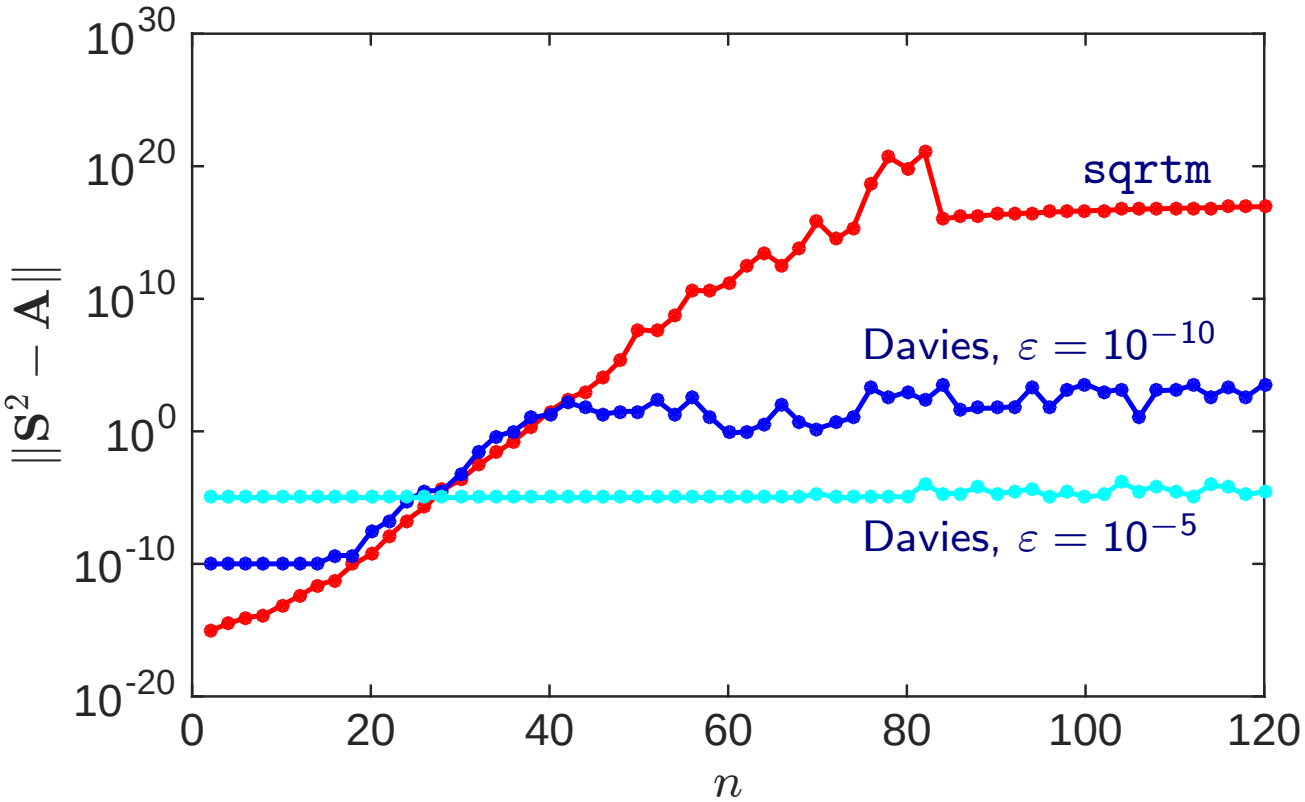
- Approximate $f(\mathbf{A})$ by $f(\mathbf{A} + \mathbf{\Delta})$ for small ε. Davies then uses perturbation theory for the resolvent (see [Rinehart, 1956]) to bound $\|f(\mathbf{A}) - f(\mathbf{A} + \mathbf{\Delta})\|$. Davies shows examples for computing fractional powers of matrices. For matrices with a highly ill-conditioned eigenvector matrix, one can obtain more accurate approximations to $f(\mathbf{A})$ by diagonalizing $\mathbf{A} + \mathbf{\Delta}$ rather than $\mathbf{A}$ itself!
- Potential applications to convergence theory of GMRES for $\mathbf{Ax} = \mathbf{b}$; cf. [Sifuentes, Embree, Morgan, 2013].

Davies's Conjecture: approximating the matrix square root

For the Toeplitz example, compute Davies's approximation:

$$\mathbf{A} + \mathbf{D} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}, \qquad \mathbf{A}^{1/2} \approx \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{-1} =: \mathbf{S}.$$

Compare these results to those produced by MATLAB's `sqrtm` command (a Schur-based algorithm due to Björck, Hammarling, and Higham).



(All these methods have large forward error for large n: $\|\mathbf{S} - \mathbf{A}^{1/2}\| \gg 1$.)

Numerical Range: Crouzeix's Conjecture

Numerical range

The *numerical range* (*field of values*) of $\mathbf{A} \in \mathbf{C}^{n \times n}$,

$$W(\mathbf{A}) \quad = \quad \{\mathbf{v}^* \mathbf{A} \mathbf{v} : \|\mathbf{v}\| = 1\},$$

is a convex subset of the complex plane that contains the spectrum.

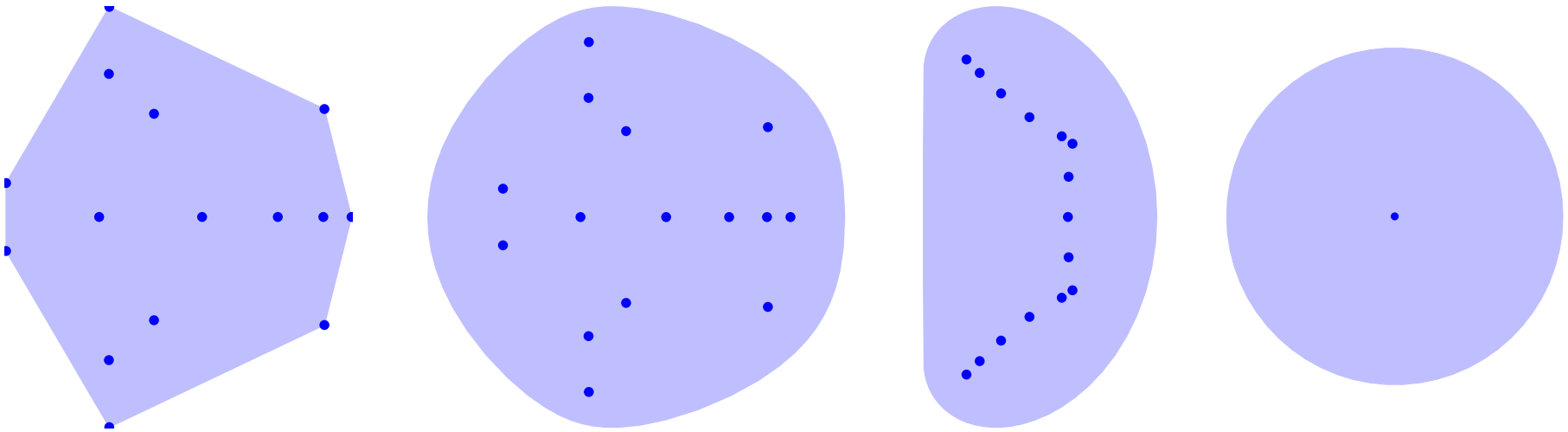
Numerical range

The *numerical range* (*field of values*) of $\mathbf{A} \in \mathbf{C}^{n \times n}$,

$$W(\mathbf{A}) \quad = \quad \{\mathbf{v}^* \mathbf{A} \mathbf{v} : \|\mathbf{v}\| = 1\},$$

is a convex subset of the complex plane that contains the spectrum.

Eigenvalues and the numerical range, for four different 15×15 matrices.



normal random Grcar Jordan

Bounding behavior of matrices with the numerical range

The largest magnitude of a point in $W(\mathbf{A})$ is the *numerical radius*,

$$\mu(\mathbf{A}) := \max_{z \in W(\mathbf{A})} |z|.$$

The numerical radius is bounded by the operator norm (e.g., [Halmos, 1982]):

$$\frac{1}{2}\|\mathbf{A}\| \leq \mu(\mathbf{A}) \leq \|\mathbf{A}\|.$$

Bounding behavior of matrices with the numerical range

The largest magnitude of a point in $W(\mathbf{A})$ is the *numerical radius*,

$$\mu(\mathbf{A}) := \max_{z \in W(\mathbf{A})} |z|.$$

The numerical radius is bounded by the operator norm (e.g., [Halmos, 1982]):

$$\frac{1}{2}\|\mathbf{A}\| \leq \mu(\mathbf{A}) \leq \|\mathbf{A}\|.$$

Theorem (Berger, "A strange dilation theorem," 1965; Pearcy, 1966)

For any $\mathbf{A} \in \mathbb{C}^{n \times n}$ $\mu(\mathbf{A}^k) \leq (\mu(\mathbf{A}))^k$ *and hence*

$$\|\mathbf{A}^k\| \leq 2(\mu(\mathbf{A}))^k.$$

Bounding behavior of matrices with the numerical range

The largest magnitude of a point in $W(\mathbf{A})$ is the *numerical radius*,

$$\mu(\mathbf{A}) := \max_{z \in W(\mathbf{A})} |z|.$$

The numerical radius is bounded by the operator norm (e.g., [Halmos, 1982]):

$$\frac{1}{2}\|\mathbf{A}\| \leq \mu(\mathbf{A}) \leq \|\mathbf{A}\|.$$

Theorem (Berger, "A strange dilation theorem," 1965; Pearcy, 1966)

For any $\mathbf{A} \in \mathbf{C}^{n \times n}$ $\mu(\mathbf{A}^k) \leq (\mu(\mathbf{A}))^k$ *and hence*

$$\|\mathbf{A}^k\| \leq 2(\mu(\mathbf{A}))^k.$$

Berger's bound can be much better than the conventional bound $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k$, and the constant is sharp: for

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

and $k = 1$ we have

$$W(\mathbf{A}) = \{z \in \mathbf{C} : |z| \leq 1/2\}, \qquad \|\mathbf{A}\| = 1, \qquad \mu(\mathbf{A}) = 1/2.$$

Bounding behavior of matrices with the numerical range

$$\|\mathbf{A}^k\| \leq 2(\mu(\mathbf{A}))^k = 2 \max_{z \in W(\mathbf{A})} |z^k|.$$

Does this result generalize from $f(z) = z^k$ to general f analytic on $W(\mathbf{A})$?

Bounding behavior of matrices with the numerical range

$$\|\mathbf{A}^k\| \leq 2(\mu(\mathbf{A}))^k = 2 \max_{z \in W(\mathbf{A})} |z^k|.$$

Does this result generalize from $f(z) = z^k$ to general f analytic on $W(\mathbf{A})$?

Partial generalizations exist if one is willing to enlarge $W(\mathbf{A})$ to Ω. Following *Von Neumann's theory of spectral sets*, we seek $K \geq 1$, $\Omega \supseteq W(\mathbf{A})$ such that

$$\|f(\mathbf{A})\| \leq K \sup_{z \in \Omega} |f(z)|.$$

see, e.g., [Paulsen, 2002], [Badea, Crouzeix, Delyon, 2006]. Partial results for disks, sectors, etc. (see [Badea, Crouzeix, Delyon, 2006]), ellipses [Eiermann, 1993].

Crouzeix's Conjecture about $\|f(\mathbf{A})\|$

Journal of Functional Analysis 244 (2007) 668–690

Numerical range and functional calculus in Hilbert space

Michel Crouzeix

Theorem

For any matrix $\mathbf{A} \in C^{n\times n}$ *and any* f *analytic on* $W(\mathbf{A})$,

$$\|f(\mathbf{A})\| \leq C \max_{z\in W(\mathbf{A})} |f(z)|,$$

where $2 \leq C \leq 11.08$.

Crouzeix's Conjecture about $\|f(\mathbf{A})\|$

Journal of Functional Analysis 244 (2007) 668–690

Numerical range and functional calculus in Hilbert space

Michel Crouzeix

Theorem

For any matrix $\mathbf{A} \in C^{n\times n}$ *and any* f *analytic on* $W(\mathbf{A})$,

$$\|f(\mathbf{A})\| \leq C \max_{z\in W(\mathbf{A})} |f(z)|,$$

where $2 \leq C \leq 11.08$.

- Crouzeix's Conjecture: $C = 2$.
- N.B. One only needs to prove this for polynomials of degree $< n$.

Crouzeix's Conjecture: partial results

Theorem

For any matrix $\mathbf{A} \in C^{n\times n}$ *and any* f *analytic on* $W(\mathbf{A})$,

$$\|f(\mathbf{A})\| \leq C \max_{z\in W(\mathbf{A})} |f(z)|,$$

where $2 \leq C \leq 11.08$.

- The conjecture holds when $\mathbf{A}$ is normal (could take $C = 1$).
- The conjecture holds when $n = 2$ (Crouzeix).
- The conjecture holds when $W(\mathbf{A})$ is a disk (Badea).
- All numerical evidence suggests that the conjecture holds.

Crouzeix's Conjecture: progress of Greenbaum and Choi

- The approach of [Greenbaum, Choi, 2012], [Choi, 2013].

  Greenbaum and Choi outline a general means of attack:
  — Construct a bijective conformal map g from $W(\mathbf{A})$ to the unit disk.
  — Write $g(\mathbf{A})$ as a similarity transformation of a contraction:

$$g(\mathbf{A}) = \mathbf{V}\mathbf{C}\mathbf{V}^{-1}, \qquad \|\mathbf{C}\| \leq 1.$$

  — If $\|\mathbf{V}\|\,\|\mathbf{V}^{-1}\| \leq 2$, then Crouzeix's conjecture holds.

Crouzeix's Conjecture: progress of Greenbaum and Choi

- The approach of [Greenbaum, Choi, 2012], [Choi, 2013].

  Greenbaum and Choi outline a general means of attack:
  — Construct a bijective conformal map g from $W(\mathbf{A})$ to the unit disk.
  — Write $g(\mathbf{A})$ as a similarity transformation of a contraction:

$$g(\mathbf{A}) = \mathbf{V}\mathbf{C}\mathbf{V}^{-1}, \qquad \|\mathbf{C}\| \leq 1.$$

  — If $\|\mathbf{V}\|\|\mathbf{V}^{-1}\| \leq 2$, then Crouzeix's conjecture holds.

  Using this technique, [Greenbaum, Choi, 2012] ($\alpha_1 = \cdots = \alpha_{n-1} = 1$) and [Choi, 2013] (general case) prove that Crouzeix's conjecture holds for all matrices of the form

$$\mathbf{A} = \begin{bmatrix} \lambda & \alpha_1 & & \\ & \lambda & \ddots & \\ & & \ddots & \alpha_{n-1} \\ \alpha_n & & & \lambda \end{bmatrix},$$

  for any $\lambda, \alpha_1, \ldots, \alpha_n \in \mathbf{C}$.

Numerical Range:
Inverse Problems for Ritz Values

Eigenvalue Approximations from Ritz values

Many techniques for approximating eigenvalues (theory and algorithms) use Rayleigh quotient estimates,

$$\frac{\langle \mathbf{A}\mathbf{v}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} = \frac{\mathbf{v}^* \mathbf{A} \mathbf{v}}{\mathbf{v}^* \mathbf{v}} \in W(\mathbf{A}).$$

The numerical range is simply the union of all such eigenvalue estimates.

We can generalize this to approximate a collection of eigenvalues (and the associated invariant subspace), as in the Ritz–Galerkin method, and subspace eigenvalue algorithms (Lanczos, Arnoldi, Jacobi–Davidson, etc.).

Eigenvalue Approximations from Ritz values

Many techniques for approximating eigenvalues (theory and algorithms) use Rayleigh quotient estimates,

$$\frac{\langle \mathbf{A}\mathbf{v}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} = \frac{\mathbf{v}^* \mathbf{A} \mathbf{v}}{\mathbf{v}^* \mathbf{v}} \in W(\mathbf{A}).$$

The numerical range is simply the union of all such eigenvalue estimates.

We can generalize this to approximate a collection of eigenvalues (and the associated invariant subspace), as in the Ritz–Galerkin method, and subspace eigenvalue algorithms (Lanczos, Arnoldi, Jacobi–Davidson, etc.).

- The columns of $\mathbf{V} \in \mathbf{C}^{n \times p}$ form an orthonormal basis for the subspace $\mathcal{V}$.
- $\mathbf{V}^* \mathbf{A} \mathbf{V}$ is called a *generalized Rayleigh quotient* and (we hope) $\sigma(\mathbf{V}^* \mathbf{A} \mathbf{V})$ approximates some subset of $\sigma(\mathbf{A})$.
- Note that all the points $\sigma(\mathbf{V}^* \mathbf{A} \mathbf{V})$ are Ritz values:

$$\mathbf{V}^* \mathbf{A} \mathbf{V} \mathbf{z} = \theta \mathbf{z} \quad \Longrightarrow \quad \theta = \frac{(\mathbf{V}\mathbf{z})^* \mathbf{A} (\mathbf{V}\mathbf{z})}{(\mathbf{V}\mathbf{z})^* (\mathbf{V}\mathbf{z})} \in W(\mathbf{A}).$$

- How do the eigenvalues of $\mathbf{V}^* \mathbf{A} \mathbf{V}$ distribute themselves across $W(\mathbf{A})$?

Inverse Field of Values Problem: single Ritz value

Uhlig [2008] introduced the "inverse field of values problem".

Problem (iFOV)

Suppose $A \in \boldsymbol{C}^{n\times n}$ and $\theta \in W(\mathbf{A})$.

- *Find a generating vector $\mathbf{v} \in \boldsymbol{C}^n$ for θ:*

$$\theta = \mathbf{v}^*\mathbf{A}\mathbf{v}, \qquad \|\mathbf{v}\| = 1.$$

- *How many such linearly independent generating vectors exist for θ?*

Inverse Field of Values Problem: single Ritz value

Uhlig [2008] introduced the "inverse field of values problem".

Problem (iFOV)

Suppose $A \in C^{n\times n}$ and $\theta \in W(\mathbf{A})$.

- *Find a generating vector $\mathbf{v} \in C^n$ for θ:*

$$\theta = \mathbf{v}^*\mathbf{A}\mathbf{v}, \qquad \|\mathbf{v}\| = 1.$$

- *How many such linearly independent generating vectors exist for θ?*

- Uhlig [2008] proposed a randomized algorithm to solve this problem.
- Carden [2009] gave a simpler method based on exact solvability for the $n = 2$ case. For θ in the interior of $W(\mathbf{A})$, it converges exactly in finitely many steps and there are n linearly independent generating vectors.
- Carden's algorithm refined by [Chorianopoulos, Psarrakos, Uhlig 2010].
- Applications?

Inverse Field of Values Problem: multiple Ritz values

The generalization of this problem to larger subspaces is much more difficult.

Problem (iFOV(p))

Given $A \in C^{n\times n}$ and p points

$$\theta_1, \theta_2, \ldots, \theta_p \in W(\mathbf{A})$$

- *Does there exist $\mathbf{V} \in C^{n\times p}$ with orthonormal columns such that*

$$\sigma(\mathbf{V}^*\mathbf{A}\mathbf{V}) = \{\theta_1, \theta_2, \ldots, \theta_p\} \quad ?$$

- *If so, how many generating subspaces* $\mathrm{Ran}(\mathbf{V})$ *exist ?*
- *How can these generating subspaces be constructed ?*

Inverse Field of Values Problem: multiple Ritz values

The generalization of this problem to larger subspaces is much more difficult.

Problem (iFOV(p))

Given $A \in C^{n\times n}$ and p points

$$\theta_1, \theta_2, \ldots, \theta_p \in W(\mathbf{A})$$

- *Does there exist $\mathbf{V} \in C^{n\times p}$ with orthonormal columns such that*

$$\sigma(\mathbf{V}^*\mathbf{A}\mathbf{V}) = \{\theta_1, \theta_2, \ldots, \theta_p\} \quad ?$$

- *If so, how many generating subspaces* $\mathrm{Ran}(\mathbf{V})$ *exist ?*
- *How can these generating subspaces be constructed ?*

- $p = 1$: Solvable for all $\theta_1 \in W(\mathbf{A})$.
- $p = n$: Solvable if and only if $\{\theta_1, \ldots, \theta_n\} = \sigma(\mathbf{A})$.
- Solvable for all p if $\mathbf{A}$ is Hermitian and Ritz values obey interlacing.
- Remaining cases are more interesting but analysis is much harder.

Inverse Field of Values Problem: Hermitian Case

Recall the interlacing result described earlier.

- Eigenvalues: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, Ritz values: $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_p$
  Interlacing gives:

$$\theta_1 \in [\lambda_1, \lambda_{n-p+1}], \quad \theta_2 \in [\lambda_2, \lambda_{n-p+2}], \quad \cdots, \quad \theta_p \in [\lambda_p, \lambda_n].$$

Interlacing limits the number of Ritz values between exterior eigenvalues.

(λ_1, λ_2)	(λ_2, λ_3)	(λ_3, λ_4)	$\cdots$	$(\lambda_{n-3}, \lambda_{n-2})$	$(\lambda_{n-2}, \lambda_{n-1})$	$(\lambda_{n-1}, \lambda_n)$
1	2	3	$\cdots$	3	2	1

Inverse Field of Values Problem: Hermitian Case

Recall the interlacing result described earlier.

- Eigenvalues: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, Ritz values: $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_p$
  Interlacing gives:

$$\theta_1 \in [\lambda_1, \lambda_{n-p+1}], \quad \theta_2 \in [\lambda_2, \lambda_{n-p+2}], \quad \cdots, \quad \theta_p \in [\lambda_p, \lambda_n].$$

  Interlacing limits the number of Ritz values between exterior eigenvalues.

(λ_1, λ_2)	(λ_2, λ_3)	(λ_3, λ_4)	$\cdots$	$(\lambda_{n-3}, \lambda_{n-2})$	$(\lambda_{n-2}, \lambda_{n-1})$	$(\lambda_{n-1}, \lambda_n)$
1	2	3	$\cdots$	3	2	1

- If $\mathbf{A}$ is Hermitian, iFOV(p) is solvable for any set of Ritz values $\{\theta_1, \ldots, \theta_p\}$ that obey interlacing. See, e.g., [Parlett 1980].

Inverse Field of Values Problem: Hermitian Case

Recall the interlacing result described earlier.

- Eigenvalues: $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, Ritz values: $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_p$
  Interlacing gives:

$$\theta_1 \in [\lambda_1, \lambda_{n-p+1}], \quad \theta_2 \in [\lambda_2, \lambda_{n-p+2}], \quad \cdots, \quad \theta_p \in [\lambda_p, \lambda_n].$$

  Interlacing limits the number of Ritz values between exterior eigenvalues.

(λ_1, λ_2)	(λ_2, λ_3)	(λ_3, λ_4)	$\cdots$	$(\lambda_{n-3}, \lambda_{n-2})$	$(\lambda_{n-2}, \lambda_{n-1})$	$(\lambda_{n-1}, \lambda_n)$
1	2	3	$\cdots$	3	2	1

- If $\mathbf{A}$ is Hermitian, iFOV(p) is solvable for any set of Ritz values $\{\theta_1, \ldots, \theta_p\}$ that obey interlacing. See, e.g., [Parlett 1980].

- There has long been interest in developing some notion of "interlacing" for non-Hermitian matrices. Results on iFOV(p) hint in this direction.
- For general normal matrices, the problem is already difficult: no all-purpose generalization of interlacing is available. (Specialized results: isometric Arnoldi process; normal $\mathbf{V}^*\mathbf{A}\mathbf{V}$ [Fan, Pall 1957; Queiró, Duarte 2008].)

Algebraic Characterization for iFOV($n-1$), Normal Case

Theorem (Thompson 1966; Malamud 2004)

Let $\mathbf{A}$ *be normal with distinct eigenvalues* $\lambda_1, \ldots, \lambda_n$*, and suppose we seek to solve iFOV(*$n-1$*) with Ritz values* $\theta_1, \ldots, \theta_{n-1}$*. Define*

$$x_k := \frac{\prod_{\ell=1}^{n-1}(\lambda_k - \theta_\ell)}{\prod_{\ell=1,\ell\neq k}^{n}(\lambda_k - \lambda_\ell)}.$$

There exists some $\mathbf{V} \in C^{n\times(n-1)}$*,* $\mathbf{V}^*\mathbf{V} = \mathbf{I}$ *such that*

$$\sigma(\mathbf{V}^*\mathbf{A}\mathbf{V}) = \{\theta_1, \ldots, \theta_{n-1}\}$$

if and only if $x_k \geq 0$ *for all* $k = 1, \ldots, n$*.*

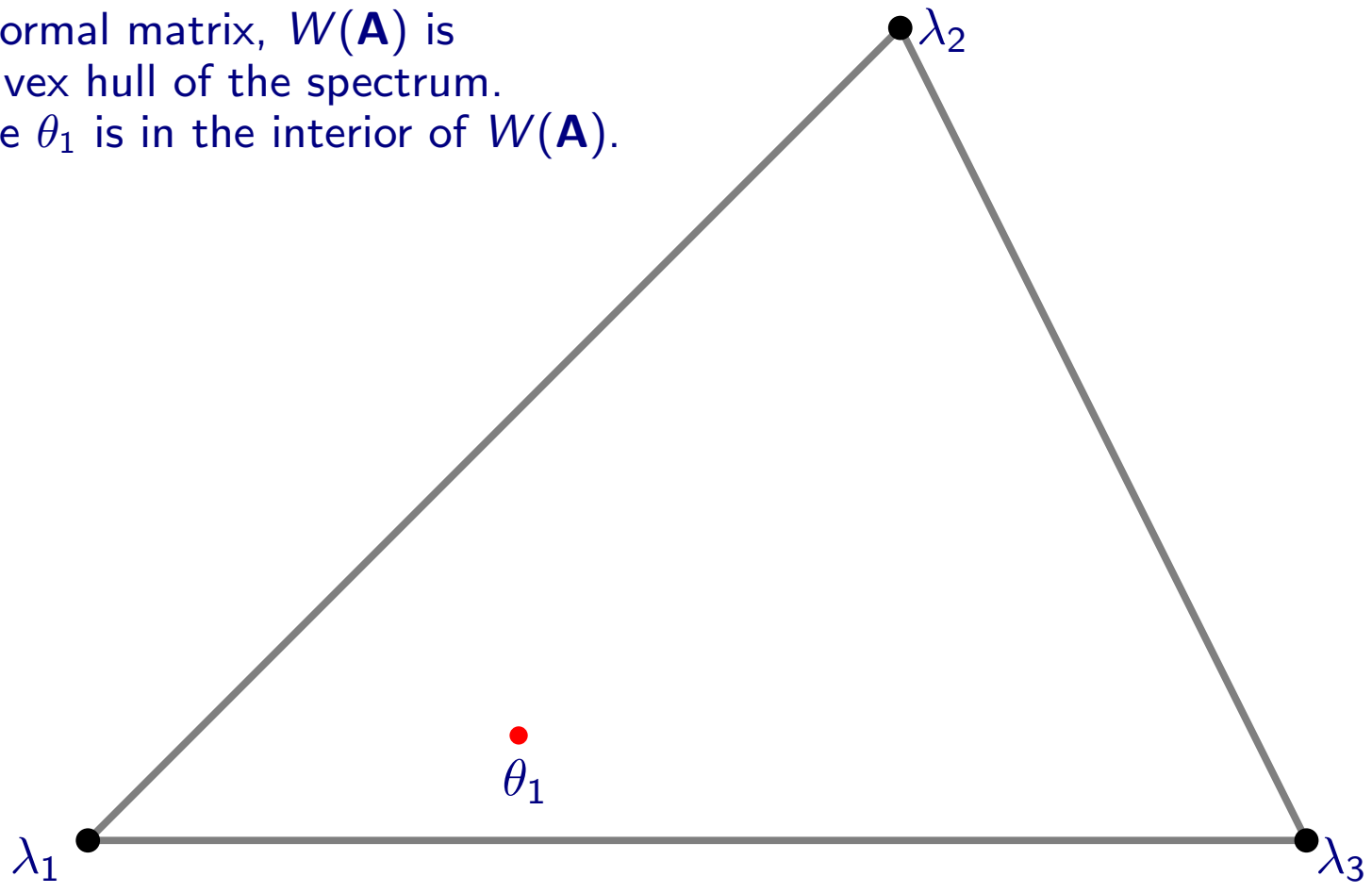
Malamud's motivation: relating zeros of polynomials and their derivatives.

Carden and Hansen use the requirement that $\arg(x_k) = 0 (\mathrm{mod}\ 2\pi)$ to give a geometric interpretation of these results when $n = 3$ using Ceva's Theorem.

iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

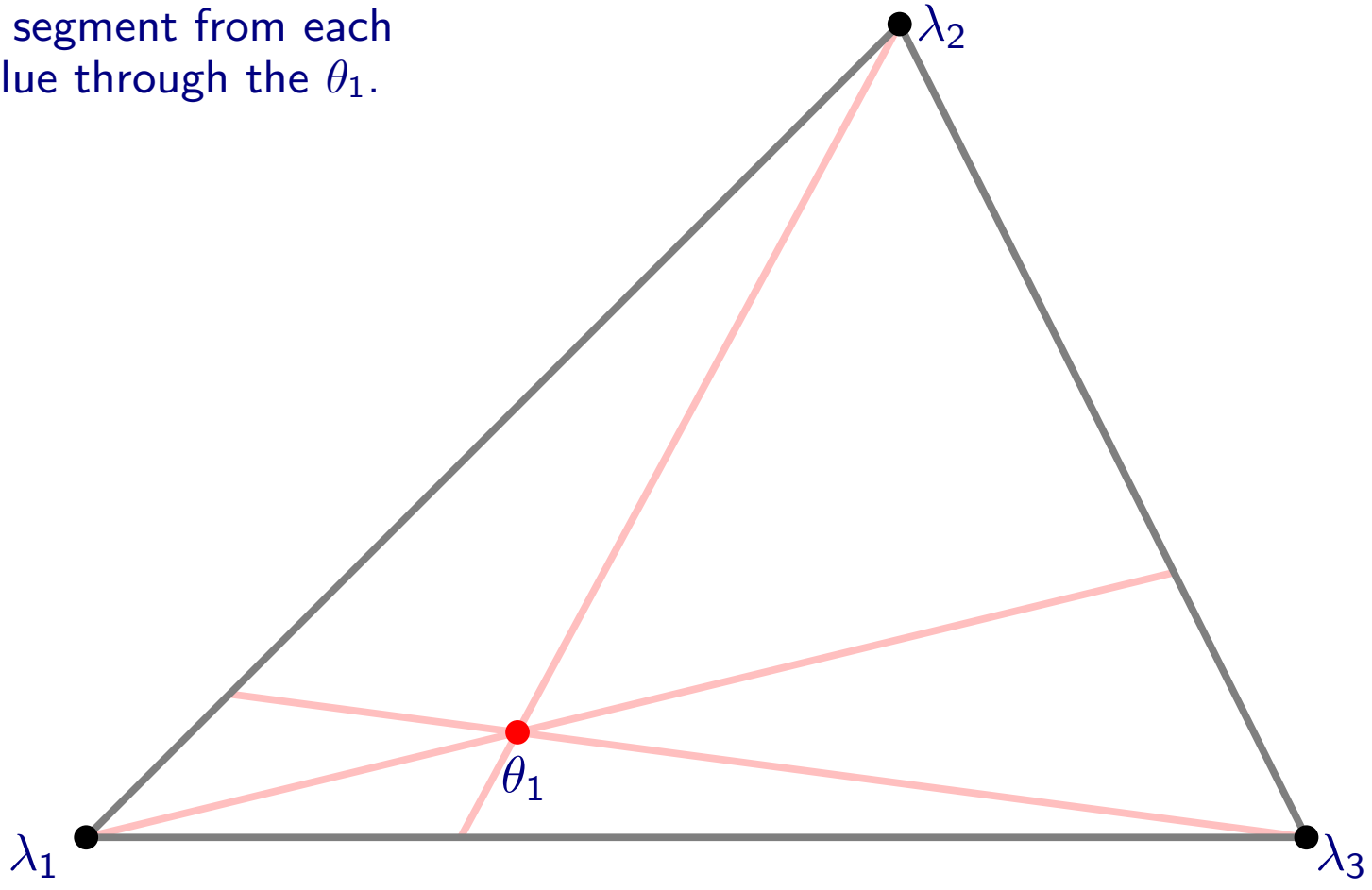
For a normal matrix, $W(\mathbf{A})$ is
the convex hull of the spectrum.
Suppose θ_1 is in the interior of $W(\mathbf{A})$.



iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

Draw a segment from each eigenvalue through the θ_1.

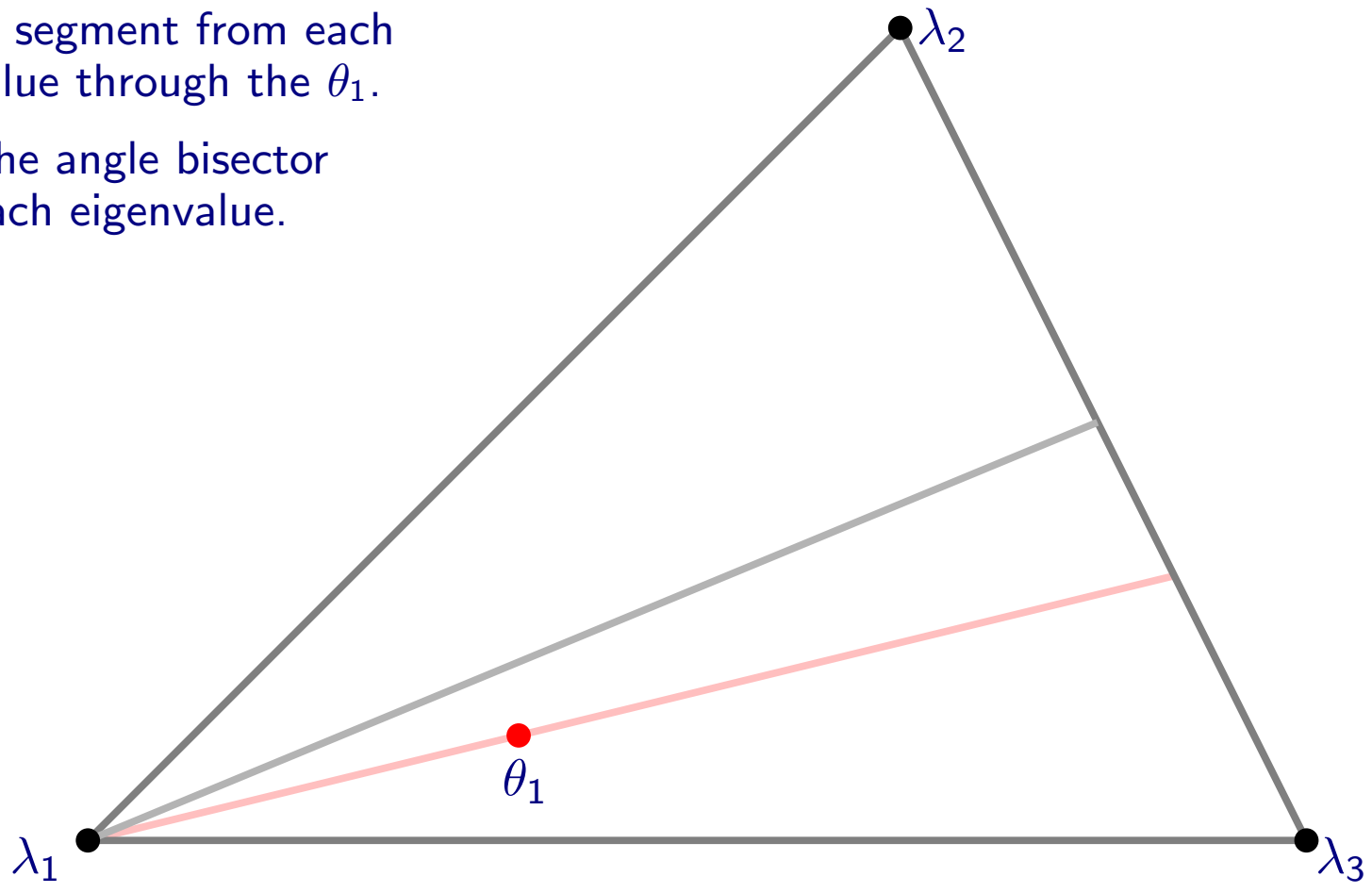


iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

Draw a segment from each eigenvalue through the θ_1.

Draw the angle bisector from each eigenvalue.



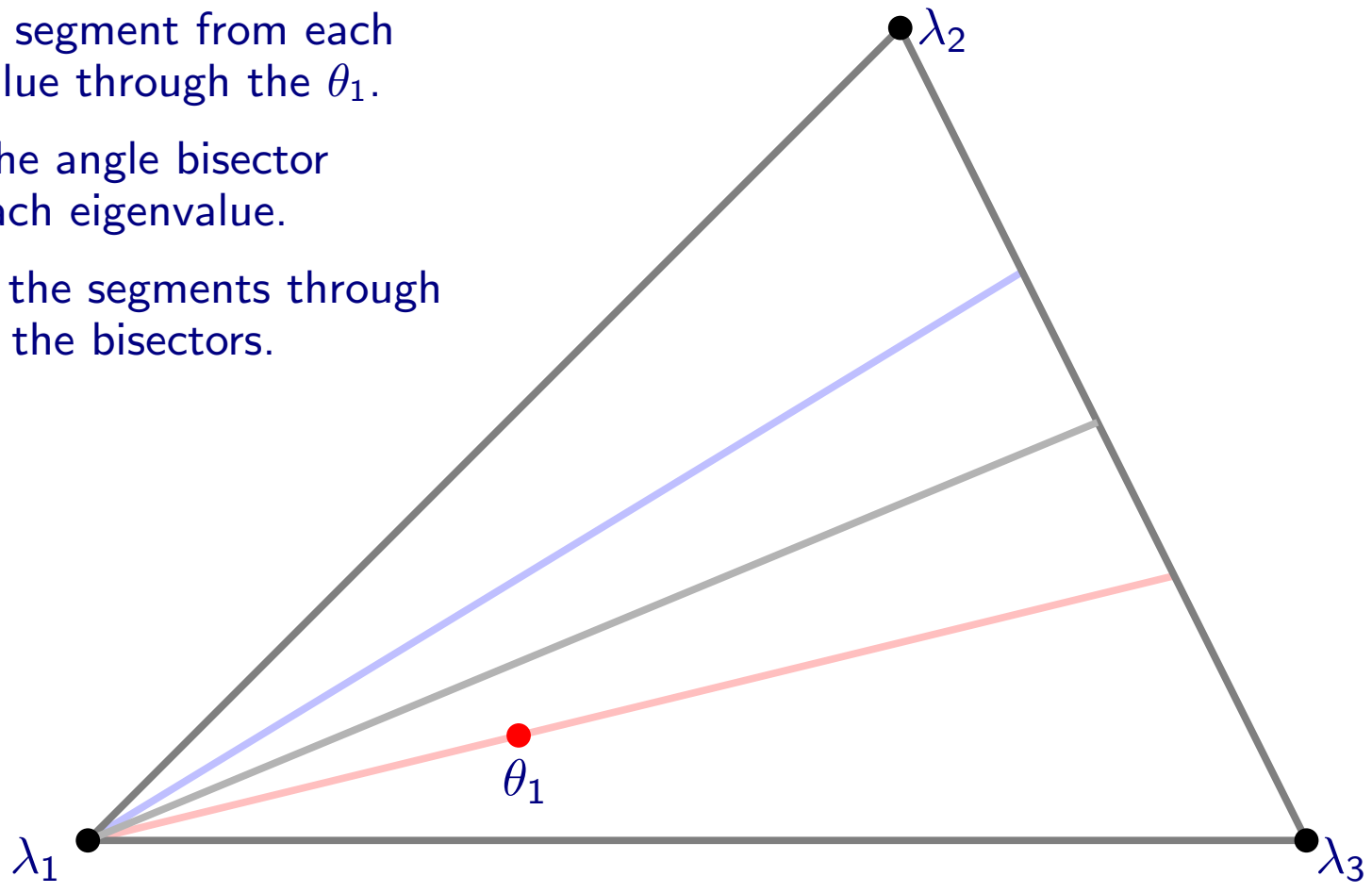
iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

Draw a segment from each eigenvalue through the θ_1.

Draw the angle bisector from each eigenvalue.

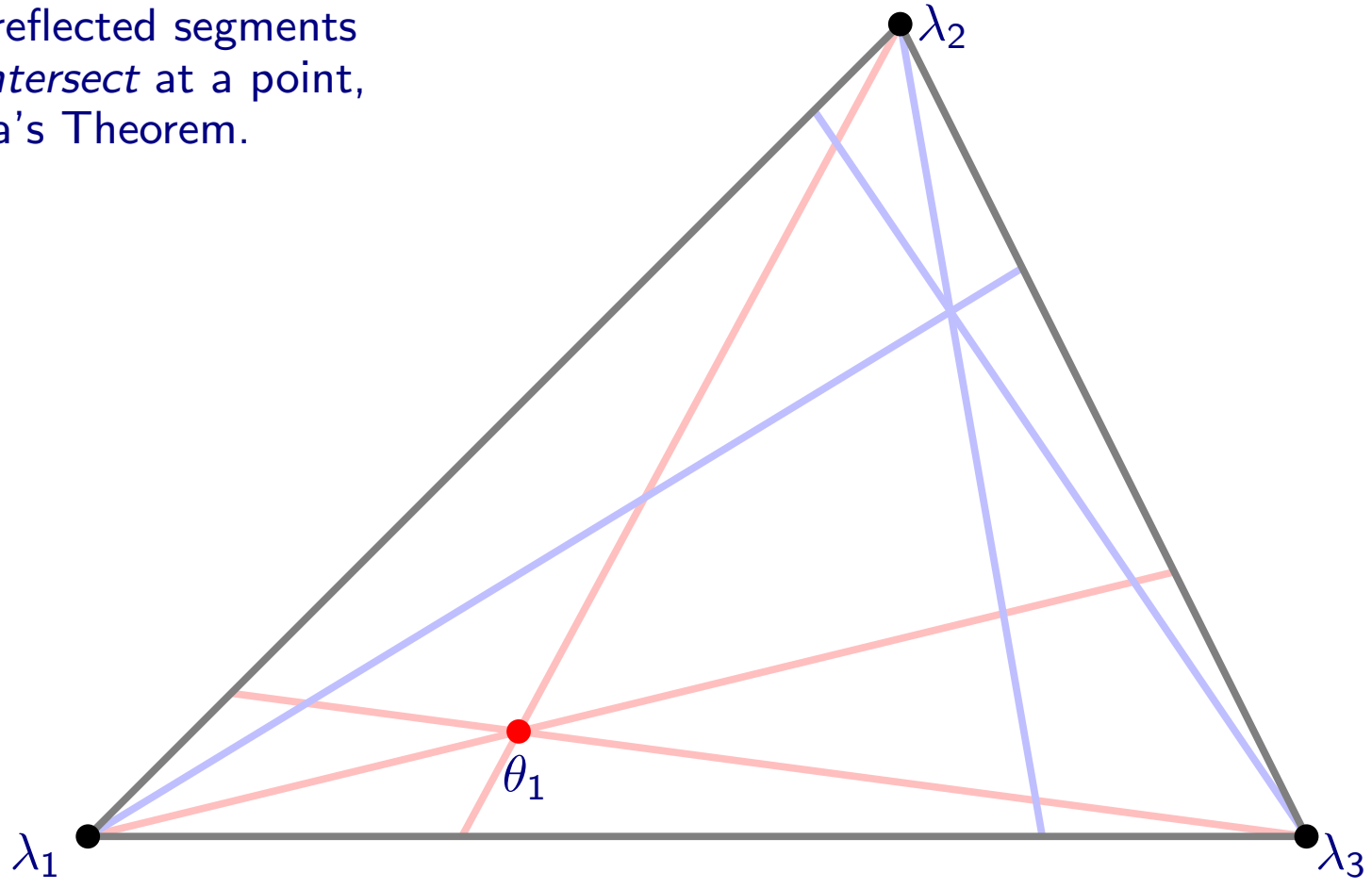
Reflect the segments through θ_1 over the bisectors.



iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

These reflected segments *must intersect* at a point, by Ceva's Theorem.

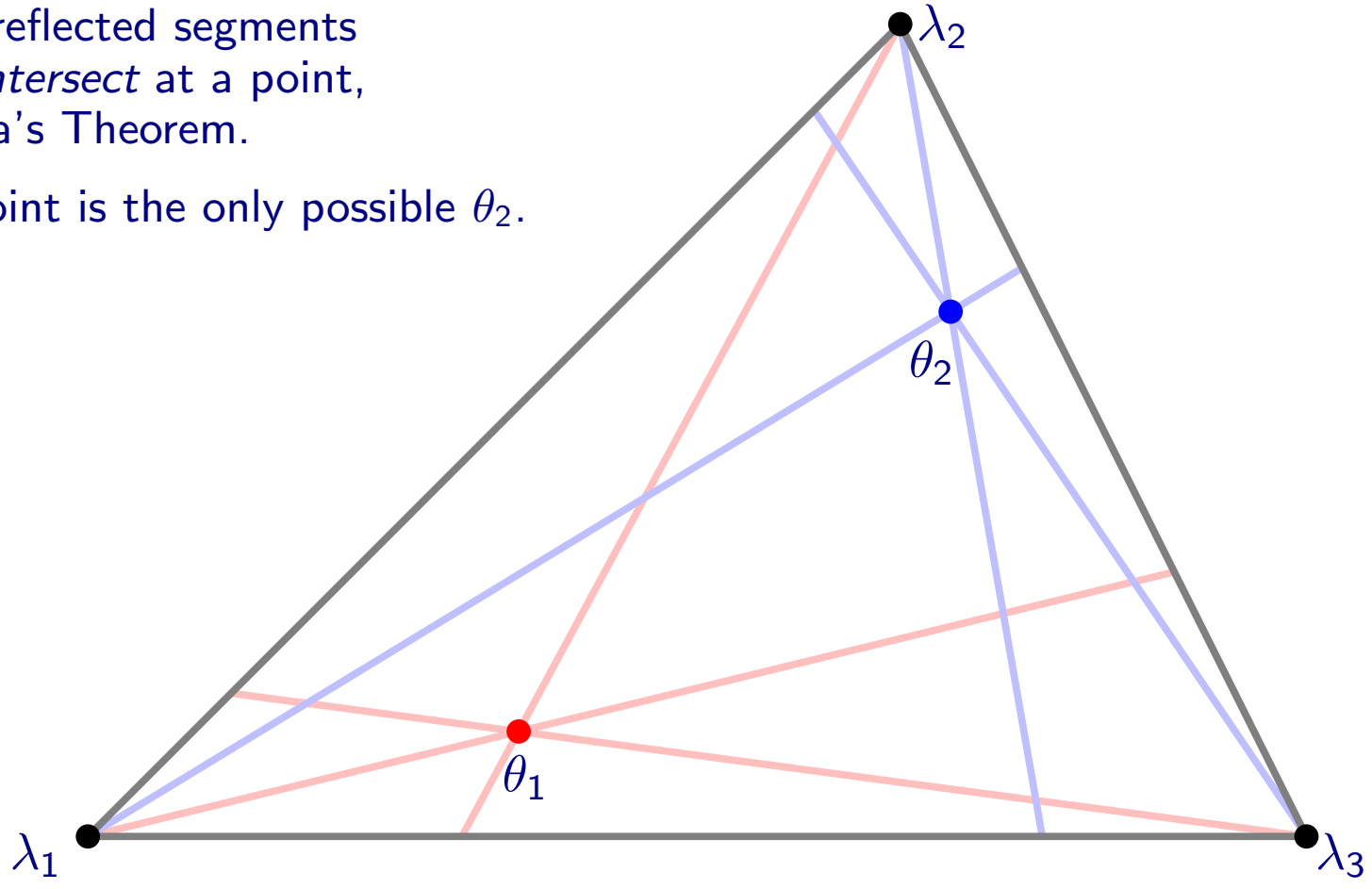


iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

These reflected segments *must intersect* at a point, by Ceva's Theorem.

This point is the only possible θ_2.



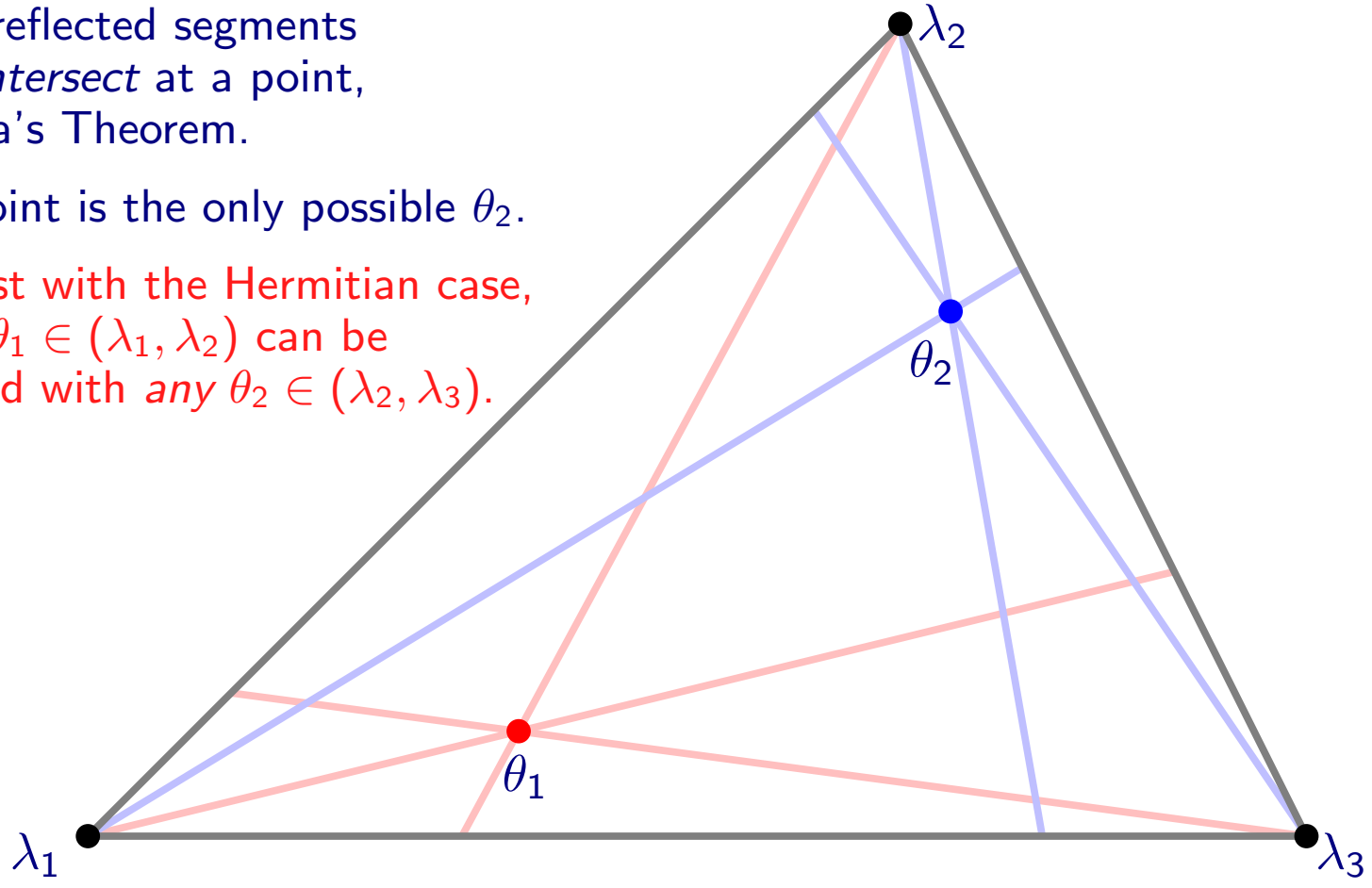
iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

These reflected segments *must intersect* at a point, by Ceva's Theorem.

This point is the only possible θ_2.

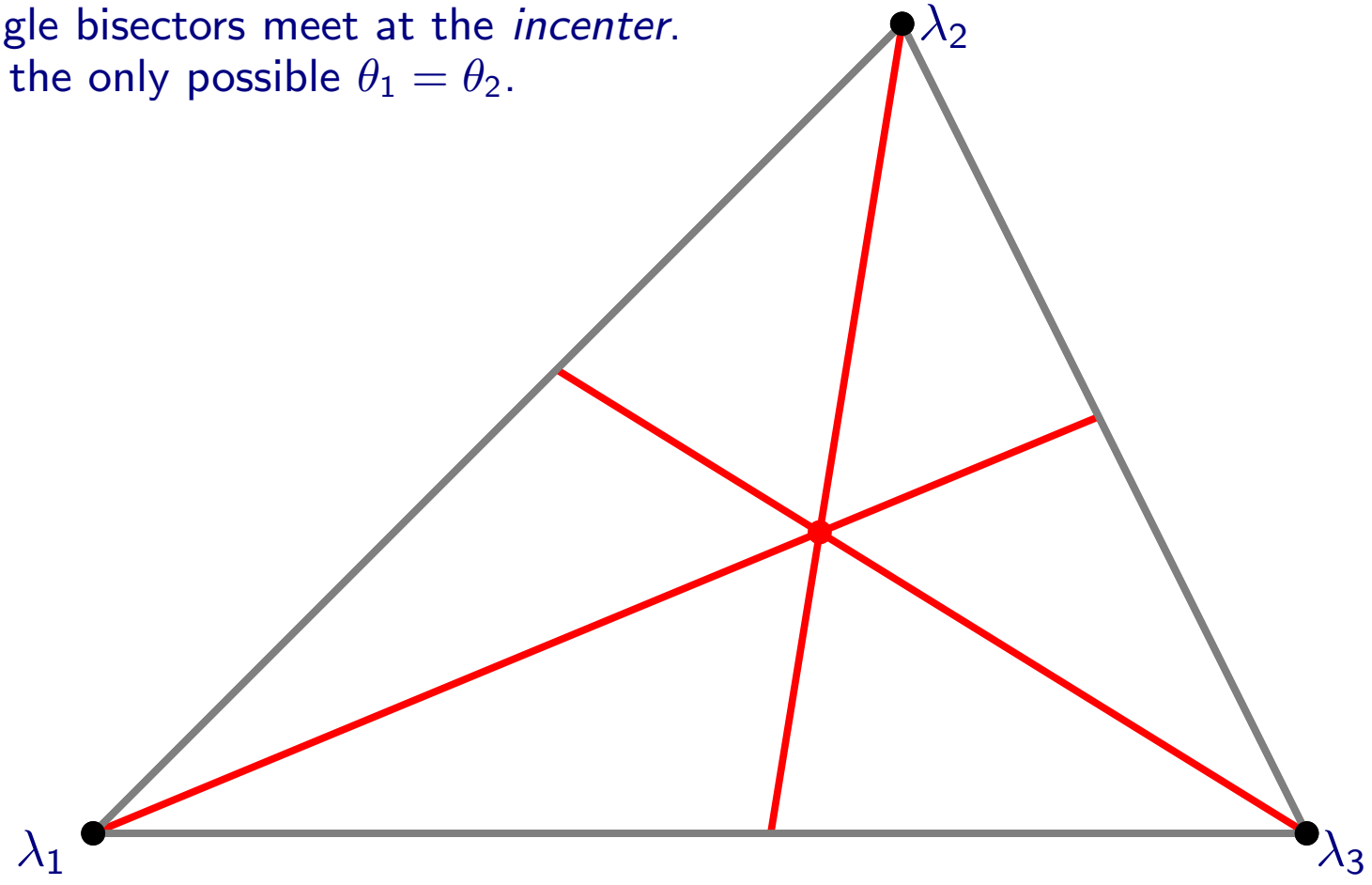
Contrast with the Hermitian case, where $\theta_1 \in (\lambda_1, \lambda_2)$ can be matched with *any* $\theta_2 \in (\lambda_2, \lambda_3)$.



iFOV($n-1$) for Normal Matrices, $n = 3$

Interpretation of this result for $n = 3$ [Carden, Hansen, 2011].

The angle bisectors meet at the *incenter*.
This is the only possible $\theta_1 = \theta_2$.



Ritz Value Localization for Non-Hermitian Matrices

Do we have any hope of some notion of "interlacing" for nonnormal matrices?

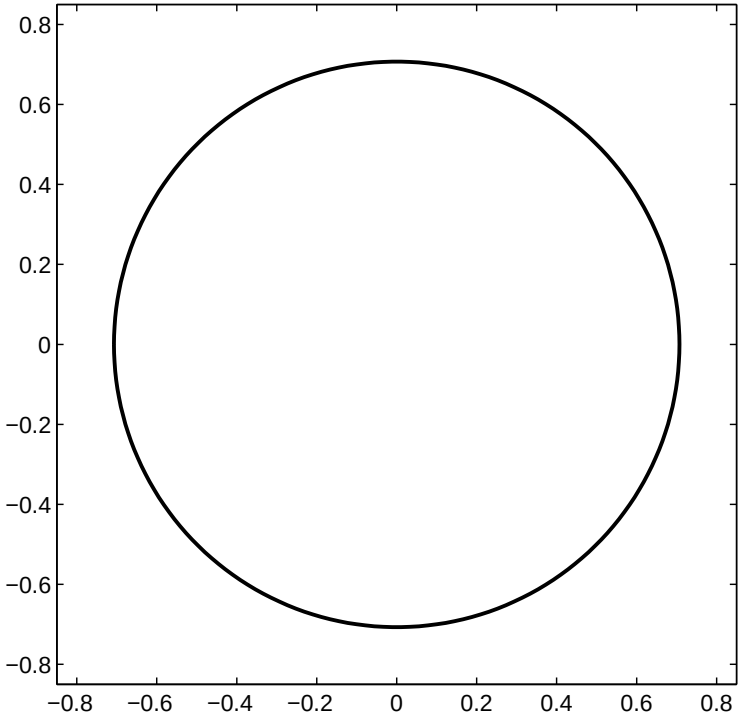
Consider an extreme example:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

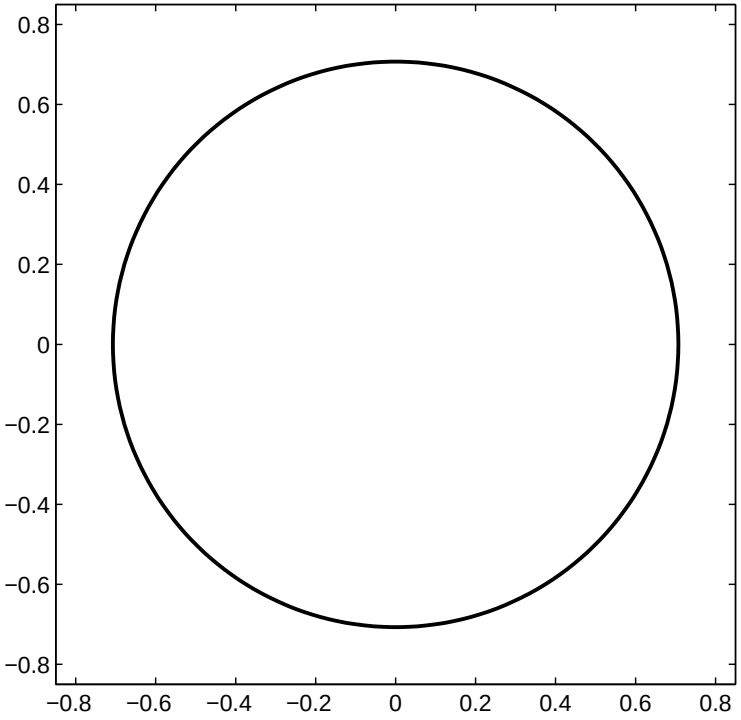
Repeat the following experiment many times:

- Generate random *two* dimensional subspaces, $\mathcal{V} = \operatorname{Ran} \mathbf{V}$, where $\mathbf{V}^* \mathbf{V} = \mathbf{I}$.
- Form $\mathbf{V}^* \mathbf{A} \mathbf{V} \in \mathbf{C}^{2 \times 2}$ and compute Ritz values $\{\theta_1, \theta_2\} = \sigma(\mathbf{V}^* \mathbf{A} \mathbf{V})$.
- Identify the leftmost and rightmost Ritz values.
- Since $\sigma(\mathbf{A}) = \{0\}$, "interlacing" is meaningless here...

Ritz Values of a Jordan Block



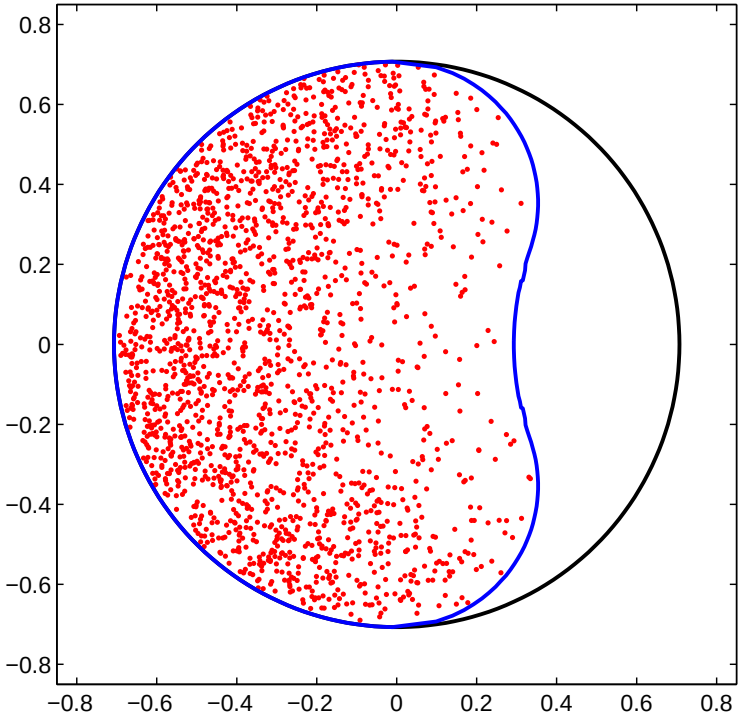
leftmost Ritz value



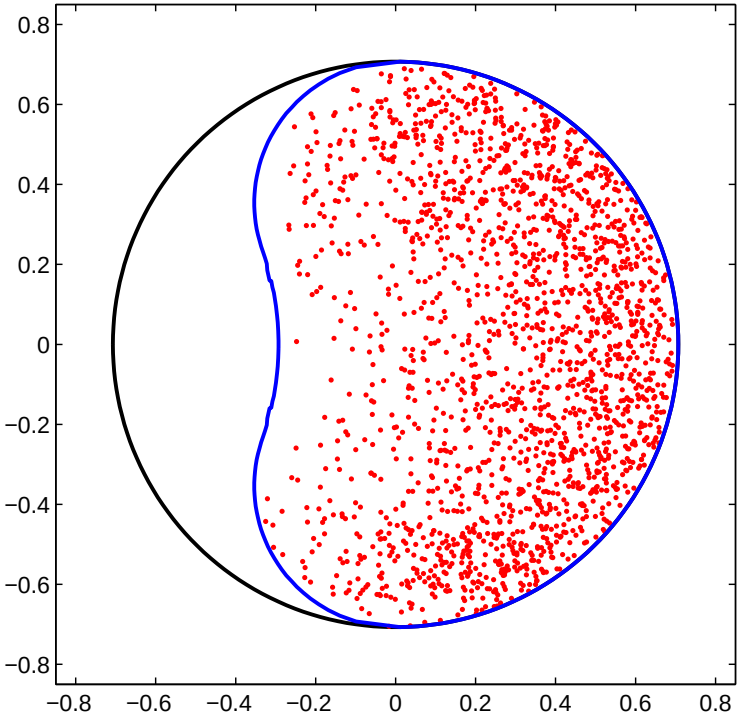
rightmost Ritz value

$$W(\mathbf{A}) = \{z \in \mathbf{C} : |z| \leq \sqrt{2}/2\}$$

Ritz Values of a Jordan Block



leftmost Ritz value



rightmost Ritz value

random (complex) two dimensional subspaces

Ritz Values for Three Matrices with Identical Fields of Values

Compute $p = 4$ Ritz values for these 8×8 matrices.

$$\mathbf{A}_1 = \gamma_1 \begin{bmatrix} 0 & 1 & & & & & & \\ & 0 & & & & & & \\ & & 0 & 1 & & & & \\ & & & 0 & & & & \\ & & & & 0 & 1 & & \\ & & & & & 0 & & \\ & & & & & & 0 & 1 \\ & & & & & & & 0 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} 0 & 1 & & & & & & \\ & 0 & 1 & & & & & \\ & & 0 & 1 & & & & \\ & & & 0 & 1 & & & \\ & & & & 0 & 1 & & \\ & & & & & 0 & 1 & \\ & & & & & & 0 & 1 \\ & & & & & & & 0 \end{bmatrix} \quad \mathbf{A}_3 = \gamma_3 \begin{bmatrix} 0 & \varrho^1 & & & & & & \\ & 0 & \varrho^2 & & & & & \\ & & 0 & \varrho^3 & & & & \\ & & & 0 & \varrho^4 & & & \\ & & & & 0 & \varrho^5 & & \\ & & & & & 0 & \varrho^6 & \\ & & & & & & 0 & \varrho^7 \\ & & & & & & & 0 \end{bmatrix}$$

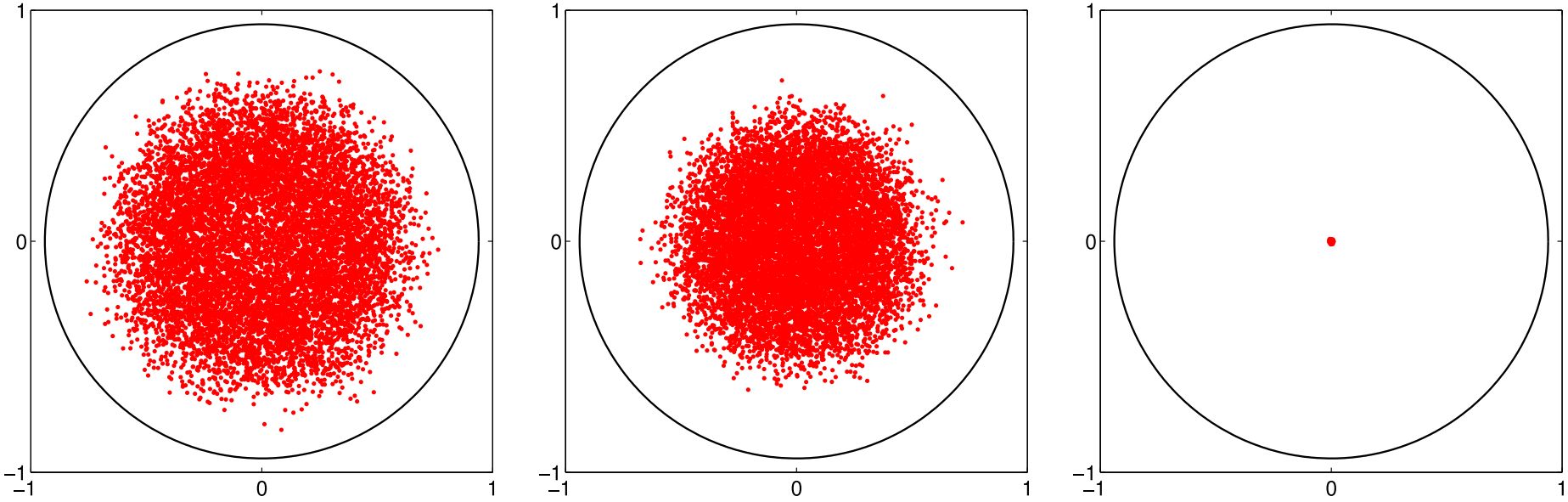
(γ_1 and γ_3 set to give same $W(\mathbf{A})$ for all examples; $\varrho = 1/8$.)

Ritz Values for Three Matrices with Identical Fields of Values

Compute $p = 4$ Ritz values for these 8×8 matrices.

$$\mathbf{A}_1 = \gamma_1 \begin{bmatrix} 0 & 1 & & & & & & \\ & 0 & & & & & & \\ & & 0 & 1 & & & & \\ & & & 0 & & & & \\ & & & & 0 & 1 & & \\ & & & & & 0 & & \\ & & & & & & 0 & 1 \\ & & & & & & & 0 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} 0 & 1 & & & & & & \\ & 0 & 1 & & & & & \\ & & 0 & 1 & & & & \\ & & & 0 & 1 & & & \\ & & & & 0 & 1 & & \\ & & & & & 0 & 1 & \\ & & & & & & 0 & 1 \\ & & & & & & & 0 \end{bmatrix} \quad \mathbf{A}_3 = \gamma_3 \begin{bmatrix} 0 & \varrho^1 & & & & & & \\ & 0 & \varrho^2 & & & & & \\ & & 0 & \varrho^3 & & & & \\ & & & 0 & \varrho^4 & & & \\ & & & & 0 & \varrho^5 & & \\ & & & & & 0 & \varrho^6 & \\ & & & & & & 0 & \varrho^7 \\ & & & & & & & 0 \end{bmatrix}$$

(γ_1 and γ_3 set to give same $W(\mathbf{A})$ for all examples; $\varrho = 1/8$.)



Smallest magnitude of $p = 4$ Ritz values, $10,000$ random complex subspaces.

Ritz Value Localization, Sorted by Real Part

Theorem (Carden, E. 2012)

Let $\theta_1, \ldots, \theta_p$ denote the Ritz values of $\mathbf{A} \in \mathbb{C}^{n \times n}$ drawn from a $p < n$ dimensional subspace, labeled by increasing real part: $\operatorname{Re}\theta_1 \leq \cdots \leq \operatorname{Re}\theta_p$. Then for $k = 1, \ldots, p$,

$$\frac{\mu_1 + \cdots + \mu_k}{k} \leq \operatorname{Re}\theta_k \leq \frac{\mu_{n-p+k} + \cdots + \mu_n}{p - k + 1},$$

where $\mu_1 \leq \cdots \leq \mu_n$ are the eigenvalues of $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$.

Ritz Value Localization, Sorted by Real Part

Theorem (Carden, E. 2012)

Let $\theta_1, \ldots, \theta_p$ denote the Ritz values of $\mathbf{A} \in \mathbb{C}^{n \times n}$ drawn from a $p < n$ dimensional subspace, labeled by increasing real part: $\operatorname{Re}\theta_1 \leq \cdots \leq \operatorname{Re}\theta_p$. Then for $k = 1, \ldots, p$,

$$\frac{\mu_1 + \cdots + \mu_k}{k} \leq \operatorname{Re}\theta_k \leq \frac{\mu_{n-p+k} + \cdots + \mu_n}{p-k+1},$$

where $\mu_1 \leq \cdots \leq \mu_n$ are the eigenvalues of $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$.

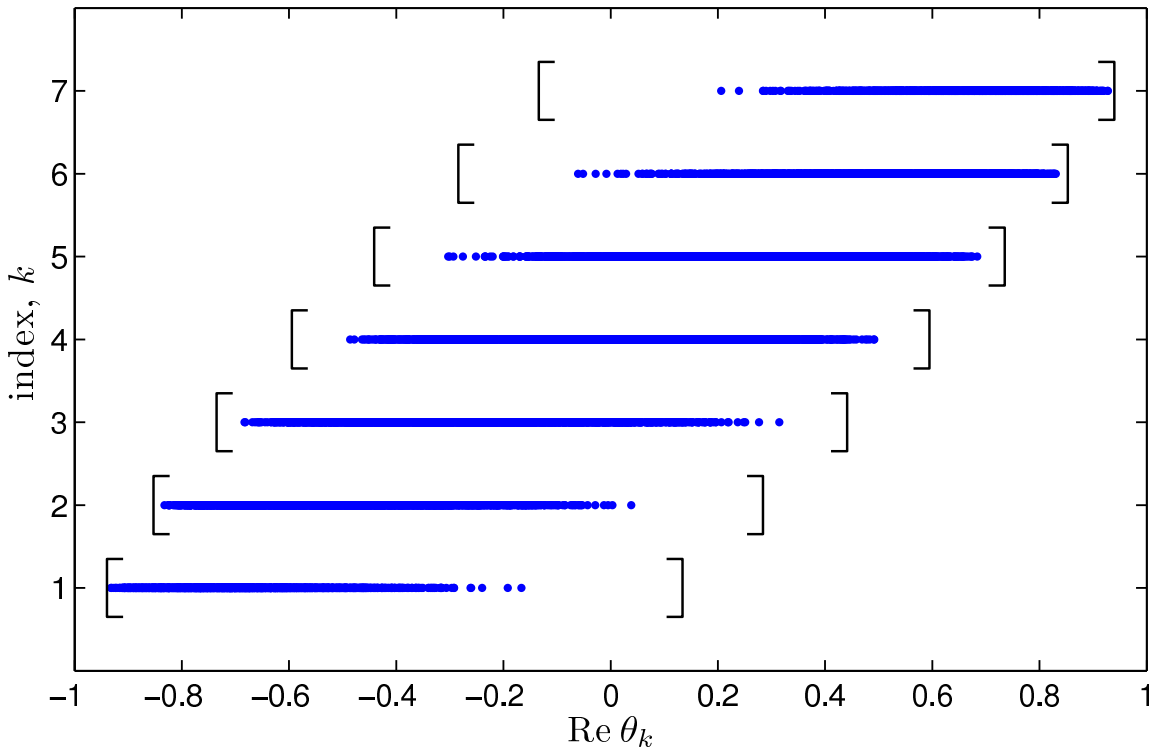
- Ky Fan similarly bounded the real parts of the eigenvalues of $\mathbf{A}$ [Fan 1950].
- The fact that $\theta_k \in W(\mathbf{A})$ gives the well-known bound

  $$\mu_1 \leq \operatorname{Re}\theta_k \leq \mu_n, \qquad k = 1, \ldots, p.$$

  The theorem provides sharper bounds for interior Ritz values.
- The interior eigenvalues of $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$ give additional insight; cf. eigenvalue inclusion regions of [Psarrakos and Tsatsomeros, 2012].

Illustration for Jordan Block

Let $\mathbf{A}$ be an 8×8 Jordan block, and take $p = 7$ Ritz values.

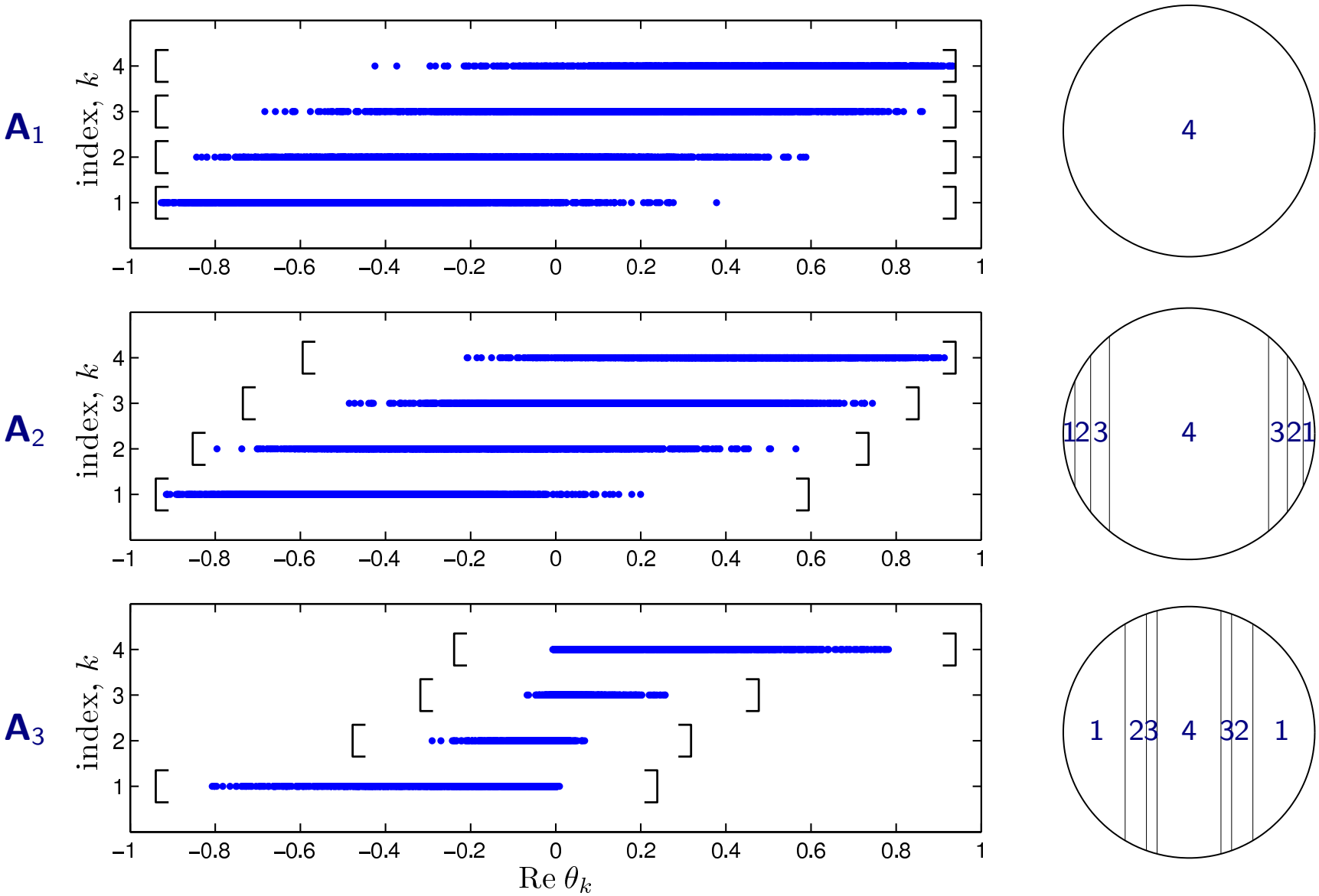


Containment intervals from previous theorem, along with Ritz values from 2000 random subspaces. (Note that 0 must be contained in all intervals, as the spectrum of the principal 7×7 submatrix is zero.)

The bounds are not sharp, but they do reveal more structure. . . .

Illustration for Three Matrices with Identical Fields of Values

Three matrices seen earlier with the same $W(\mathbf{A})$, different interior structure.
For $p = 4$, numbers on right indicate max Ritz values in each region.



Ritz Value Localization, Sorted by Magnitude

Similar results hold when Ritz values are sorted by magnitude.

Theorem (Carden, E. 2012)

Let $\theta_1, \ldots, \theta_p$ denote the Ritz values of $\mathbf{A} \in \boldsymbol{C}^{n \times n}$ drawn from a $p < n$ dimensional subspace, labeled by decreasing magnitude: $|\theta_1| \geq \cdots \geq |\theta_p|$. Then for $k = 1, \ldots, p$,

$$|\theta_k| \leq \left(s_1 \cdots s_k\right)^{1/k},$$

where $s_1 \geq \cdots \geq s_n$ are the singular values of $\mathbf{A}$.

Ritz Value Localization, Sorted by Magnitude

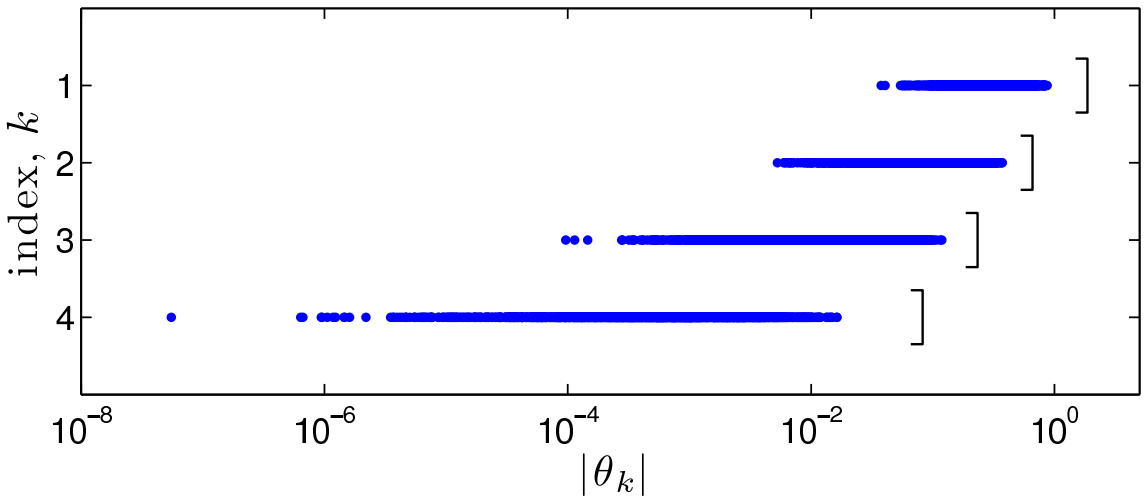
Similar results hold when Ritz values are sorted by magnitude.

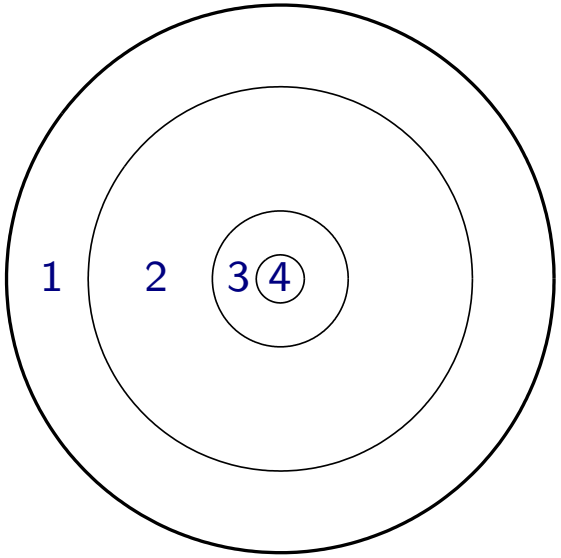
Theorem (Carden, E. 2012)

Let $\theta_1, \ldots, \theta_p$ denote the Ritz values of $\mathbf{A} \in \mathbb{C}^{n \times n}$ drawn from a $p < n$ dimensional subspace, labeled by decreasing magnitude: $|\theta_1| \geq \cdots \geq |\theta_p|$. Then for $k = 1, \ldots, p$,

$$|\theta_k| \leq (s_1 \cdots s_k)^{1/k},$$

where $s_1 \geq \cdots \geq s_n$ are the singular values of $\mathbf{A}$.





The numbers on the right denote the max Ritz values that can fall in each region.

Numerical range: practical example

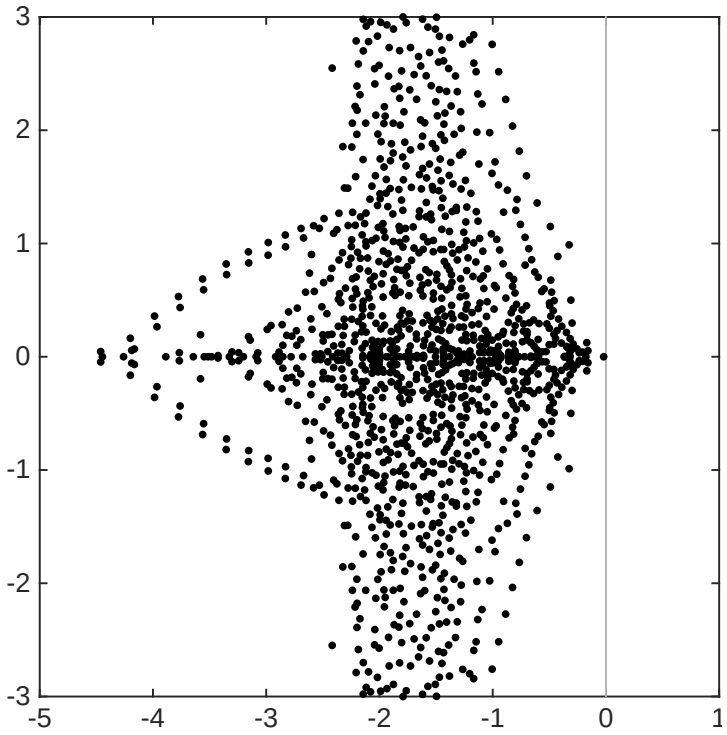
The numerical range can give good insight, but For many non-self-adjoint problems, it is too large. For example, the requirement that f be analytic on $W(\mathbf{A})$ can become quite restrictive, e.g., when f is a rational function with poles somewhat near eigenvalues.

- 2d flow over an backward-facing step, viscosity $\nu = 1/400$, discretized using Q_2–Q_1 finite elements via IFISS [Elman, Silvester, Ramage].

Numerical range: practical example

The numerical range can give good insight, but For many non-self-adjoint problems, it is too large. For example, the requirement that f be analytic on $W(\mathbf{A})$ can become quite restrictive, e.g., when f is a rational function with poles somewhat near eigenvalues.

- 2d flow over an backward-facing step, viscosity $\nu = 1/400$, discretized using Q_2–Q_1 finite elements via IFISS [Elman, Silvester, Ramage].
- The resulting matrix is *nondiagonalizable,*

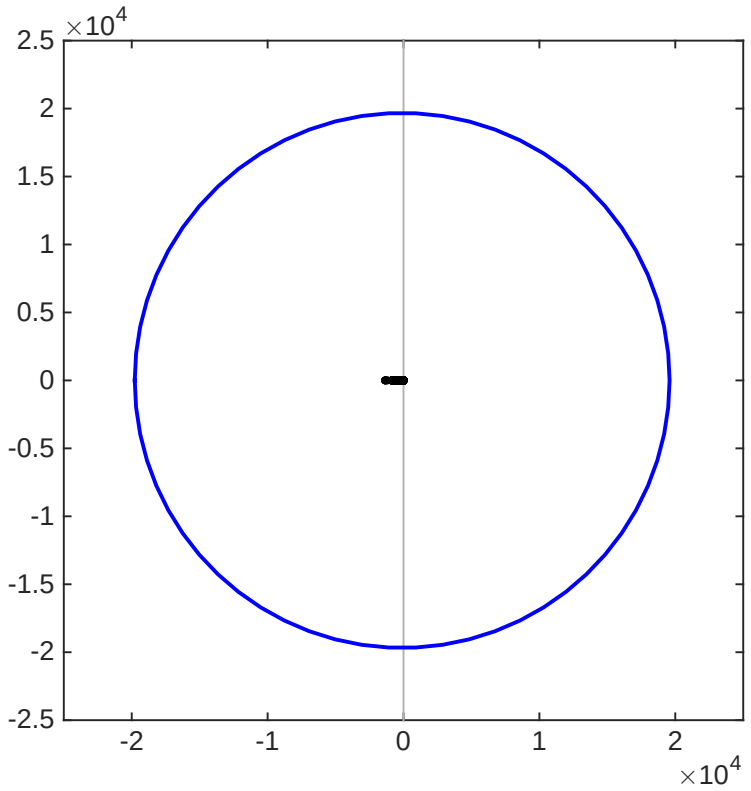


rightmost eigenvalues of $\mathbf{A}$

Numerical range: practical example

The numerical range can give good insight, but For many non-self-adjoint problems, it is too large. For example, the requirement that f be analytic on $W(\mathbf{A})$ can become quite restrictive, e.g., when f is a rational function with poles somewhat near eigenvalues.

- 2d flow over an backward-facing step, viscosity $\nu = 1/400$, discretized using Q_2–Q_1 finite elements via IFISS [Elman, Silvester, Ramage].
- The resulting matrix is *nondiagonalizable, and has a large numerical range.*



$W(\mathbf{A})$ and $\sigma(\mathbf{A})$

Pseudospectra and Matrix Behavior

Pseudospectra and matrix behavior

Question

If $\mathbf{A}, \mathbf{B}$ *have identical pseudospectra,* $\sigma_\varepsilon(\mathbf{A}) = \sigma_\varepsilon(\mathbf{B})$ *for all* $\varepsilon > 0$*, then does* $\|f(\mathbf{A})\| = \|f(\mathbf{B})\|$ *for all functions* f *analytic on* $\sigma(\mathbf{A}) = \sigma(\mathbf{B})$*?*

Do the pseudospectra of a matrix determine its behavior?

Anne Greenbaum*
Courant Institute of Mathematical Sciences
251 Mercer St.
New York, NY 10012
greenbau@cims.nyu.edu

Lloyd N. Trefethen†
Department of Computer Science
Cornell University
Ithaca, NY 14853
LNT@cs.cornell.edu

Unpublished Cornell tech report, 1993 (see *Spectra and Pseudospectra*, Ch. 47.)

Pseudospectra and matrix behavior

Greenbaum and Trefethen show that the matrices

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & & \\ & & & 0 & 0 \\ & & & 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 1 & 0 & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & & \\ & & & 0 & \alpha \\ & & & 0 & 0 \end{bmatrix}$$

for $1 < |\alpha| \leq \sqrt{2}$, have the same pseudospectra, $\sigma_\varepsilon(\mathbf{A}) = \sigma_\varepsilon(\mathbf{B})$ for all $\varepsilon > 0$,

$$\|\mathbf{A}\| = 1 < |\alpha| = \|\mathbf{B}\|,$$

so *pseudospectra do not determine behavior* (in the 2-norm).

A and **B** are *derogatory* (multiple Jordan blocks associated with $\lambda = 0$).

Will all such examples be derogatory?

Extreme results about behavior and pseudospectra

Borque and Ransford (2009) give a beautifully alarming example.

For distinct $\alpha, \beta \in (0, \pi/4]$, the *non-derogatory* matrices

$$\mathbf{A} = \begin{bmatrix} 0 & \sec\alpha & 0 & 1 \\ 0 & 0 & \sec\beta\csc\beta & 0 \\ 0 & 0 & 0 & \csc\alpha \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0 & \sec\beta & 0 & 1 \\ 0 & 0 & \sec\alpha\csc\alpha & 0 \\ 0 & 0 & 0 & \csc\beta \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

have *super-identical pseudospectra*, i.e., for all $z \in \mathbf{C}$, *all* singular values of $z - \mathbf{A}$ and $z - \mathbf{B}$ match:

$$s_j(z - \mathbf{A}) = s_j(z - \mathbf{B}), \qquad j = 1, \ldots, n,$$

yet

$$\frac{\|\mathbf{A}^2\|}{\|\mathbf{B}^2\|} = \frac{\cos\alpha}{\cos\beta}.$$

Will all such examples be non-diagonalizable?

Extreme results about behavior and pseudospectra

Borque and Ransford (2009) give a beautifully alarming example.

For distinct $\alpha, \beta \in (0, \pi/4]$, the *non-derogatory* matrices

$$\mathbf{A} = \begin{bmatrix} 0 & \sec\alpha & 0 & 1 \\ 0 & 0 & \sec\beta\csc\beta & 0 \\ 0 & 0 & 0 & \csc\alpha \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0 & \sec\beta & 0 & 1 \\ 0 & 0 & \sec\alpha\csc\alpha & 0 \\ 0 & 0 & 0 & \csc\beta \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

have *super-identical pseudospectra*, i.e., for all $z \in \mathbf{C}$, *all* singular values of $z - \mathbf{A}$ and $z - \mathbf{B}$ match:

$$s_j(z - \mathbf{A}) = s_j(z - \mathbf{B}), \qquad j = 1, \ldots, n,$$

yet

$$\frac{\|\mathbf{A}^2\|}{\|\mathbf{B}^2\|} = \frac{\cos\alpha}{\cos\beta}.$$

Will all such examples be non-diagonalizable?

No! [Ransford, Rostand 2011] give a 4×4 example with *simple eigenvalues* and super-identical pseudospectra with $\|\mathbf{A}^2\| \neq \|\mathbf{B}^2\|$.

Extreme results about behavior and pseudospectra

How extreme can the difference in behavior be?

Theorem (Ransford, 2007)

Given n, specify two positive-valued submultiplicative sequences $\alpha_2, \ldots, \alpha_n$ and $\beta_2, \ldots, \beta_n$ (i.e., $\alpha_{j+k} \leq \alpha_j \alpha_k$ and $\beta_{j+k} \leq \beta_j \beta_k$ for all j, k). Then one can construct $\mathbf{A}, \mathbf{B} \in C^{N \times N}$ with $N = 2n + 3$ such that

$$\sigma_\varepsilon(\mathbf{A}) = \sigma_\varepsilon(\mathbf{B}) \quad \text{for all } \varepsilon > 0,$$

but

$$\|\mathbf{A}^k\| = \alpha_k, \qquad \|\mathbf{B}^k\| = \beta_k, \quad \text{for } k = 2, \ldots, n.$$

Pseudospectra and behavior in the Hilbert–Schmidt norm

There is one positive result lurking that has not been much digested, in light of these extreme examples.

Consider the Hilbert–Schmidt (Frobenius) norm

$$\|\mathbf{Z}\|_{\mathrm{HS}} = \mathrm{trace}(\mathbf{Z}^*\mathbf{Z}) = \left(\sum_{j,k=1}^{n} |z_{j,k}|^2\right)^{1/2} = \sqrt{s_1(\mathbf{Z})^2 + \cdots + s_n(\mathbf{Z})^2}.$$

Define the ε-pseudospectrum of $\mathbf{A}$ in this norm to be

$$\sigma_\varepsilon^{\mathrm{HS}}(\mathbf{A}) = \{z \in \mathbf{C} : \|(z - \mathbf{A})^{-1}\|_{\mathrm{HS}} > 1/\varepsilon\}.$$

Theorem (Greenbaum, Trefethen, 1993)

If $\sigma_\varepsilon^{\mathrm{HS}}(\mathbf{A}) = \sigma_\varepsilon^{\mathrm{HS}}(\mathbf{B})$ for all $\varepsilon > 0$, then $\|f(\mathbf{A})\|_{\mathrm{HS}} = \|f(\mathbf{B})\|_{\mathrm{HS}}$ for all f analytic on $\sigma(\mathbf{A}) = \sigma(\mathbf{B})$.

N.B. If pseudospectra are defined in terms of the eigenvalue perturbation definition, then $\sigma_\varepsilon^{\mathrm{HS}}(\mathbf{A})$ is the same as the usual 2-norm pseudospectrum.

Pseudospectra for Matrix Pencils
Transient Analysis of DAEs

Linear stability analysis for fluid flows

Pseudospectra/nornormality have provided a compelling tool for analyzing subcritical transition to turbulence in fluid flows, particularly for classical problems where the dynamics can be reduced to simple ODEs, e.g., Orr–Sommerfeld; e.g., [Butler, Farrell 1992], [Trefethen, Trefethen, Reddy, Driscoll 1993], [Reddy, Schmid, Henningson 1993], [Schmid, Henningson 2001].

More generally, for a given flow regime one needs to:

- Find a steady-state flow (Picard/Newton iterations).
- Linearize the flow about this steady-state to obtain

$$\begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}'(t) \\ \mathbf{p}'(t) \end{bmatrix} = \begin{bmatrix} \mathbf{F} & \mathbf{C}^* \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}(t) \\ \mathbf{p}(t) \end{bmatrix},$$

  which we write as $\mathbf{B}\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$.
- Analyze the spectral properties of the pencil $(\mathbf{A}, \mathbf{B})$.
- Need a generalization of pseudospectra for matrix pencils.
- For 2d examples we use the IFISS package [Elman, Silvester, Ramage].

See, e.g., [Gunzberger 1989].

Pseudospectra of matrix pencils

- Many definitions of pseudospectra of matrix pencils have been proposed: [Riedel 1994], [Ruhe 1995], [Frayssé, Gueury, Nicoud, Toumazou 1996], etc.,
- Further generalizations (polynomial, delay, nonlinear EVPs): [Tisseur, Higham 2001], [Green, Wagenknecht 2006], [Bindel, Hood 2013].
- For example, one often sees

$$\sigma_\varepsilon(\mathbf{A}, \mathbf{B}) = \{z \in \mathbf{C} : \|(z\mathbf{B} - \mathbf{A})^{-1}\| > 1/\varepsilon\},$$

but even though $\mathbf{SBx}'(t) = \mathbf{SAx}(t)$ has the same dynamics as $\mathbf{Bx}'(t) = \mathbf{Ax}(t)$, *in general* $\sigma_\varepsilon(\mathbf{SA}, \mathbf{SB}) \neq \sigma_\varepsilon(\mathbf{A}, \mathbf{B})$ *using this definition*.

Pseudospectra of matrix pencils

- Many definitions of pseudospectra of matrix pencils have been proposed: [Riedel 1994], [Ruhe 1995], [Frayssé, Gueury, Nicoud, Toumazou 1996], etc.,
- Further generalizations (polynomial, delay, nonlinear EVPs): [Tisseur, Higham 2001], [Green, Wagenknecht 2006], [Bindel, Hood 2013].
- For example, one often sees

$$\sigma_\varepsilon(\mathbf{A}, \mathbf{B}) = \{z \in \mathbf{C} : \|(z\mathbf{B} - \mathbf{A})^{-1}\| > 1/\varepsilon\},$$

  but even though $\mathbf{SBx}'(t) = \mathbf{SAx}(t)$ has the same dynamics as $\mathbf{Bx}'(t) = \mathbf{Ax}(t)$, *in general* $\sigma_\varepsilon(\mathbf{SA}, \mathbf{SB}) \neq \sigma_\varepsilon(\mathbf{A}, \mathbf{B})$ *using this definition*.

- Key: We use pseudospectra to analyze *dynamics*, rather than perturbations in eigenvalue computations.
- If $\mathbf{B}$ is invertible, the 'right' approach (cf. [Ruhe 1995]) considers

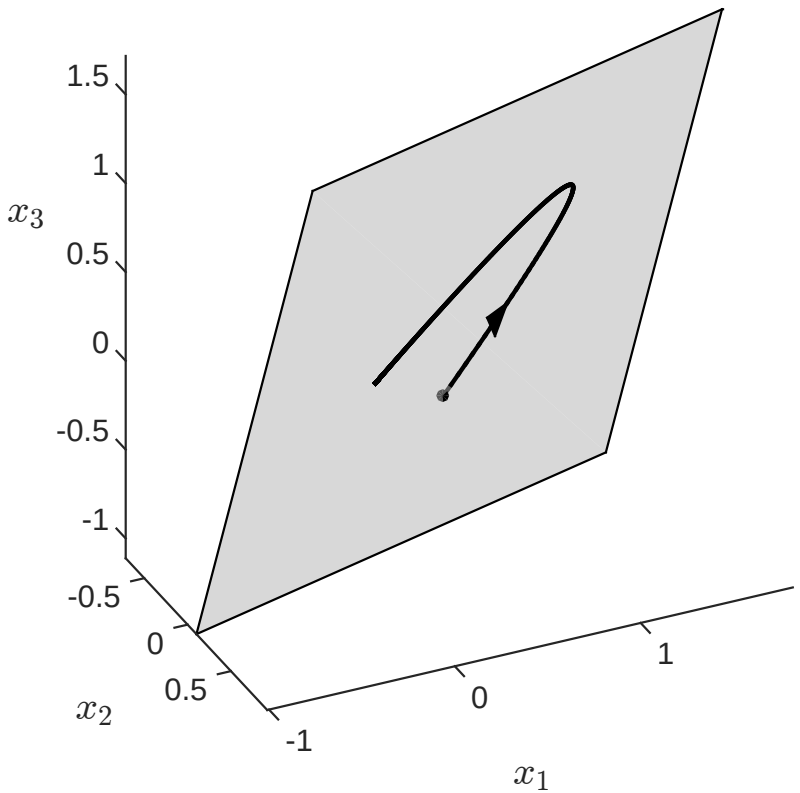
$$\mathbf{x}'(t) = \mathbf{B}^{-1}\mathbf{Ax}(t)$$

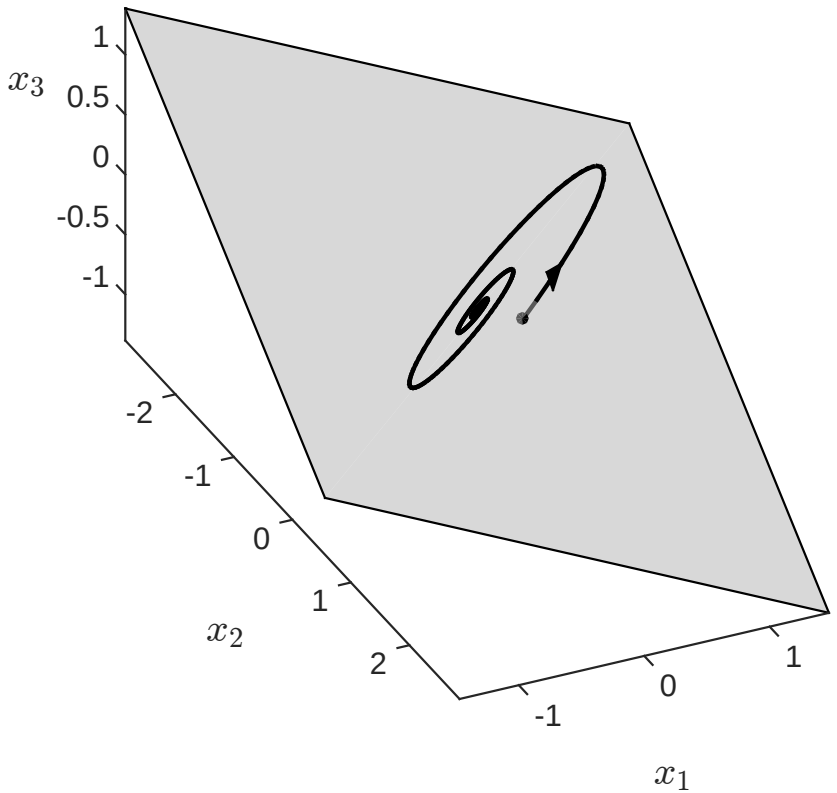
  and analyzes $\sigma_\varepsilon(\mathbf{B}^{-1}\mathbf{A})$ *in the correct physical norm*.

Transient dynamics of differential algebraic equations

When $\mathbf{B}$ is singular and $\mathbf{A} - \lambda\mathbf{B}$ is invertible for some λ, we have a *differential algebraic equation (DAE)*. The examples below show stable DAE solutions exhibiting transient growth [E., Keeler].

$$\mathbf{A} = \begin{bmatrix} -1 & 10 & 0 \\ -1 & -1 & 0 \\ 1 & -1 & -1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} -1 & 25 & 0 \\ -1 & -1 & 0 \\ 1 & -1 & -1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$





Pseudospectra of matrix pencils

- When $\mathbf{B}$ is singular, as it is when

$$\mathbf{B} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

  we must use tools from DAEs to understand transient dynamics [Cambpell, Meyer 1979], [Kunkel, Mehrmann 2006].

Pseudospectra of matrix pencils

- When $\mathbf{B}$ is singular, as it is when

$$\mathbf{B} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

  we must use tools from DAEs to understand transient dynamics [Cambpell, Meyer 1979], [Kunkel, Mehrmann 2006].
- Simplest case: for invertible $\mathbf{A}$ we can write the Schur form

$$\mathbf{A}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{G} & \mathbf{S} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^* \\ \mathbf{U}_2^* \end{bmatrix}$$

  for $[\mathbf{U}_1\ \mathbf{U}_2]$ unitary, $\mathbf{G}$ invertible, and $\mathbf{N}$ nilpotent.

Pseudospectra of matrix pencils

- When $\mathbf{B}$ is singular, as it is when

$$\mathbf{B} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

  we must use tools from DAEs to understand transient dynamics [Cambpell, Meyer 1979], [Kunkel, Mehrmann 2006].
- Simplest case: for invertible $\mathbf{A}$ we can write the Schur form

$$\mathbf{A}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{G} & \mathbf{S} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^* \\ \mathbf{U}_2^* \end{bmatrix}$$

  for $[\mathbf{U}_1\ \mathbf{U}_2]$ unitary, $\mathbf{G}$ invertible, and $\mathbf{N}$ nilpotent.
- Then the dynamics evolve as

$$\mathbf{x}(t) = \mathbf{U}_1 \mathrm{e}^{t\mathbf{G}^{-1}} \mathbf{U}_1^* \mathbf{x}(0)$$

  for initial conditions that satisfy the algebraic constraints, $\mathbf{x}(0) \in \mathrm{Ran}(\mathbf{U}_1)$.

Pseudospectra of matrix pencils

- When $\mathbf{B}$ is singular, as it is when

$$\mathbf{B} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

  we must use tools from DAEs to understand transient dynamics [Cambpell, Meyer 1979], [Kunkel, Mehrmann 2006].
- Simplest case: for invertible $\mathbf{A}$ we can write the Schur form

$$\mathbf{A}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{G} & \mathbf{S} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^* \\ \mathbf{U}_2^* \end{bmatrix}$$

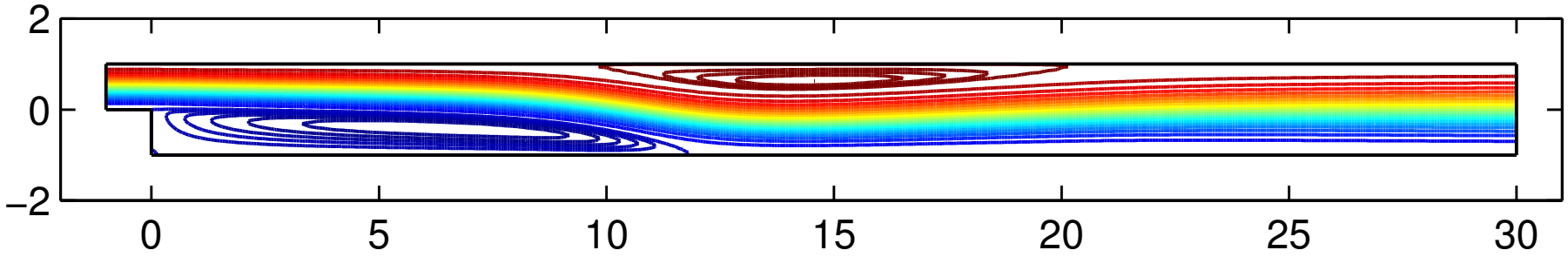
  for $[\mathbf{U}_1\ \mathbf{U}_2]$ unitary, $\mathbf{G}$ invertible, and $\mathbf{N}$ nilpotent.
- Then the dynamics evolve as

$$\mathbf{x}(t) = \mathbf{U}_1 \mathrm{e}^{t\mathbf{G}^{-1}} \mathbf{U}_1^* \mathbf{x}(0)$$

  for initial conditions that satisfy the algebraic constraints, $\mathbf{x}(0) \in \mathrm{Ran}(\mathbf{U}_1)$.

- *To understand the transient dynamics, study $\sigma_\varepsilon(\mathbf{G}^{-1})$ in the right norm.*

Pseudospectra for flow over a backward facing step



(viscosity $\nu = 1/600$)

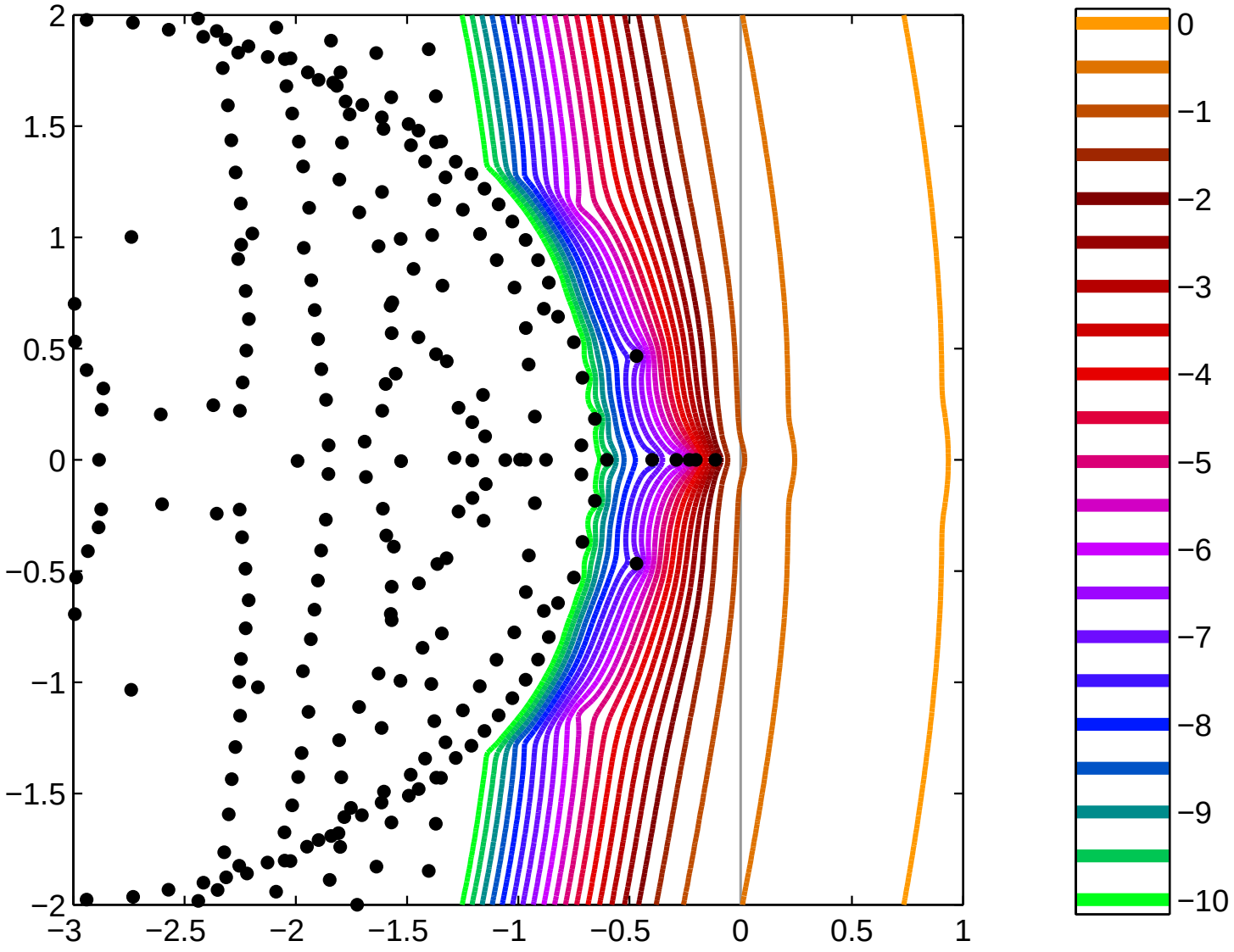
This is a notorious fluid stability problem; see [Gresho et al. 1993].

To compute pseudospectra $\sigma_\varepsilon(\mathbf{G}^{-1})$:

- Transform coordinates so the vector 2-norm approximates the energy norm for the PDE.
- Use the implicitly restarted Arnoldi algorithm (ARPACK/eigs) to compute the portion of $\mathbf{G}^{-1}$ active on the invariant subspace associated with the 1000 smallest magnitude eigenvalues.
- Numerous helpful tools are available: [Cliffe, Garratt, Spence 1994], [Stykel 2008], [Heinkenschloss, Sorensen, Sun 2008].

Pseudospectra for flow over a backward facing step

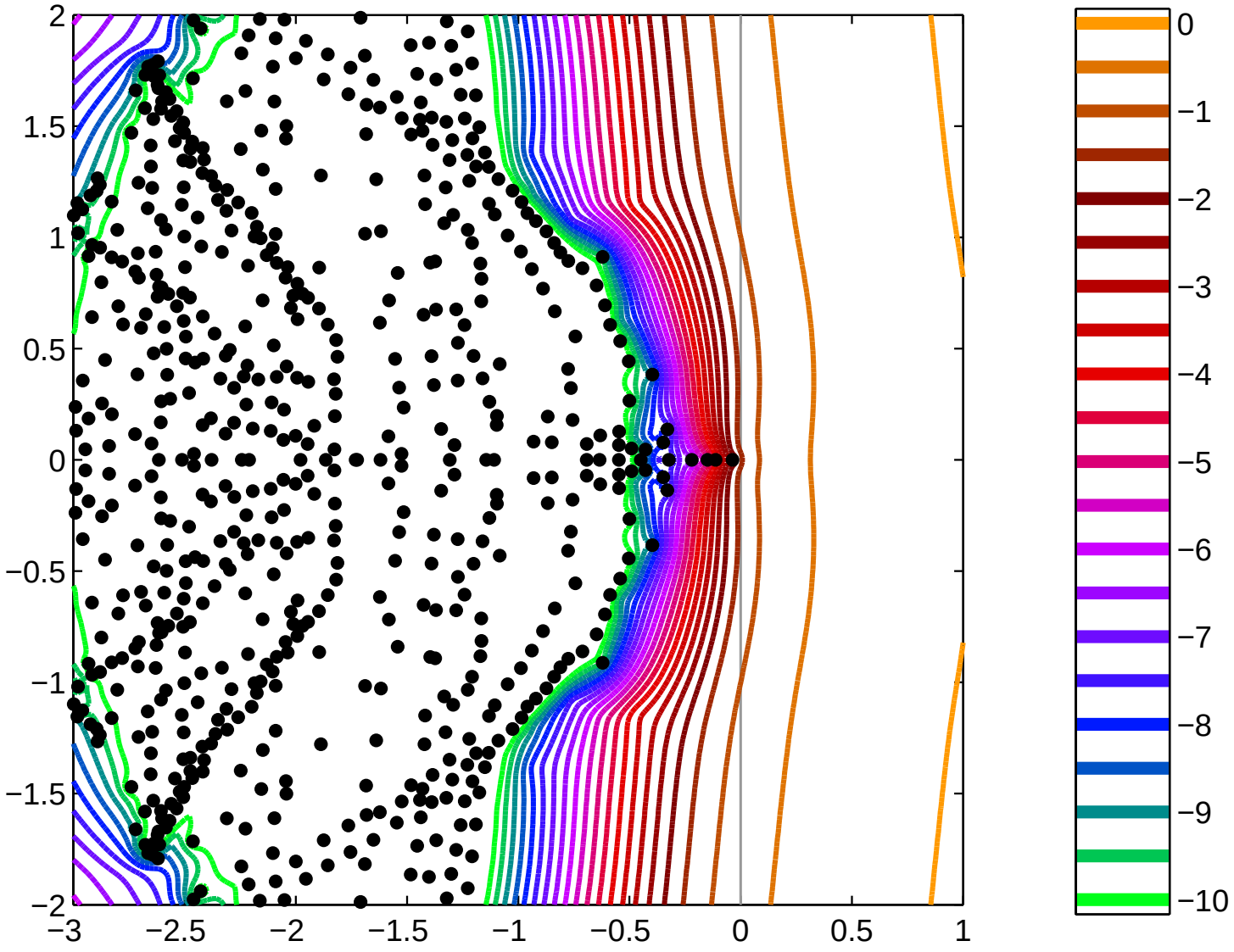
$\nu = 1/100$



[E., Keeler]

Pseudospectra for flow over a backward facing step

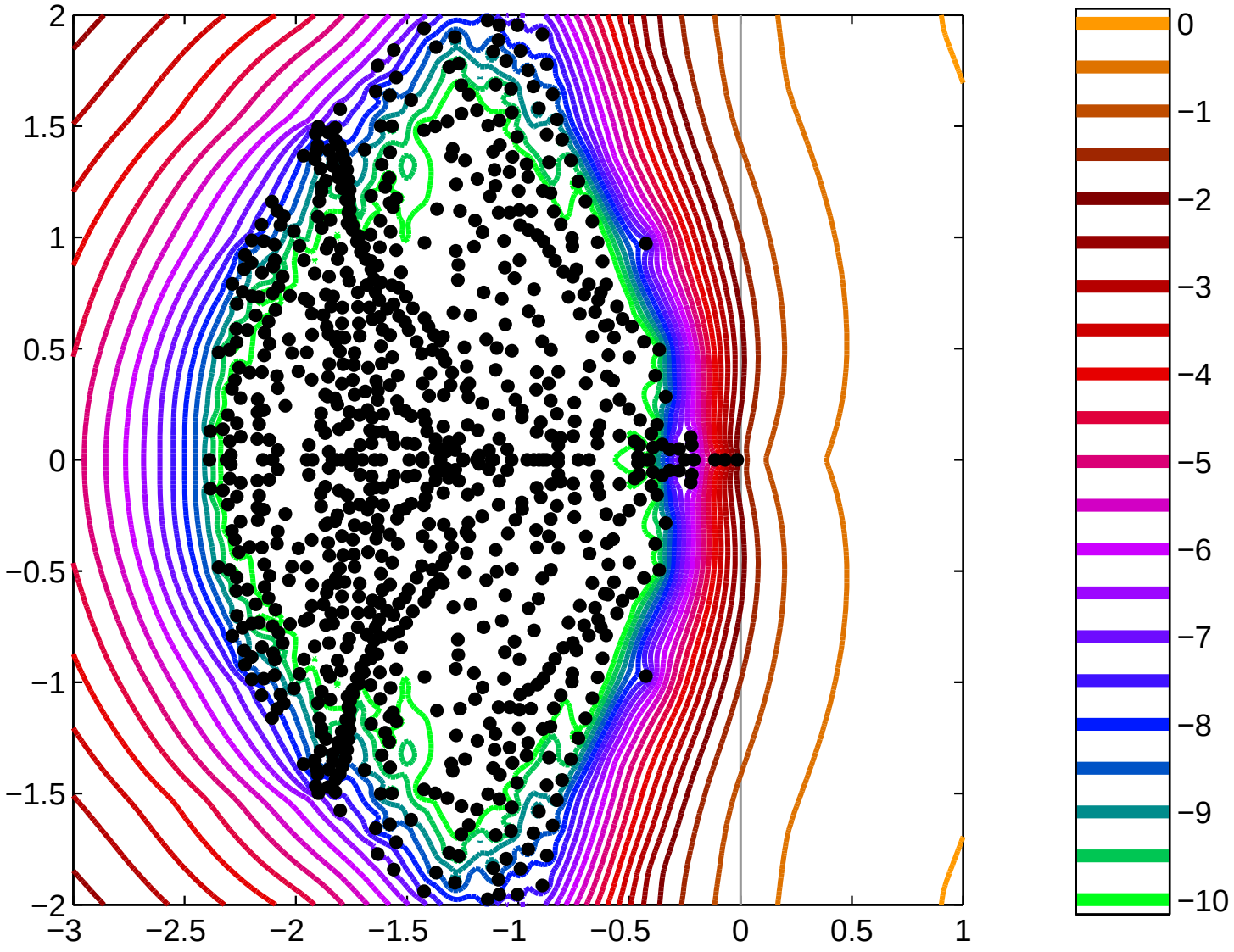
$\nu = 1/200$



[E., Keeler]

Pseudospectra for flow over a backward facing step

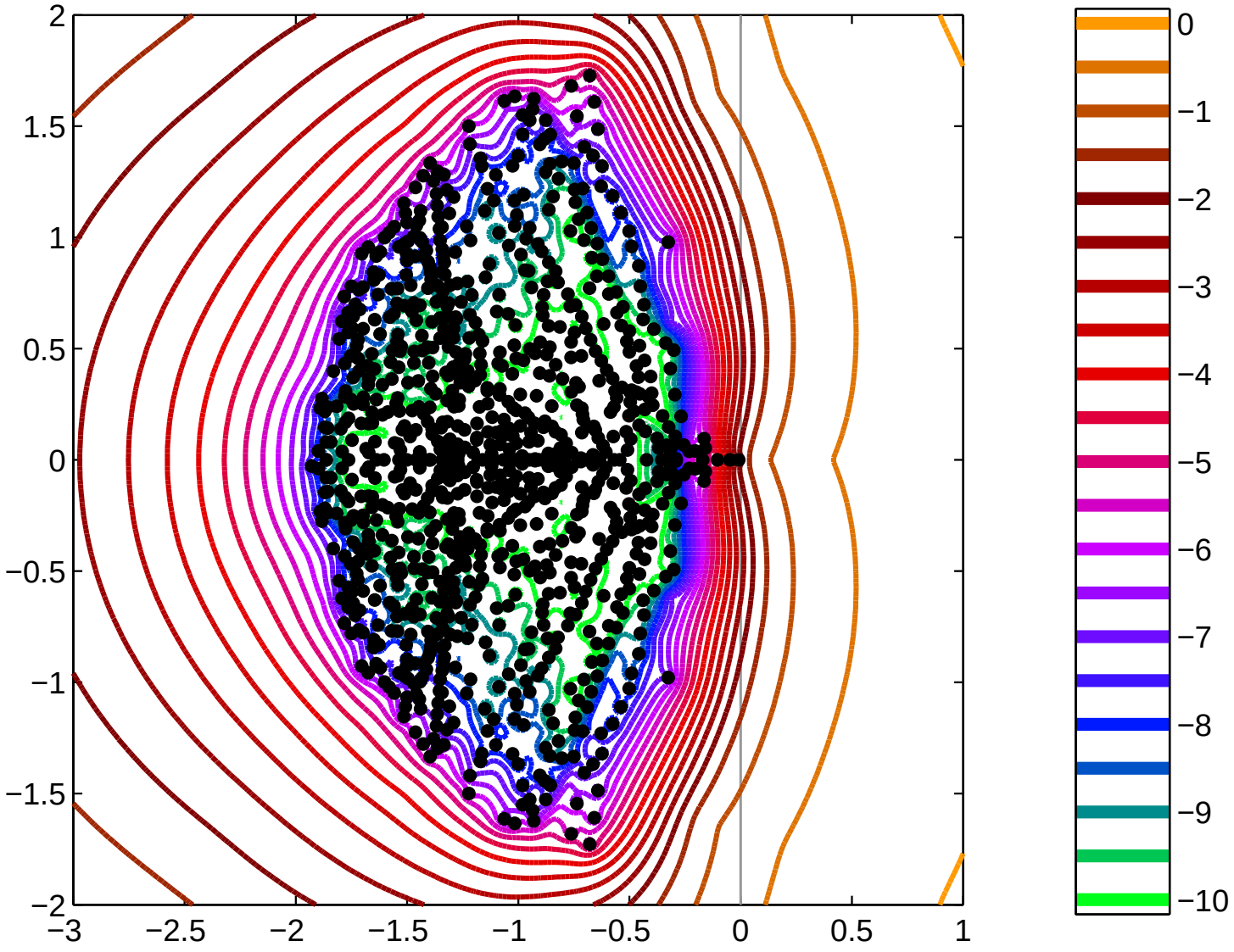
$\nu = 1/300$



[E., Keeler]

Pseudospectra for flow over a backward facing step

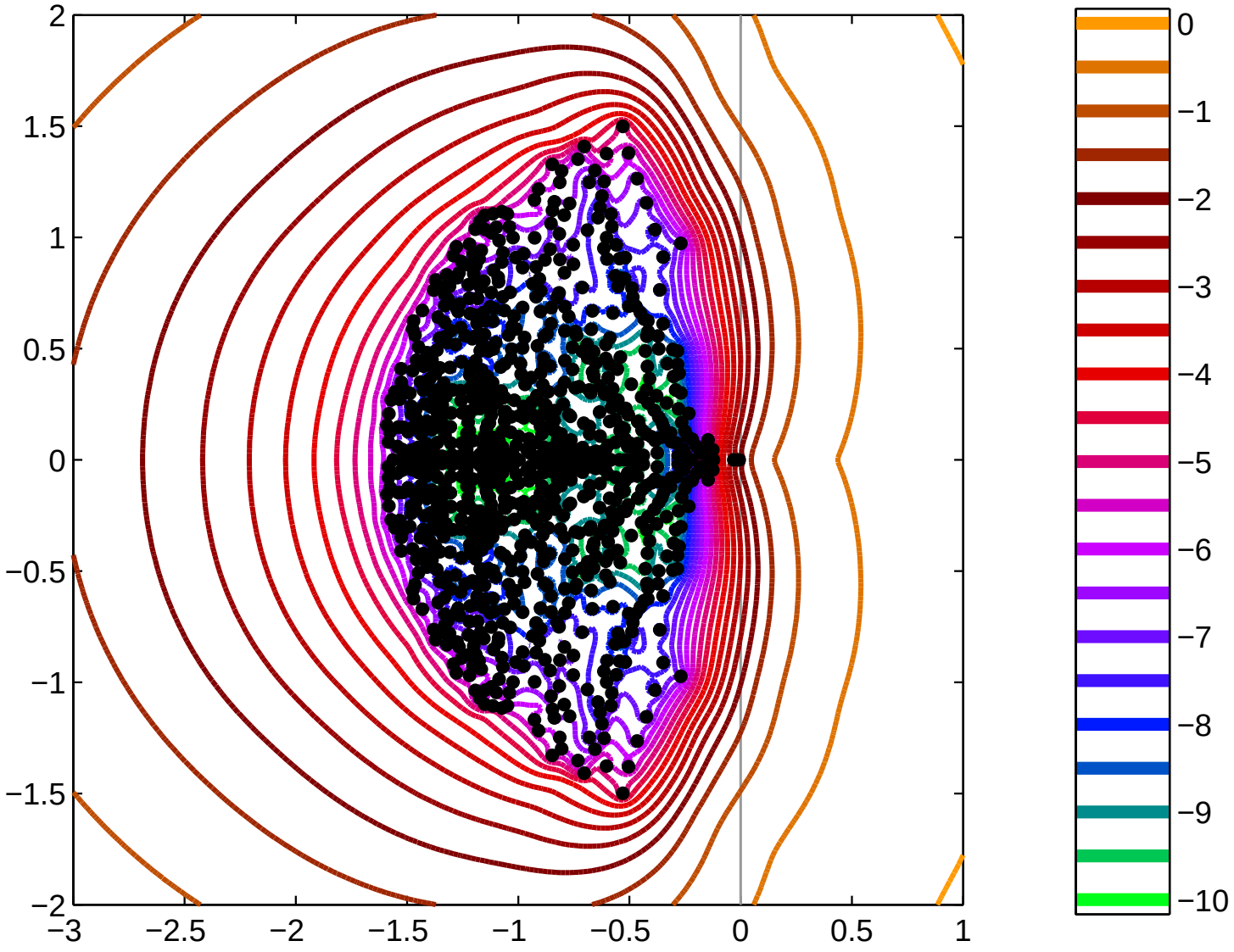
$\nu = 1/400$



[E., Keeler]

Pseudospectra for flow over a backward facing step

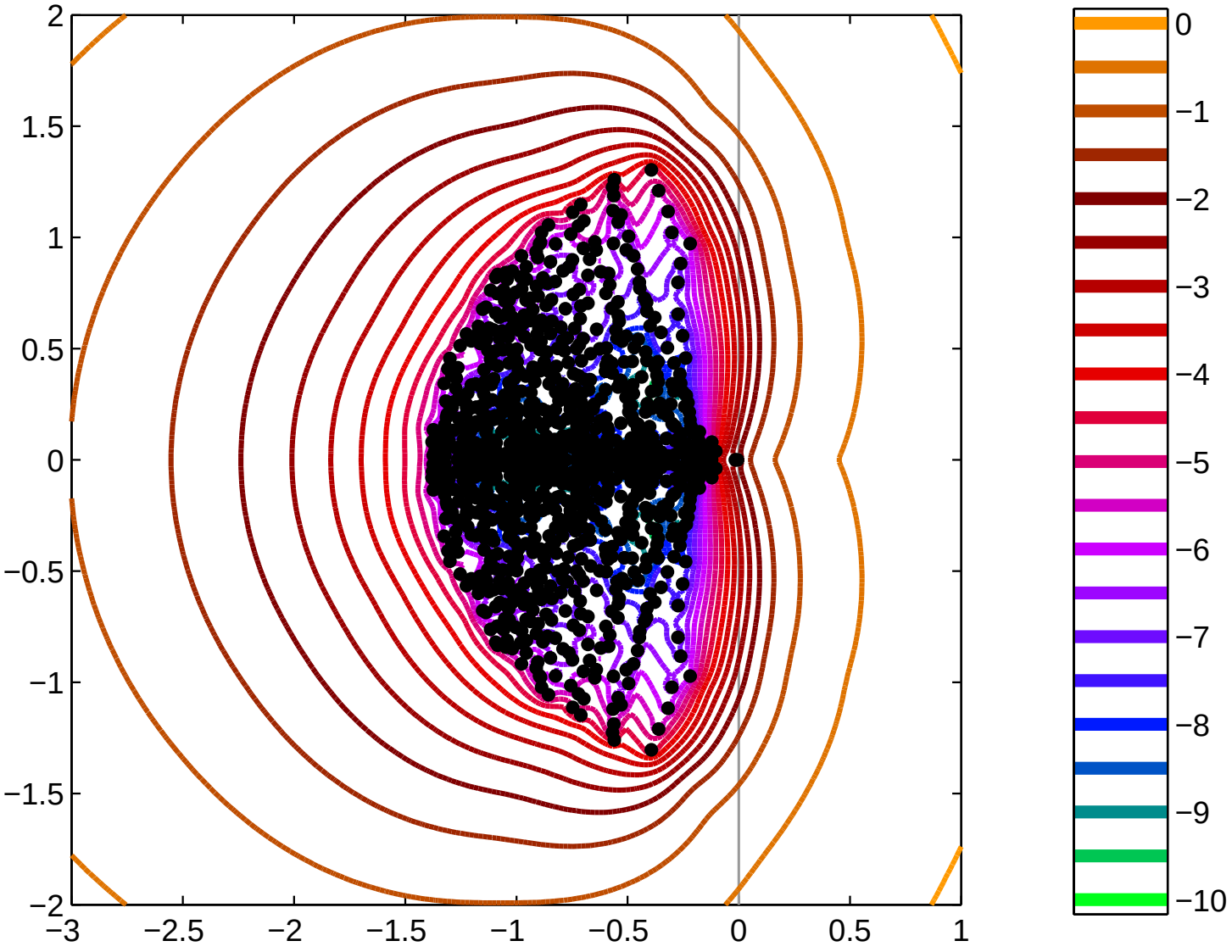
$\nu = 1/500$



[E., Keeler]

Pseudospectra for flow over a backward facing step

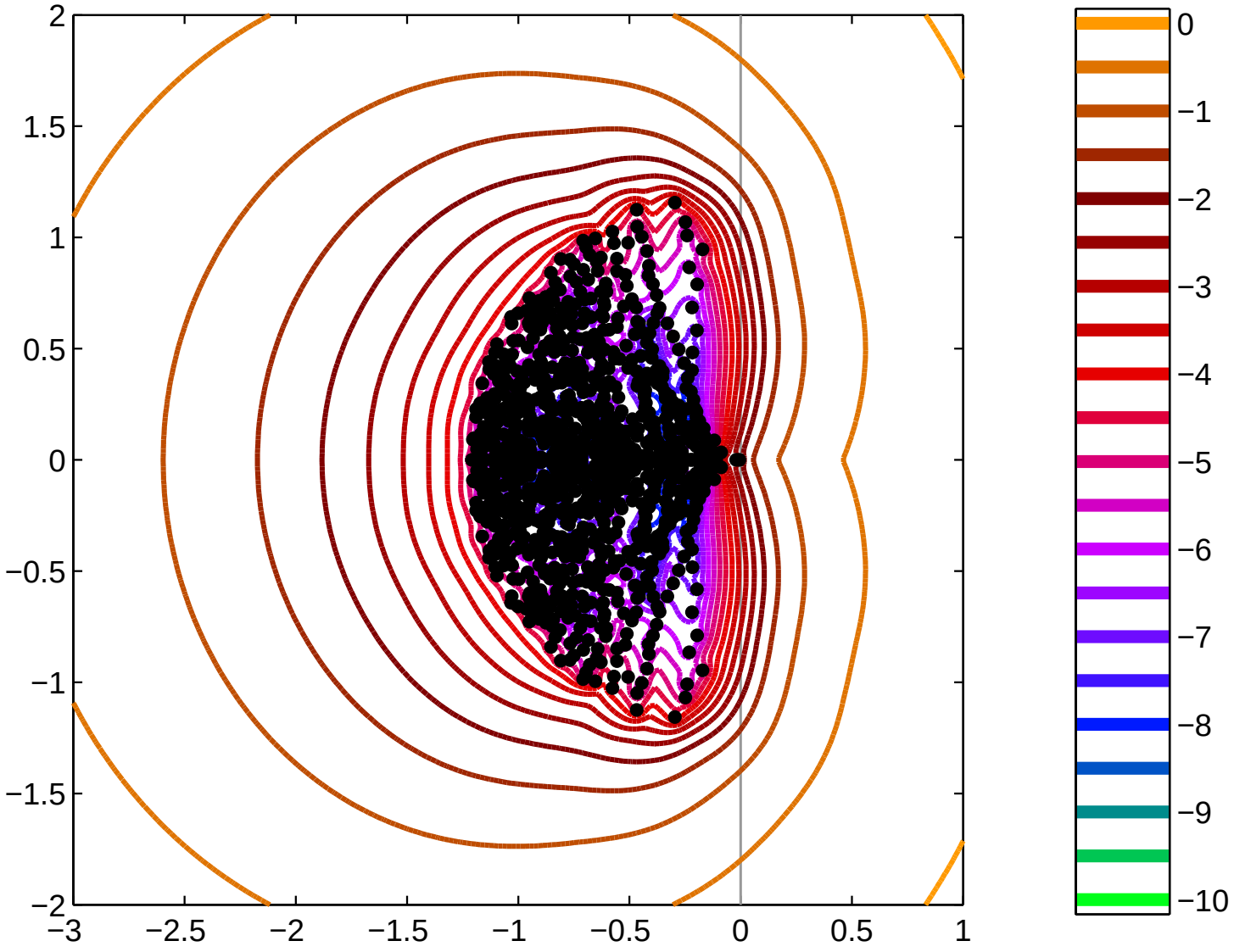
$\nu = 1/600$



[E., Keeler]

Pseudospectra for flow over a backward facing step

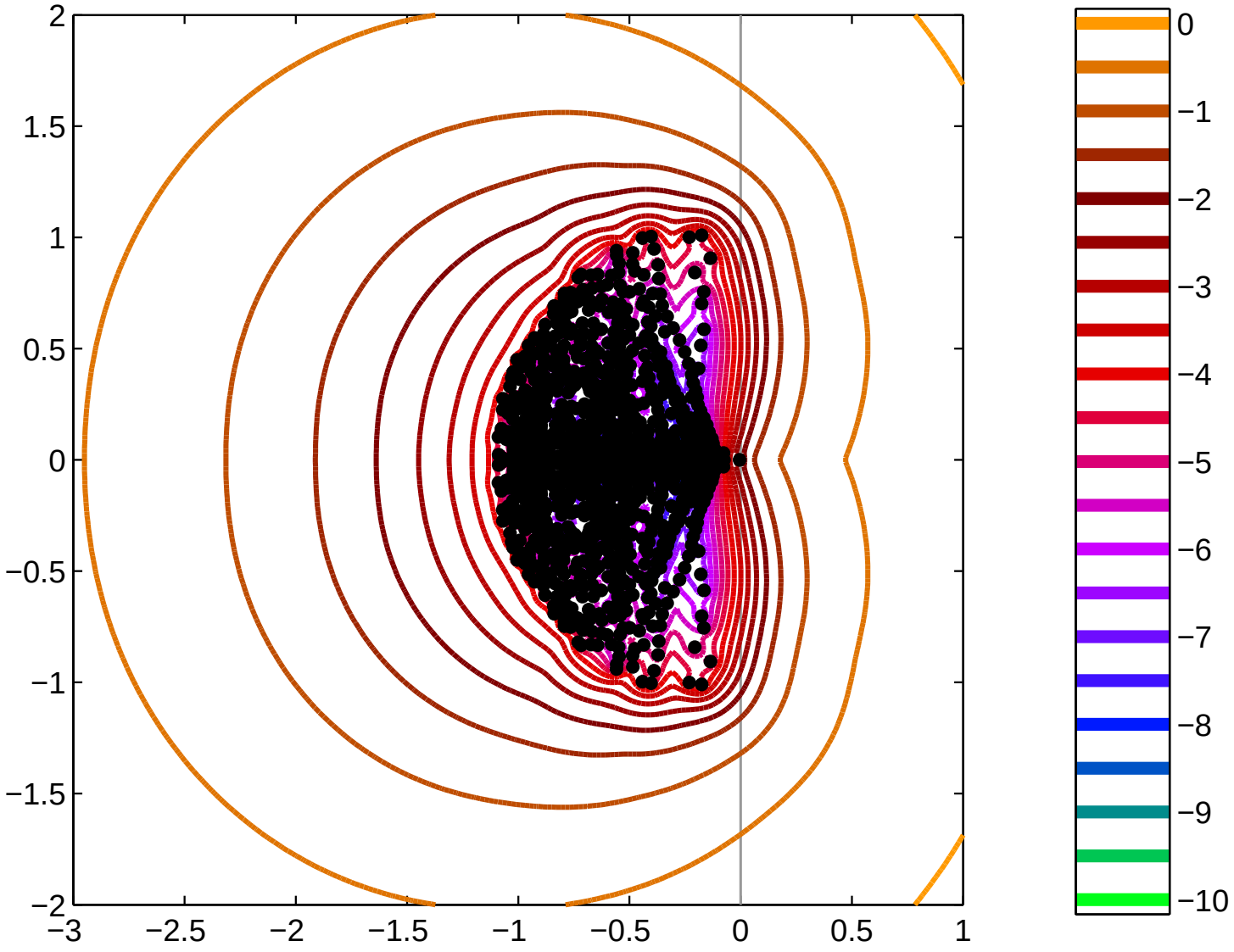
$\nu = 1/700$



[E., Keeler]

Pseudospectra for flow over a backward facing step

$\nu = 1/800$



[E., Keeler]

Singular Values of Solutions of Lyapunov Equations

Matrix Equations in Dynamical Systems

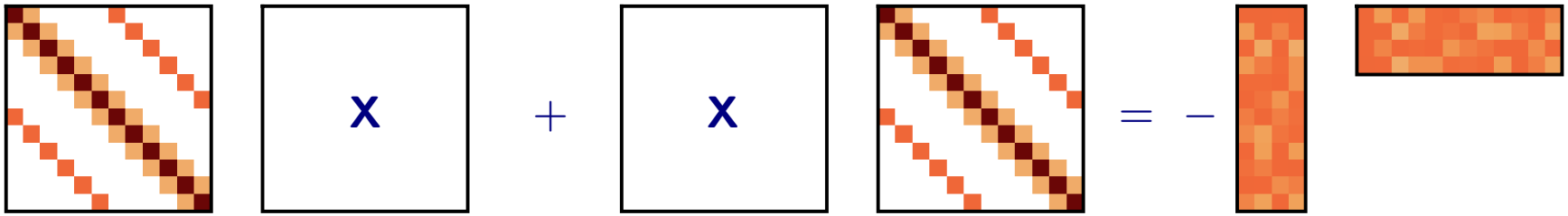
Many problems in model reduction, and control/dynamical systems in general, lead to matrix equations, the most common being the Lyapunov equation. (See the recent survey on linear matrix equations by [Simoncini].)
Assume that $\mathbf{A} \in \mathbf{C}^{n \times n}$ is stable: all eigenvalues have negative real part.

$$\boxed{\mathbf{A}}\ \boxed{\mathbf{X}} + \boxed{\mathbf{X}}\ \boxed{\mathbf{A}^*} = -\boxed{\mathbf{B}}\ \boxed{\mathbf{B}^*}$$

Given the $n \times n$ matrix $\mathbf{A}$ and the $n \times m$ matrix $\mathbf{B}$ ($m \ll n$), solve for the square $n \times n$ matrix $\mathbf{X}$.

Matrix Equations in Dynamical Systems

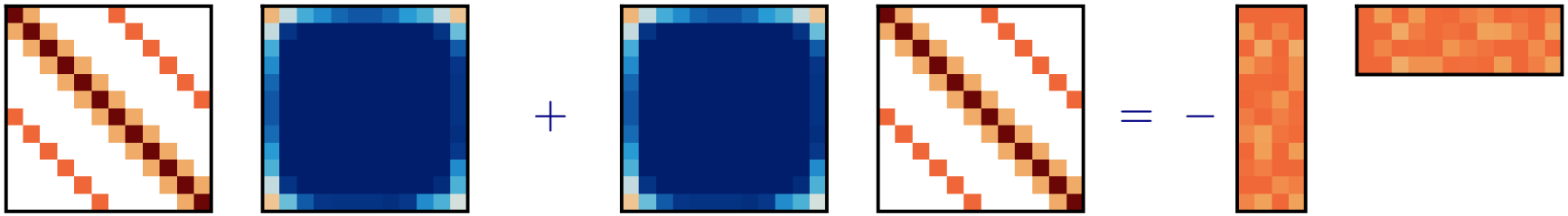
Many problems in model reduction, and control/dynamical systems in general, lead to matrix equations, the most common being the Lyapunov equation. (See the recent survey on linear matrix equations by [Simoncini].)
Assume that $\mathbf{A} \in \mathbf{C}^{n \times n}$ is stable: all eigenvalues have negative real part.



Given the $n \times n$ matrix $\mathbf{A}$ and the $n \times m$ matrix $\mathbf{B}$ ($m \ll n$), solve for the square $n \times n$ matrix $\mathbf{X}$.

Matrix Equations in Dynamical Systems

Many problems in model reduction, and control/dynamical systems in general, lead to matrix equations, the most common being the Lyapunov equation. (See the recent survey on linear matrix equations by [Simoncini].)
Assume that $\mathbf{A} \in \mathbf{C}^{n \times n}$ is stable: all eigenvalues have negative real part.

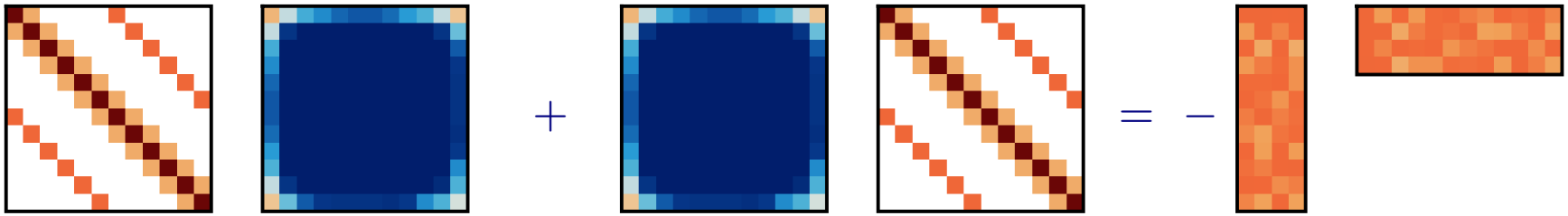


Given the $n \times n$ matrix $\mathbf{A}$ and the $n \times m$ matrix $\mathbf{B}$ ($m \ll n$), solve for the square $n \times n$ matrix $\mathbf{X}$.

- The solution $\mathbf{X}$ is a Hermitian matrix.
- Under mild conditions ($(\mathbf{A}, \mathbf{B})$ controllable), $\mathbf{X}$ is positive definite.
- Typically $\mathbf{X}$ has n^2 nonzeros: cannot directly store $\mathbf{X}$ for large n.

Matrix Equations in Dynamical Systems

Many problems in model reduction, and control/dynamical systems in general, lead to matrix equations, the most common being the Lyapunov equation. (See the recent survey on linear matrix equations by [Simoncini].)
Assume that $\mathbf{A} \in \mathbf{C}^{n \times n}$ is stable: all eigenvalues have negative real part.



Given the $n \times n$ matrix $\mathbf{A}$ and the $n \times m$ matrix $\mathbf{B}$ ($m \ll n$), solve for the square $n \times n$ matrix $\mathbf{X}$.

- The solution $\mathbf{X}$ is a Hermitian matrix.
- Under mild conditions ($(\mathbf{A}, \mathbf{B})$ controllable), $\mathbf{X}$ is positive definite.
- Typically $\mathbf{X}$ has n^2 nonzeros: cannot directly store $\mathbf{X}$ for large n.
- When m is small, *the singular values of* $\mathbf{X}$ *often decay quickly*, depending on eigenvalues of $\mathbf{A}$ (and related quantities) [Penzl 2000a, 2000b].

Low Rank Approximations from Iterative Methods

- How do spectral properties of $\mathbf{A}$ affect the singular values of $\mathbf{X}$?
- Iterative methods for solving the Lyapunov equation naturally construct low-rank approximations to $\mathbf{X}$. (Take $\text{rank}(\mathbf{B}) = 1$ for simplicity.)
- Denote the singular values of $\mathbf{X}$ by

$$s_1 \geq s_2 \geq \cdots \geq s_n > 0.$$

- Let $\mathbf{X}_k$ be a rank-k approximation to $\mathbf{X}_k$ (e.g., from Galerkin or ADI).
- Any bound on $\|\mathbf{X} - \mathbf{X}_k\|$ becomes a bound on s_{k+1} by the Schmidt–Mirsky–Eckart–Young theorem:

$$s_{k+1} \;=\; \min_{\text{rank}(\widehat{\mathbf{X}}) \leq k} \|\mathbf{X} - \widehat{\mathbf{X}}\| \;\leq\; \|\mathbf{X} - \mathbf{X}_k\|.$$

- Similarly, s_{k+1} bounds the best performance attainable by any iterative method that constructs a rank-k approximation $\mathbf{X}_k$. (This is helpful for understanding if the methods are near-optimal.)

Nonnormality and Singular Values Decay Bounds

All known bounds predict slower decay of singular values of $\mathbf{X}$.

$$\frac{s_{k+1}}{s_1} \leq \|\mathbf{V}\|^2 \, \|\mathbf{V}^{-1}\|^2 \max_{z \in \sigma(\mathbf{A})} \prod_{j=1}^{k} \frac{|z + \mu_k|^2}{|z - \overline{\mu_k}|^2}$$

$$\frac{s_{k+1}}{s_1} \leq C^2 \max_{z \in W(\mathbf{A})} \prod_{j=1}^{k} \frac{|z + \mu_k|^2}{|z - \overline{\mu_k}|^2}$$

$$\frac{s_{k+1}}{s_1} \leq \frac{L_\varepsilon^2}{4\pi^2 \varepsilon^2} \max_{z \in \sigma_\varepsilon(\mathbf{A})} \prod_{j=1}^{k} \frac{|z + \mu_k|^2}{|z - \overline{\mu_k}|^2}$$

Nonnormality and Singular Values Decay Bounds

All known bounds predict slower decay of singular values of $\mathbf{X}$.

$$\boxed{\frac{s_{k+1}}{s_1} \leq \|\mathbf{V}\|^2 \, \|\mathbf{V}^{-1}\|^2 \max_{z \in \sigma(\mathbf{A})} \prod_{j=1}^{k} \frac{|z+\mu_k|^2}{|z-\overline{\mu_k}|^2}}$$

$$\boxed{\frac{s_{k+1}}{s_1} \leq C^2 \max_{z \in W(\mathbf{A})} \prod_{j=1}^{k} \frac{|z+\mu_k|^2}{|z-\overline{\mu_k}|^2}}$$

$$\boxed{\frac{s_{k+1}}{s_1} \leq \frac{L_\varepsilon^2}{4\pi^2\varepsilon^2} \max_{z \in \sigma_\varepsilon(\mathbf{A})} \prod_{j=1}^{k} \frac{|z+\mu_k|^2}{|z-\overline{\mu_k}|^2}}$$

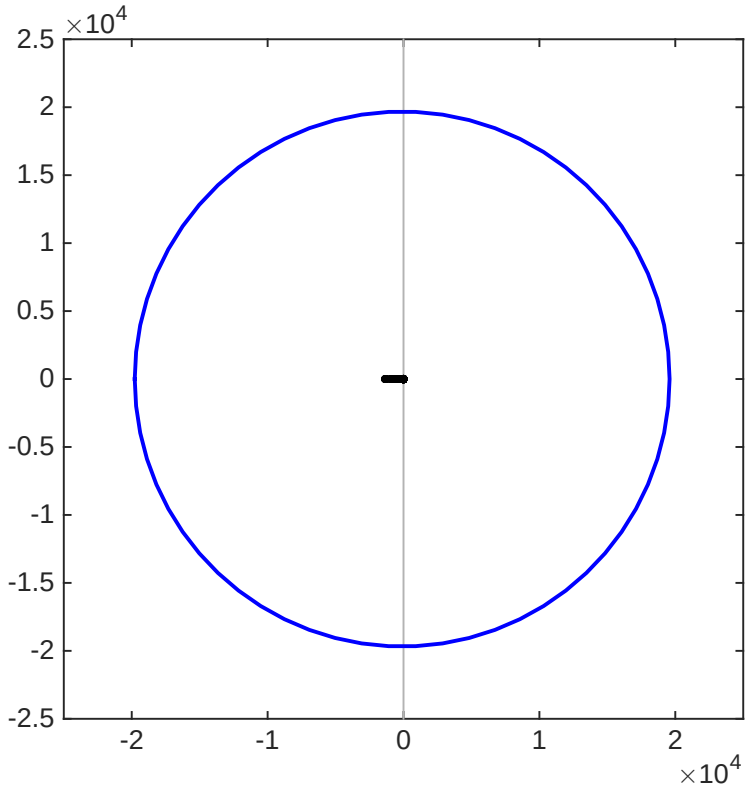
Consider this experiment:
Fix the spectrum $\sigma(\mathbf{A})$ but let *the departure of* $\mathbf{A}$ *from normality* increase.

- There are many essentially equivalent ways to measure departure from normality [Grone et al. 1987; Elsner & Paardekooper 1987].
- As the departure of $\mathbf{A}$ from normality increases, typically:
  – $\kappa(\mathbf{V})$ increases;
  – $W(\mathbf{A})$ gets larger;
  – $\sigma_\varepsilon(\mathbf{A})$ gets larger and/or $L_\varepsilon/(2\pi\varepsilon)$ increases.

An Example from Bifurcation Detection

An example from [Elman, Meerbergen, Spence, Wu, 2012; Elman, Wu, 2013]:

- 2d flow over an backward-facing step, viscosity $\nu = 1/400$, discretized using Q_2–Q_1 finite elements via IFISS [Elman, Silvester, Ramage].
- Problem can be recast as a standard Lyapunov inverse iteration problem (linearize about steady state; map infinite eigenvalues; invert mass matrix).
- The resulting matrix is *nondiagonalizable, and has a large numerical range,*

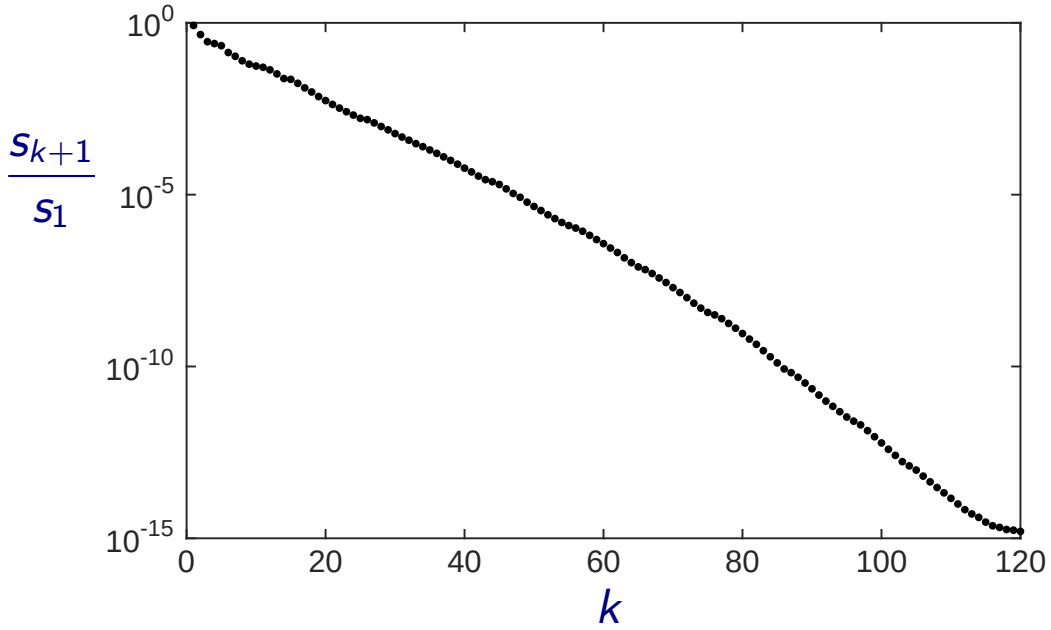


$W(\mathbf{A})$ and $\sigma(\mathbf{A})$

An Example from Bifurcation Detection

An example from [Elman, Meerbergen, Spence, Wu, 2012; Elman, Wu, 2013]:

- 2d flow over an backward-facing step, viscosity $\nu = 1/400$, discretized using Q_2–Q_1 finite elements via IFISS [Elman, Silvester, Ramage].
- Problem can be recast as a standard Lyapunov inverse iteration problem (linearize about steady state; map infinite eigenvalues; invert mass matrix).
- The resulting matrix is *nondiagonalizable, and has a large numerical range, but the singular values still decay very rapidly.*



The connection between $W(\mathbf{A})$ and $\frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$

The *Hermitian part of* $\mathbf{A}$ is $\frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$.

$$\begin{aligned} \text{eigenvalues of } \mathbf{A}: &\quad \lambda_1, \lambda_2, \ldots, \lambda_n \\ \text{eigenvalues of } \tfrac{1}{2}(\mathbf{A}+\mathbf{A}^*): &\quad \omega_n \le \omega_{n-1} \le \cdots \le \omega_1 \end{aligned}$$

Recall that the numerical range $W(\mathbf{A})$ is the set of all Rayleigh quotients:

$$W(\mathbf{A}) = \{\mathbf{v}^*\mathbf{A}\mathbf{v} : \|\mathbf{v}\| = 1\}.$$

Now if $z \in W(\mathbf{A})$, then

$$\operatorname{Re} z = \frac{z+z^*}{2} = \frac{\mathbf{v}^*\mathbf{A}\mathbf{v} + (\mathbf{v}^*\mathbf{A}^*\mathbf{v})^*}{2} = \mathbf{v}^*\left(\frac{\mathbf{A}+\mathbf{A}^*}{2}\right)\mathbf{v}.$$

The connection between $W(\mathbf{A})$ and $\frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$

The *Hermitian part of* $\mathbf{A}$ is $\frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$.

$$\begin{aligned} \text{eigenvalues of } \mathbf{A}: &\quad \lambda_1, \lambda_2, \ldots, \lambda_n \\ \text{eigenvalues of } \tfrac{1}{2}(\mathbf{A}+\mathbf{A}^*): &\quad \omega_n \le \omega_{n-1} \le \cdots \le \omega_1 \end{aligned}$$

Recall that the numerical range $W(\mathbf{A})$ is the set of all Rayleigh quotients:

$$W(\mathbf{A}) = \{\mathbf{v}^*\mathbf{A}\mathbf{v} : \|\mathbf{v}\| = 1\}.$$

Now if $z \in W(\mathbf{A})$, then

$$\operatorname{Re} z = \frac{z+z^*}{2} = \frac{\mathbf{v}^*\mathbf{A}\mathbf{v} + (\mathbf{v}^*\mathbf{A}^*\mathbf{v})^*}{2} = \mathbf{v}^*\Big(\frac{\mathbf{A}+\mathbf{A}^*}{2}\Big)\mathbf{v}.$$

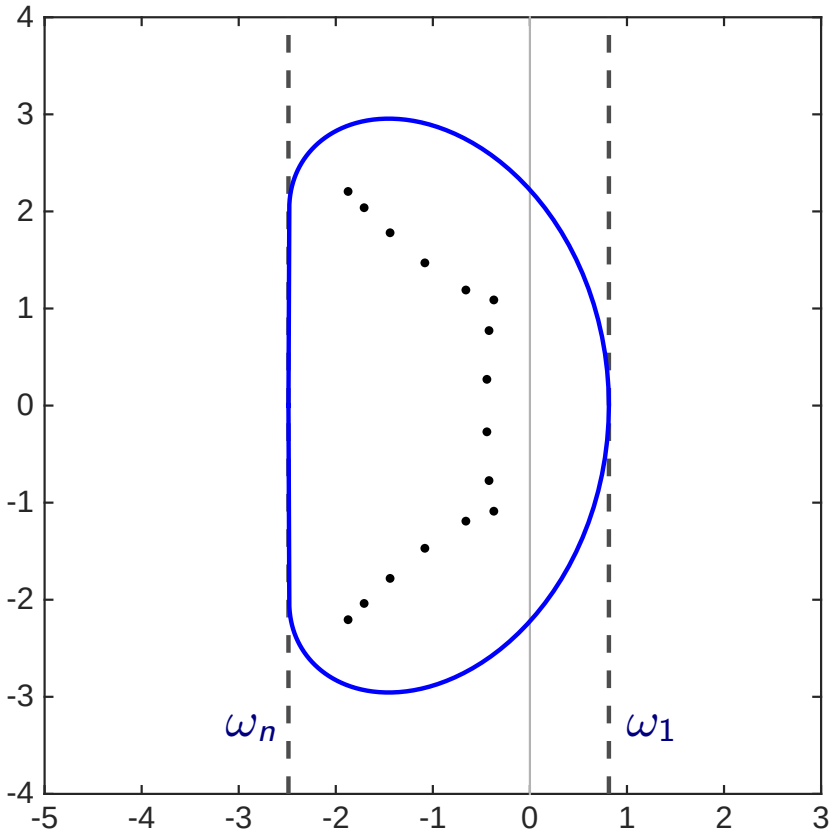
Hence the extreme eigenvalues of $\frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$ *dictate the real extent of* $W(\mathbf{A})$*:*

$$\operatorname{Re} W(\mathbf{A}) = [\omega_n, \omega_1].$$

The Connection Between $W(\mathbf{A})$ and $\frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$

The extreme eigenvalues of $\frac{1}{2}(\mathbf{A}+\mathbf{A}^*)$ dictate the real extent of $W(\mathbf{A})$:

$$\text{Re } W(\mathbf{A}) = [\omega_n, \omega_1].$$



$W(\mathbf{A})$ computed with Higham's Test Matrix Toolbox

An Extreme Example Illuminates: No Decay

- What properties of $\mathbf{A}$ permit a solution $\mathbf{X}$ with no singular value decay?

An Extreme Example Illuminates: No Decay

- What properties of $\mathbf{A}$ permit a solution $\mathbf{X}$ with no singular value decay?
- No decay $\Longrightarrow$ $\mathbf{X}$ is a Hermitian matrix with $s_1 = \cdots = s_n$, i.e.,

$$\mathbf{X} = \xi \mathbf{I}$$

  for some real $\xi > 0$.
- Substituting this $\mathbf{X}$ into the Lyapunov equation $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{BB}^*$,

$$\frac{1}{2}(\mathbf{A} + \mathbf{A}^*) = -\frac{1}{2\xi}\mathbf{BB}^*.$$

- $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$ is a negative semidefinite matrix of rank equal to rank($\mathbf{B}$).

An Extreme Example Illuminates: No Decay

- What properties of $\mathbf{A}$ permit a solution $\mathbf{X}$ with no singular value decay?
- No decay $\Longrightarrow$ $\mathbf{X}$ is a Hermitian matrix with $s_1 = \cdots = s_n$, i.e.,

$$\mathbf{X} = \xi \mathbf{I}$$

  for some real $\xi > 0$.
- Substituting this $\mathbf{X}$ into the Lyapunov equation $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{BB}^*$,

$$\frac{1}{2}(\mathbf{A} + \mathbf{A}^*) = -\frac{1}{2\xi}\mathbf{BB}^*.$$

- $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$ is a negative semidefinite matrix of rank equal to rank($\mathbf{B}$).

Worst case singular value decay $\iff$ $\operatorname{Re} W(\mathbf{A}) = [\omega_n, 0]$.

If $W(\mathbf{A})$ extends into the right-half plane, the singular values *must* decay.

Solvable Example: Jordan Block

An intriguing example from [Sabino 2006]:

$$\mathbf{A} = \begin{bmatrix} -1 & \alpha \\ 0 & -1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} t \\ 1 \end{bmatrix}.$$

Increasing α increases the distance of $\mathbf{A}$ from normality.

Solvable Example: Jordan Block

An intriguing example from [Sabino 2006]:

$$\mathbf{A} = \begin{bmatrix} -1 & \alpha \\ 0 & -1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} t \\ 1 \end{bmatrix}.$$

Increasing α increases the distance of $\mathbf{A}$ from normality.

The Lyapunov equation $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{BB}^*$ has solution

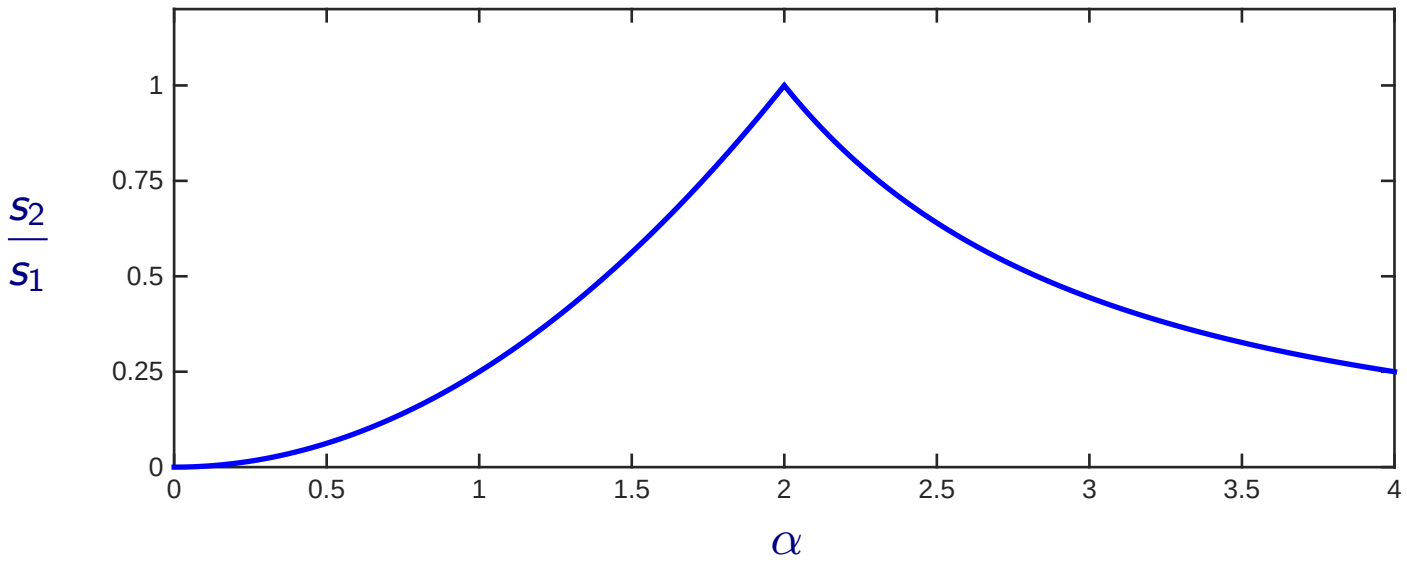
$$\mathbf{X} = \frac{1}{4} \begin{bmatrix} 2t^2 + 2\alpha t + \alpha^2 & \alpha + 2t \\ \alpha + 2t & 2 \end{bmatrix}.$$

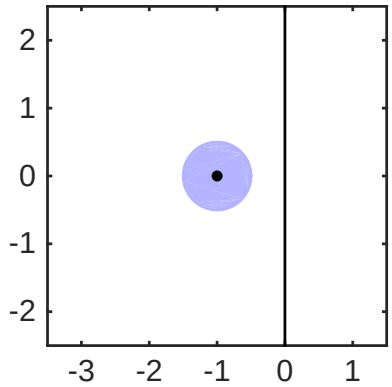
Maximizing over all $t \in \mathbb{R}$ gives the worst case singular value 'decay'

$$\frac{s_2}{s_1} = \frac{\mathrm{tr}(\mathbf{X}) - \sqrt{\mathrm{tr}(\mathbf{X})^2 - 4\,\det(\mathbf{X})}}{\mathrm{tr}(\mathbf{X}) + \sqrt{\mathrm{tr}(\mathbf{X})^2 - 4\,\det(\mathbf{X})}} = \begin{cases} \alpha^2/4, & 0 < \alpha \leq 2; \\ 4/\alpha^2, & 2 \leq \alpha. \end{cases}$$

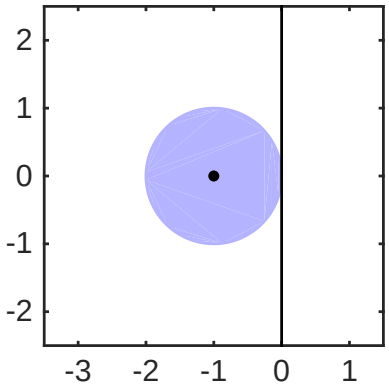
Solvable Example: Jordan Block

$$\frac{s_2}{s_1} = \begin{cases} \alpha^2/4, & 0 < \alpha \leq 2; \\ 4/\alpha^2, & 2 \leq \alpha. \end{cases}$$

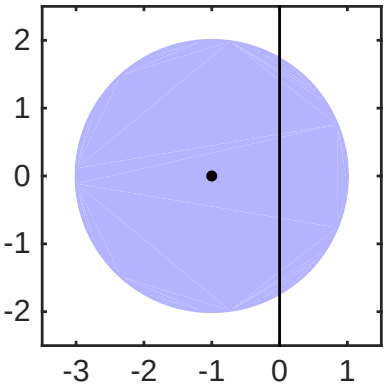




$\alpha = 1$



$\alpha = 2$



$\alpha = 4$

A More Nuanced Approach to Decay Bounds

If the singular values of $\mathbf{X}$ decay slowly, what must be true of $\mathbf{A}$?

Theorem (Baker, E., Sabino, 2014)

Suppose $\mathbf{A}$ *is a stable matrix with* $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{BB}^*$.
Let $s_1 \geq s_2 \geq \cdots \geq s_n$ *denote the singular values of of* $\mathbf{X}$,
and $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_n$ *denote the eigenvalues of* $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$. *Then*

$$\boxed{\frac{s_k}{s_1} - 1 - \frac{\|\mathbf{B}\|^2}{2s_1\|\mathbf{A}\|} \leq \frac{\omega_k}{\|\mathbf{A}\|} \leq 1 - \frac{s_{n-k+1}}{s_1}.}$$

This implies a bound on the trailing singular values:

$$\boxed{\frac{s_{n-k+1}}{s_1} \leq 1 - \frac{\omega_k}{\|\mathbf{A}\|},}$$

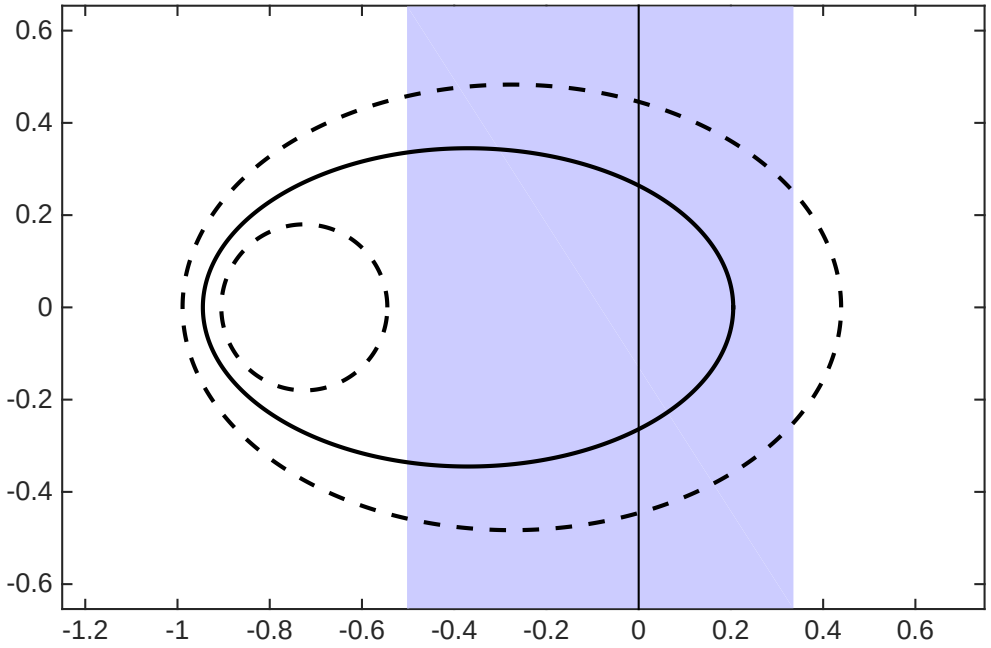
which gives faster singular value decay as the departure of $\mathbf{A}$ *from normality increases.*

Possible and Impossible $W(\mathbf{A})$

Corollary.

$$\boxed{-\frac{\|\mathbf{B}\|^2}{2s_1} \le \omega_1 \le \frac{s_1 - s_n}{s_1 + s_n}\|\mathbf{A}\|}$$

Suppose that $\|\mathbf{A}\| = \|\mathbf{B}\| = s_1 = 1$ and $s_n = 1/2$.



Given this data, *the two dashed curves are not possible boundaries of $W(\mathbf{A})$*, while the solid curve could be the boundary of $W(\mathbf{A})$.

Summary of Lyapunov Singular Value Decay

$$\boxed{\frac{s_{n-k+1}}{s_1} \leq 1 - \frac{\omega_k}{\|\mathbf{A}\|}.}$$

- The bound *does not depend on* rank($\mathbf{B}$).
- The departure from normality (as reflected by $\omega_k > 0$) plays a very different role from the previously known bounds.
- The bound is not necessarily sharp. Take $\alpha \to \infty$ in the Jordan example:

$$|\mathbf{A}\| \sim \alpha, \qquad \omega_1(\mathbf{A}) = \frac{\alpha}{2} - 1,$$

so

$$\frac{s_n}{s_1} \to 0 \qquad \text{while} \qquad 1 - \frac{\omega_1}{\|\mathbf{A}\|} \sim \frac{1}{2}.$$

- There is more to understand about the solutions to Lyapunov (and Sylvester) equations with coefficients that are far from normal.
- The eigenvalues of $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$ reveal a great deal!

Summary

- Linear Stability Analysis and Transient Dynamics
- Diagonalization and Davies's Conjecture
- Numerical Range
  – Crouzeix's Conjecture on $\|f(\mathbf{A})\|$
  – Inverse numerical range problems / localization of Ritz values
- Pseudospectra
  – "Do pseudospectra determine behavior?"
  – Pseudospectra of matrix pencils for Differential Algebraic Equations
- Lyapunov Equations
  – Singular values of Lyapunov solutions with nonnormal coefficients