Unstable modes in projection-based reduced-order models: How many can there be, and what do they tell you?

Mark Embree[1]

Department of Mathematics and Computational Modeling and Data Analytics Division, Academy of Integrated Science, Virginia Tech, Blacksburg, VA 24061, USA

ARTICLE INFO

ABSTRACT

Projection methods provide an appealing way to construct reduced-order models of large-scale linear dynamical systems: they are intuitively motivated and fairly easy to compute. Unfortunately, the resulting reduced models need not inherit the stability of the original system. How many unstable modes can these reduced models have? This note investigates this question, using theory originally motivated by iterative methods for linear algebraic systems and eigenvalue problems, and illustrating the theory with a number of small examples. From these results follow rigorous upper bounds on the number of unstable modes in reduced models generated via orthogonal projection, for both continuous- and discrete-time systems. Can anything be learned from the unstable modes in reduced-order models? Several examples illustrate how such instability can helpfully signal transient growth in the original system.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Reduced-order models are an enabling technology for simulation and design. One seeks simple low-order models that mimic the dynamics of a system with a high-dimensional state space. Asymptotic stability is the most fundamental property the reduced system should capture, but several popular algorithms can construct unstable reduced-order models for stable systems.

In this note, we investigate the potential instability of reduced-order models (ROMs) derived from projection methods. For simplicity of presentation, consider the standard continuous-time, single-input, single-output (SISO) linear system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \tag{1}$$
$$y(t) = \mathbf{c}^*\mathbf{x}(t) + du(t); \tag{2}$$

here $\mathbf{A} \in \mathbb{C}^{n \times n}$, $\mathbf{b}, \mathbf{c} \in \mathbb{C}^{n \times 1}$, and $d \in \mathbb{C}$ ($\cdot^*$ denotes the conjugate-transpose). For details about projection-based reduced-order modeling, see, e.g., [1].

Orthogonal projection algorithms restrict the state to evolve in the k-dimensional subspace $\mathcal{V} \subset \mathbb{C}^n$, then close the system by imposing a *Galerkin condition*: the reduced system's misfit should be orthogonal to the subspace $\mathcal{V}$. While the choice of $\mathcal{V}$ is crucial to the quality and properties of the resulting ROM, many of the results

we discuss apply to any choice of $\mathcal{V}$ (including $\mathcal{V}$ derived from moment-matching reduction of multi-input, multi-output (MIMO) systems).

Let the columns of $\mathbf{V} \in \mathbb{C}^{n \times k}$ form an orthonormal basis for $\mathcal{V}$, so $\mathbf{V}^*\mathbf{V} = \mathbf{I}$. To reduce the dimension of the system (1)–(2), approximate $\mathbf{x}(t) \approx \mathbf{V}\hat{\mathbf{x}}(t) \in \mathcal{V}$. One cannot simply replace $\mathbf{x}(t)$ by $\mathbf{V}\hat{\mathbf{x}}(t)$ in (1), since in general $\mathbf{A}\mathbf{V}\hat{\mathbf{x}}(t) + \mathbf{b}u(t) \notin \mathcal{V}$. To obtain a well-determined equation, impose the Galerkin condition

$$\mathbf{V}^*\left(\mathbf{V}\dot{\hat{\mathbf{x}}}(t) - \left(\mathbf{A}\mathbf{V}\hat{\mathbf{x}}(t) + \mathbf{b}u(t)\right)\right) = \mathbf{0}, \tag{3}$$

which yields the reduced system

$$\dot{\hat{\mathbf{x}}}(t) = (\mathbf{V}^*\mathbf{A}\mathbf{V})\hat{\mathbf{x}}(t) + (\mathbf{V}^*\mathbf{b})u(t) \tag{4}$$
$$\hat{y}(t) = (\mathbf{c}^*\mathbf{V})\hat{\mathbf{x}}(t) + du(t). \tag{5}$$

(Oblique projection methods impose the orthogonality in (3) against a different subspace; see Section 6 for details.)

For an effective ROM, one seeks a subspace $\mathcal{V}$ of smallest possible dimension k for which the reduced output $\hat{y}(t)$ mimics the true output $y(t)$, i.e., to make $\|y - \hat{y}\|$ small in an appropriate norm. Taking $\mathcal{V}$ to be a Krylov subspace gives particularly appealing properties, but the framework we describe also applies to the Galerkin proper orthogonal decomposition (POD) method (see, e.g., [2,3]) applied to a linear system; one might extrapolate some insight about the behavior of POD for nonlinear systems.

E-mail address: embree@vt.edu.

1.1. Moment-matching projection

The degree-k Krylov subspace generated by the matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ and vector $\mathbf{b} \in \mathbb{C}^n$ is

$$\mathcal{K}_k(\mathbf{A}, \mathbf{b}) = \text{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\}. \qquad (6)$$

Under mild conditions, $\dim(\mathcal{K}_k(\mathbf{A}, \mathbf{b})) = k$. If we take $\mathcal{V} = \mathcal{K}_k(\mathbf{A}, \mathbf{b})$ and let $\mathbf{V}$ have orthonormal columns, then *the resulting ROM matches the first k moments of the original model*:

$$\mathbf{c}^* \mathbf{A}^s \mathbf{b} = (\mathbf{c}^* \mathbf{V})(\mathbf{V}^* \mathbf{A} \mathbf{V})^s (\mathbf{V}^* \mathbf{b}), \quad s = 0, \ldots, k-1.$$

That is, the first k terms of the Taylor expansion of the transfer function

$$H(z) := \mathbf{c}^*(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$$

match those of the reduced transfer function

$$\widehat{H}(z) := (\mathbf{V}\mathbf{c})^*(z\mathbf{I} - \mathbf{V}^* \mathbf{A} \mathbf{V})^{-1}(\mathbf{V}^* \mathbf{b})$$

when expanded about $z = \infty$ [1, Section 11.2]. To better capture the frequency response about some finite point $\mu \in \mathbb{C}$, one can instead select

$$\mathcal{V} = \mathcal{K}_k((\mu\mathbf{I} - \mathbf{A})^{-1}, \mathbf{b})$$

to match moments at μ. (The oblique projection method based on the bi-Lanczos method addressed in Section 6 gives ROMs that match twice as many moments as these orthogonal projection Krylov methods, but this extra measure of accuracy can come at the cost of numerical challenges and additional unstable modes.)

This elegant moment matching gives a compelling motivation for Krylov projection methods, but these techniques have a crucial weakness: even when the matrix $\mathbf{A}$ is stable (all eigenvalues in the left half-plane), the reduced model $\mathbf{V}^* \mathbf{A} \mathbf{V}$ can have eigenvalues in the right half-plane: the original system is asymptotically stable (all solutions to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ converge to zero), but the ROM supports solutions that diverge as $t \to \infty$.

In this note we investigate this phenomenon, drawing on results that have been developed to explain the behavior of iterative methods for the solution of linear algebraic systems and eigenvalue problems. After reviewing spectral properties associated with transient dynamics in Section 2, in Sections 3 and 4 we give upper bounds on the number of unstable eigenvalues the reduced matrix $\mathbf{V}^* \mathbf{A} \mathbf{V}$ can have for continuous- and discrete-time systems. Sections 5 and 6 describe adversarial constructions that can produce many unstable modes for orthogonal and oblique projection methods. Throughout these sections, we illustrate theory with toy examples that are easy to analyze.

Are unstable modes merely a scourge? In Section 7 we argue that unstable modes can actually give valuable insight about the transient behavior of the original system, and efforts to tame these unstable modes can result in stable ROMs that fail to accurately model the short-term behavior of the original system.

Throughout, we use $\mathbf{e}_j$ to denote the jth column of the identity matrix (whose dimension should be clear from the context), and, unless noted otherwise, $\|\cdot\|$ to denote the vector 2-norm and the associated matrix norm.

2. Spectral preliminaries

Since we seek to understand the asymptotic and transient behavior of dynamical systems (both full- and reduced-order models), we review some helpful quantities associated with the spectrum. For more detailed descriptions and illustrative examples, see [4].

Denote the spectrum (set of eigenvalues) of $\mathbf{A}$ by

$$\sigma(\mathbf{A}) := \{\lambda_1, \ldots, \lambda_n\}.$$

Two scalar quantities dictate the asymptotic stability of continuous- and discrete-time systems, the *spectral abscissa* $\alpha(\mathbf{A})$ and the *spectral radius* $\rho(\mathbf{A})$:

$$\alpha(\mathbf{A}) := \max_{\lambda \in \sigma(\mathbf{A})} \text{Re}\, \lambda, \qquad \rho(\mathbf{A}) := \max_{\lambda \in \sigma(\mathbf{A})} |\lambda|.$$

The *numerical range* (or *field of values*)

$$W(\mathbf{A}) := \{\mathbf{v}^* \mathbf{A} \mathbf{v} : \mathbf{v} \in \mathbb{C}^n, \|\mathbf{v}\| = 1\} \qquad (7)$$

is a closed, convex subset of $\mathbb{C}$ that contains $\sigma(\mathbf{A})$; for details, see [5, Chapter 1]. We denote its maximal real extent and magnitude as the *numerical abscissa* $\omega(\mathbf{A})$ and the *numerical radius* $\nu(\mathbf{A})$:

$$\omega(\mathbf{A}) := \max_{z \in W(\mathbf{A})} \text{Re}\, z, \qquad \nu(\mathbf{A}) := \max_{z \in W(\mathbf{A})} |z|. \qquad (8)$$

For any $\varepsilon > 0$, the *ε-pseudospectrum* of $\mathbf{A}$,

$$\begin{aligned} \sigma_\varepsilon(\mathbf{A}) &:= \{z \in \mathbb{C} : \|(z\mathbf{I} - \mathbf{A})^{-1}\| > 1/\varepsilon\} \\ &= \{z \in \sigma(\mathbf{A} + \mathbf{E}) \text{ for some } \mathbf{E} \in \mathbb{C}^{n \times n} \text{with } \|\mathbf{E}\| < \varepsilon\}, \end{aligned}$$

contains $\sigma(\mathbf{A})$, but also potentially points that are a distance much greater than ε from the spectrum. To analyze transient behavior of solutions to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$, we can use the *$\varepsilon$-pseudospectral abscissa* $\alpha_\varepsilon(\mathbf{A})$ and the *ε-pseudospectral radius* $\rho_\varepsilon(\mathbf{A})$:

$$\alpha_\varepsilon(\mathbf{A}) := \sup z \in \sigma_\varepsilon(\mathbf{A}) \text{Re}\, z, \qquad \rho_\varepsilon(\mathbf{A}) := \sup_{z \in \sigma_\varepsilon(\mathbf{A})} |z|. \qquad (9)$$

A theorem of Stone (see [4, eq. (17.9)]) shows that the ε-pseudospectrum cannot be more than ε larger than the numerical range:

$$\sigma_\varepsilon(\mathbf{A}) \subseteq W(\mathbf{A}) + \Delta_\varepsilon,$$

where $\Delta_\varepsilon = \{z \in \mathbb{C} : |z| < \varepsilon\}$ is the open ball of radius ε.

(The definition of $\sigma_\varepsilon(\mathbf{A})$ permits general perturbations $\mathbf{E} \in \mathbb{C}^{n \times n}$. If $\mathbf{A}$ is real valued, $\mathbf{A} \in \mathbb{R}^{n \times n}$, might one gain insight by restricting perturbations to $\mathbf{E} \in \mathbb{R}^{n \times n}$? This question motivates the study of *structured pseudospectra*, or *spectral value sets* [6–8]. Considering only real perturbations can significantly reduce the set $\sigma_\varepsilon(\mathbf{A})$, but cannot improve the condition number of any eigenvalue by more than a factor of $1/\sqrt{2}$ [9]. For analyzing transient behavior of a linear system, Example (49.9) in [4] shows that complex perturbations are necessary to reveal the potential for transient growth of real-valued linear systems.)

We seek to use the sets $\sigma(\mathbf{A})$, $W(\mathbf{A})$, and $\sigma_\varepsilon(\mathbf{A})$ to gain insight into projection-based ROMs. A class of matrices is worth singling out for their clean properties: a matrix is *normal* provided $\mathbf{A}^* \mathbf{A} = \mathbf{A}\mathbf{A}^*$. Equivalently, a normal matrix has a unitary basis of eigenvectors. This latter property makes it easy to show that

$$\mathbf{A} \text{ normal} \implies \begin{cases} W(\mathbf{A}) = \text{ convex hull of } \sigma(\mathbf{A}); \\ \sigma_\varepsilon(\mathbf{A}) = \sigma(\mathbf{A}) + \Delta_\varepsilon. \end{cases}$$

(Hermitian ($\mathbf{A} = \mathbf{A}^*$), skew-Hermitian ($\mathbf{A} = -\mathbf{A}^*$), and unitary ($\mathbf{A}^* \mathbf{A} = \mathbf{I}$) matrices are all normal.) We will refer to the "departure from normality" as a gauge of how far a matrix is from the set of normal matrices.

2.1. Potential for unstable modes

We shall say that a continuous-time system is *stable* (i.e., *asymptotically stable*) provided

$$\alpha(\mathbf{A}) < 0,$$

i.e., $\sigma(\mathbf{A})$ is contained in the open left half of the complex plane. This condition implies that all solutions $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}(0)$ to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ converge to zero as $t \to \infty$. Similarly, a discrete-time system is *stable* provided

$$\rho(\mathbf{A}) < 1,$$

i.e., $\sigma(\mathbf{A})$ is contained in the open unit disk, so all solutions $\mathbf{x}_k = \mathbf{A}^k \mathbf{x}_0$ to $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$ converge to zero as $k \to \infty$.

Where can eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$ fall, relative to these spectral quantities associated with $\mathbf{A}$? We begin with a fundamental property of non-Hermitian eigenvalue approximation; cf. [5, Prop. 1.2.13], [10, Thm. 3.1].

Proposition 1. *Suppose the columns of $\mathbf{V} \in \mathbb{C}^{n \times k}$ are orthonormal. Then*

$$\sigma(\mathbf{V}^*\mathbf{A}\mathbf{V}) \subseteq W(\mathbf{A}), \qquad (10)$$

and so any $\theta \in \sigma(\mathbf{V}^*\mathbf{A}\mathbf{V})$ must satisfy

$$\mathrm{Re}\,\theta \leq \omega(\mathbf{A}), \qquad |\theta| \leq \nu(\mathbf{A}).$$

The proof of (10) is simple: if $\theta \in \sigma(\mathbf{V}^*\mathbf{A}\mathbf{V})$, there exists a unit vector $\mathbf{y}$ such that $(\mathbf{V}^*\mathbf{A}\mathbf{V})\mathbf{y} = \theta\mathbf{y}$. Then $\|\mathbf{V}\mathbf{y}\|^2 = \|\mathbf{y}\|^2 = \mathbf{y}^*\mathbf{y} = 1$ since $\mathbf{V}$ has orthonormal columns, and

$$(\mathbf{V}\mathbf{y})^*\mathbf{A}(\mathbf{V}\mathbf{y}) = \mathbf{y}^*\mathbf{V}^*\mathbf{A}\mathbf{V}\mathbf{y} = \theta\,\mathbf{y}^*\mathbf{y} = \theta;$$

use the definition (7) to conclude that $\theta \in W(\mathbf{A})$.

It follows that if $W(\mathbf{A})$ is contained in the left half-plane (i.e., $\omega(\mathbf{A}) < 0$), then $\mathbf{V}^*\mathbf{A}\mathbf{V}$ is guaranteed to be stable. All stable normal matrices satisfy this property: if $\sigma(\mathbf{A})$ is contained in the left half-plane, so too is its convex hull, which equals $W(\mathbf{A})$ for normal $\mathbf{A}$.

Proposition 2. *If $\mathbf{A}$ is stable and normal, then $\mathbf{V}^*\mathbf{A}\mathbf{V}$ is stable for any choice of the subspace $\mathcal{V}$.*

Unfortunately, for many interesting stable models we find that $\omega(\mathbf{A}) > 0$. These are the matrices in which we are primarily concerned here.

2.2. Transient behavior

The numerical abscissa $\omega(\mathbf{A})$ does not simply bound the rightmost extent of $\theta \in \sigma(\mathbf{V}^*\mathbf{A}\mathbf{V})$, as in Proposition 1; it also signals whether solutions $e^{t\mathbf{A}}\mathbf{x}(0)$ to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ can initially exhibit transient growth [4, Chapter 17]:

$$\max_{\|\mathbf{x}(0)\|=1} \left.\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{x}(t)\|\right|_{t=0} = \left.\frac{\mathrm{d}}{\mathrm{d}t}\|e^{t\mathbf{A}}\|\right|_{t=0} = \omega(\mathbf{A}). \qquad (11)$$

The possibility for stable systems to grow on transient time scales has important physical implications, especially for systems that arise as linearizations of nonlinear systems; see [4, Part V] for examples from fluid dynamics, and Section 7 for an example involving a nonlinear heat equation.

The formula (11) based on the numerical range describes the system's performance near $t = 0$; pseudospectra give insight into the maximum transient growth. The simplest result gives a lower bound [4, Theorem 15.4]:

$$\sup_{t \geq 0} \|e^{t\mathbf{A}}\| \geq \frac{\alpha_\varepsilon(\mathbf{A}, \mathbf{E})}{\varepsilon}$$

for all $\varepsilon > 0$. The sets $W(\mathbf{A})$ and $\sigma_\varepsilon(\mathbf{A})$ generalize to matrix pencils, informing the transient dynamics of differential–algebraic equations and descriptor systems [11].

3. An upper bound on unstable modes for orthogonal projection ROMs (continuous time case)

Let $\mathbf{A}$ be a stable matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ all satisfying $\mathrm{Re}\,\lambda_j < 0$, and let

$$\mathbf{V}^*\mathbf{A}\mathbf{V} \in \mathbb{C}^{k \times k}$$

denote an order-k ROM constructed via orthogonal projection. The columns of $\mathbf{V} \in \mathbb{C}^{n \times k}$ are orthonormal, but we make no assumptions about the projection subspace $\mathrm{range}(\mathbf{V})$; it could derive from a Krylov method, POD, or any other algorithm. For example, while we introduced moment matching model reduction for SISO systems, the results in this section also apply to moment matching for MIMO systems of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$
$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t),$$

where $\mathrm{range}(\mathbf{V}) = \mathrm{range}([\mathbf{B}, \mathbf{A}\mathbf{B}, \ldots, \mathbf{A}^{\ell-1}\mathbf{B}])$ is a block Krylov subspace of dimension k.

We begin with a simple example that shows how $W(\mathbf{A})$ can extend into the right half-plane, even when $\mathbf{A}$ is stable. The Hermitian part of $\mathbf{A}$,

$$\mathbf{H} := \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^*),$$

plays a critical role in stability theory. Even when a non-Hermitian $\mathbf{A}$ is stable, $\mathbf{H}$ need not be. For example, for

$$\mathbf{A} = \begin{bmatrix} -1 & 4 \\ 0 & -1 \end{bmatrix}, \qquad \mathbf{H} = \begin{bmatrix} -1 & 2 \\ 2 & -1 \end{bmatrix},$$

giving

$$\sigma(\mathbf{A}) = \{-1, -1\}, \qquad \sigma(\mathbf{H}) = \{-3, +1\};$$

and hence $\mathbf{H}$ is not stable despite the stability of $\mathbf{A}$.

Since $\mathbf{H}$ is Hermitian, its eigenvalues are real. Label them in decreasing order as

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n.$$

Notice that $\mu_1 = \omega(\mathbf{A})$, the numerical abscissa (8) that describes the rightmost extent of the numerical range $W(\mathbf{A})$. To see this, take any $z \in W(\mathbf{A})$, for which there must exist some unit vector $\mathbf{v} \in \mathbb{C}^n$ such that $z = \mathbf{v}^*\mathbf{A}\mathbf{v}$. Then

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

Thus the real part of any z in the numerical range of $\mathbf{A}$ is a Rayleigh quotient for $\mathbf{H}$. By the variational characterization of eigenvalues of Hermitian matrices,

$$\mu_n \leq \mathrm{Re}\,z \leq \mu_1,$$

with equality attained when $\mathbf{v}$ is an eigenvector of $\mathbf{H}$ associated with μ_n or μ_1; see, e.g., [12, Theorem 4.2.6].

Via (11), the rightmost eigenvalue of $\mathbf{H}$ gives insight into the initial behavior of solutions to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$. Only recently has it been appreciated that the *interior eigenvalues of $\mathbf{H}$* help bound the eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$. We first state a result from [13, Theorem 2.1], which was developed to support convergence analysis for the restarted Arnoldi method for computing eigenvalues of large matrices. Theorem 1 establishes vertical strips in the complex plane where eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$ must fall. We follow the statement with a small example to illustrate its application.

Theorem 1. *Denote the eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V} \in \mathbb{C}^{k \times k}$ by $\theta_1, \ldots, \theta_k$, labeled by decreasing real part:*

$$\mathrm{Re}\,\theta_1 \geq \mathrm{Re}\,\theta_2 \geq \cdots \geq \mathrm{Re}\,\theta_k.$$

Let $M_{\pm j}$ denote the arithmetic mean of the j largest and smallest eigenvalues of $\mathbf{H}$,

$$M_j := \frac{\mu_1 + \cdots + \mu_j}{j}, \qquad 1 \leq j \leq n,$$
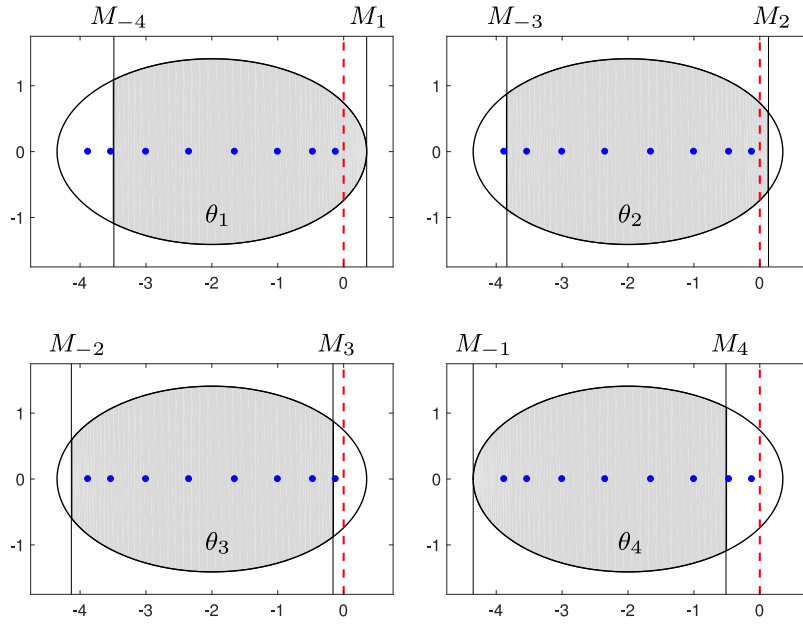$$M_{-j} := \frac{\mu_{n-j+1} + \cdots + \mu_n}{j}, \qquad 1 \leq j \leq n,$$

Fig. 1. Illustration of *Theorem* 1 applied to the matrix (13) in *Example* 1. The blue dots denote $\sigma(\mathbf{A})$; the solid oval-shaped curve in the complex plane is the boundary of the numerical range $W(\mathbf{A})$; the dashed red line shows the imaginary axis. The theorem ensures that the eigenvalue θ_j of $\mathbf{V}^*\mathbf{AV}$ falls within the associated jth gray subregion of $W(\mathbf{A})$. Since only the θ_1 and θ_2 regions extend into the right half-plane, $\mathbf{V}^*\mathbf{AV}$ can have *at most two unstable modes*.

so $M_1 \geq M_2 \geq \cdots \geq M_n$ and $M_{-1} \leq M_{-2} \leq \cdots \leq M_{-n}$. Then for $1 \leq j \leq k$,

$$M_{-k+j-1} \leq \operatorname{Re}\theta_j \leq M_j. \tag{12}$$

Thus, the jth rightmost eigenvalue θ_j of $\mathbf{V}^*\mathbf{AV}$ must fall in the intersection of $W(\mathbf{A})$ with the vertical strip

$$M_{-k+j-1} \leq \operatorname{Re}z \leq M_j.$$

Example 1. Consider the tridiagonal Toeplitz matrix

$$\mathbf{A} = \begin{bmatrix} -2 & 2 & & \\ 1/2 & -2 & \ddots & \\ & \ddots & \ddots & 2 \\ & & 1/2 & -2 \end{bmatrix} \in \mathbb{C}^{8\times 8}. \tag{13}$$

This matrix is stable, with negative eigenvalues[2]

$$\sigma(\mathbf{A}) = \{-2 + 2\cos(j\pi/9) : j = 1, \ldots, 8\}.$$

However, $W(\mathbf{A})$ extends into the right half-plane; indeed,

$$\sigma(\mathbf{H}) = \{-2 + \tfrac{5}{2}\cos(j\pi/9) : j = 1, \ldots, 8\},$$

and so $\mu_1 = \omega(\mathbf{A}) = 0.3492\ldots$. Suppose we seek a ROM of dimension $k = 4$. To five digits, we compute

$$\begin{array}{llll} M_1 = & 0.34923, & M_{-1} = & -4.34923, \\ M_2 = & 0.13217, & M_{-2} = & -4.13217, \\ M_3 = & -0.16189, & M_{-3} = & -3.83811, \\ M_4 = & -0.51288, & M_{-4} = & -3.48712. \end{array}$$

Since only two values of M_j are positive, Theorem 1 guarantees that *no more than two eigenvalues of* $\mathbf{V}^*\mathbf{AV}$ *can be in the right half-plane*. Fig. 1 illustrates $W(\mathbf{A})$ (bounded by the oval curve), with shaded regions indicating where Theorem 1 permits θ_j to fall. Since $M_1, M_2 > 0$, the theorem permits θ_1, θ_2 to be in the right half-plane; since $M_3 < 0$, all other eigenvalues θ_j for $j > 2$ *must* be in the left half-plane. (Computational experiments yield rare examples where θ_1 and $\theta_2 = \overline{\theta}_1$ fall in the right half-plane.)

---

2 For eigenvalues of tridiagonal Toeplitz matrices, see [14, p. 59].

Corollary 1. *Given the notation of Theorem* 1, *let p denote the largest integer such that* $M_p \geq 0$, *taking* $p = 0$ *if* $M_1 < 0$.

The orthogonal projection ROM $\mathbf{V}^*\mathbf{AV}$ *can never have more than p unstable modes (i.e., eigenvalues with nonnegative real part).*

Theorem 1 and Corollary 1 need not be sharp; they only give an upper bound on the maximal number of unstable modes. We have let $\mathbf{A}$ and $\mathbf{V}$ have general complex entries. Additional assumptions on these matrices could lead to sharper bounds on the eigenvalues of $\mathbf{V}^*\mathbf{AV}$. Most basically, if $\mathbf{A}$ and $\mathbf{V}$ have real entries, then complex eigenvalues of $\mathbf{V}^*\mathbf{AV}$ must occur in conjugate pairs.

The following theorem uses the eigenvalues $\{\mu_j\}$ of the Hermitian part $\mathbf{A}$ to get a lower bound on the maximal number of unstable modes.

Theorem 2. *Let q denote the number of positive eigenvalues of* $\mathbf{H} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$, *with* $0 \leq q \leq n$.

If $q \geq 1$, *there exists a q-dimensional subspace of* $\mathbb{C}^n$, *spanned by the orthonormal columns of* $\mathbf{V} \in \mathbb{C}^{n\times q}$, *such that all eigenvalues of* $\mathbf{V}^*\mathbf{AV}$ *are in the open right half-plane.*

Proof. The proof follows from an explicit construction. Once again denote the eigenvalues of $\mathbf{H}$ by $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$; label the associated orthonormal eigenvectors by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Construct the projection basis via

$$\mathbf{V} := [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_q] \in \mathbb{C}^{n\times q},$$

with $\mathbf{V}^*\mathbf{V} = \mathbf{I}$. We seek to show that $\mathbf{V}^*\mathbf{AV} \in \mathbb{C}^{q\times q}$ has q positive eigenvalues. Each $\theta \in \sigma(\mathbf{V}^*\mathbf{AV})$ has a unit eigenvector $\mathbf{y} = [y_1\ \cdots\ y_q]^T \in \mathbb{C}^q$ with $\mathbf{V}^*\mathbf{AVy} = \theta\mathbf{y}$. Notice that

$$\mathbf{Vy} = \sum_{j=1}^q y_j\mathbf{v}_j$$

and

$$\mathbf{y}^*\mathbf{V}^*\mathbf{AVy} = \theta\,\mathbf{y}^*\mathbf{y} = \theta.$$

Use these expressions to compute the real part of θ:

$$\operatorname{Re}\theta = \frac{\theta + \overline{\theta}}{2} = \frac{1}{2}\mathbf{y}^*\mathbf{V}^*(\mathbf{A} + \mathbf{A}^*)\mathbf{V}\mathbf{y}$$

$$= (\mathbf{V}\mathbf{y})^*\mathbf{H}(\mathbf{V}\mathbf{y}) = \sum_{j=1}^{q} \mu_j |y_j|^2,$$

where we have used the orthonormality of the eigenvector $\mathbf{v}_j$ of $\mathbf{H}$ for this last step. Since $\mu_1 \geq \cdots \geq \mu_q > 0$,

$$\operatorname{Re}\theta = \sum_{j=1}^{q} \mu_j |y_j|^2 \geq \mu_q \sum_{j=1}^{q} |y_j|^2 = \mu_q > 0.$$

Thus there exists $\mathbf{V} \in \mathbb{C}^{n \times q}$ with orthonormal columns for which all q eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$ are positive. ∎

This theorem suggests one way to design orthogonal projection subspaces $\mathcal{V}$ that yield ROMs with unstable modes, when the Hermitian part of $\mathbf{A}$ has positive eigenvalues. In Section 5, we shall see a different approach that takes $\mathcal{V}$ to be a Krylov subspace.

4. An upper bound on unstable modes for orthogonal projection ROMs (discrete time case)

A bound akin to Corollary 1 holds for orthogonal projection ROMs for the discrete-time system

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{b}u_k \tag{14}$$

$$y_{k+1} = \mathbf{c}^*\mathbf{x}_k + du_k. \tag{15}$$

Again let the columns of $\mathbf{V} \in \mathbb{C}^{n \times k}$ form an orthonormal basis for the subspace $\mathcal{V}$. Now assume that $\mathbf{A}$ is stable in the discrete-time sense, i.e., the spectral radius $\rho(\mathbf{A})$ is less than one. What can be said of the spectrum of $\mathbf{V}^*\mathbf{A}\mathbf{V}$?

In Section 3, arithmetic means of the eigenvalues of the Hermitian part of $\mathbf{A}$ bounded the real parts of the eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$. Now, *geometric* means of the *singular values* of $\mathbf{A}$ will bound the *magnitudes* of the eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$.

Let $s_1, \ldots, s_n$ denote the singular values of $\mathbf{A}$. We recall the following result from [13, Theorem 2.3].

Theorem 3. *Denote the eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V} \in \mathbb{C}^{k \times k}$ by $\theta_1, \ldots, \theta_k$, labeled by decreasing magnitude:*

$$|\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_k|.$$

Let G_j denote the geometric mean of the j largest singular values of $\mathbf{A}$,

$$G_j := (s_1 \cdots s_j)^{1/j}, \quad 1 \leq j \leq n. \tag{16}$$

Then

$$|\theta_j| \leq G_j, \quad 1 \leq j \leq n.$$

This theorem immediately gives a bound on the number of unstable modes in a discrete-time ROM generated via orthogonal projection.

Corollary 2. *Given the notation of Theorem 3, let p denote the largest index for which $G_j \geq 1$, taking $p = 0$ if $G_1 < 1$.*

The orthogonal projection ROM $\mathbf{V}^*\mathbf{A}\mathbf{V}$ for the discrete-time system (14)–(15) can never have more than p unstable modes (i.e., eigenvalues with magnitude at least one).

Example 2. To illustrate the bounds in Theorem 3 and Corollary 2, consider the stable matrix

$$\mathbf{A} = \begin{bmatrix} 1/2 & \gamma & & & \\ 1/8 & 1/2 & \gamma^2 & & \\ & 1/8 & \ddots & \ddots & \\ & & \ddots & 1/2 & \gamma^{n-1} \\ & & & 1/8 & 1/2 \end{bmatrix} \tag{17}$$

with $\gamma = 3/4$ and dimension $n = 128$. This matrix is stable, with $\rho(\mathbf{A}) = 0.94822\ldots$, but the numerical range extends beyond the unit disk, with the numerical radius $\mu(\mathbf{A}) = 1.09127\ldots$. What can be said of the stability of associated ROMs?

Label the eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V} \in \mathbb{C}^{k \times k}$ as $\theta_1, \ldots, \theta_k$, ordered by decreasing magnitude:

$$|\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_k|.$$

For each of these eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$ we know

$$\theta_j \in W(\mathbf{A}) \quad \text{and} \quad |\theta_j| \leq G_j.$$

Fig. 2 shows the regions

$$\Omega_j := W(\mathbf{A}) \cap \{z \in \mathbb{C} : |z| \leq G_j\}$$

for $j = 1, \ldots, 4$. By the monotonicity of the G_j values, these sets are nested:

$$\Omega_k \subseteq \cdots \subseteq \Omega_2 \subseteq \Omega_1.$$

To five digits, we compute

$$\begin{aligned} G_1 &= 1.13227 \\ G_2 &= 0.99258 \\ G_3 &= 0.90029 \\ G_4 &= 0.83738. \end{aligned}$$

Since $W(\mathbf{A})$ extends beyond the unit circle, it is possible that $|\theta_1| > 1$. However, since $G_2 < 1$, Corollary 2 ensures that *all other eigenvalues $\theta_2, \ldots, \theta_k$ of $\mathbf{V}^*\mathbf{A}\mathbf{V}$ must be contained within the unit disk: $\mathbf{V}^*\mathbf{A}\mathbf{V}$ can have at most one unstable mode.*

5. Orthogonal projection: adversarial construction

The last two sections describe how, for a given $\mathbf{A}$, one can get rigorous limits on the number of unstable modes in a ROM constructed using orthogonal projection from a generic subspace. Here we describe a construction for probing extreme limits of instability, provided one is content to only fix the eigenvalues of $\mathbf{A}$ but let the departure from normality vary. This result was proved by Duintjer Tebbens and Meurant [15, Corollary 2.3], a contribution to the convergence theory for Arnoldi's algorithm for computing eigenvalues; it builds on earlier work of Greenbaum, Pták, and Strakoš [16,17].

Theorem 4. *Let $\Sigma = \{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{C}$ denote a collection of desired eigenvalues, and specify any values for*

$$\begin{aligned} \Sigma_1 &:= \{\theta_1^{(1)}\}; \\ \Sigma_2 &:= \{\theta_1^{(2)}, \theta_2^{(2)}\}; \\ \Sigma_3 &:= \{\theta_1^{(3)}, \theta_2^{(3)}, \theta_3^{(3)}\}; \\ &\vdots \\ \Sigma_{n-1} &:= \{\theta_1^{(n-1)}, \theta_2^{(n-1)}, \ldots, \theta_{n-1}^{(n-1)}\}. \end{aligned}$$

There exists a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ and a vector $\mathbf{b} \in \mathbb{C}^n$ such that

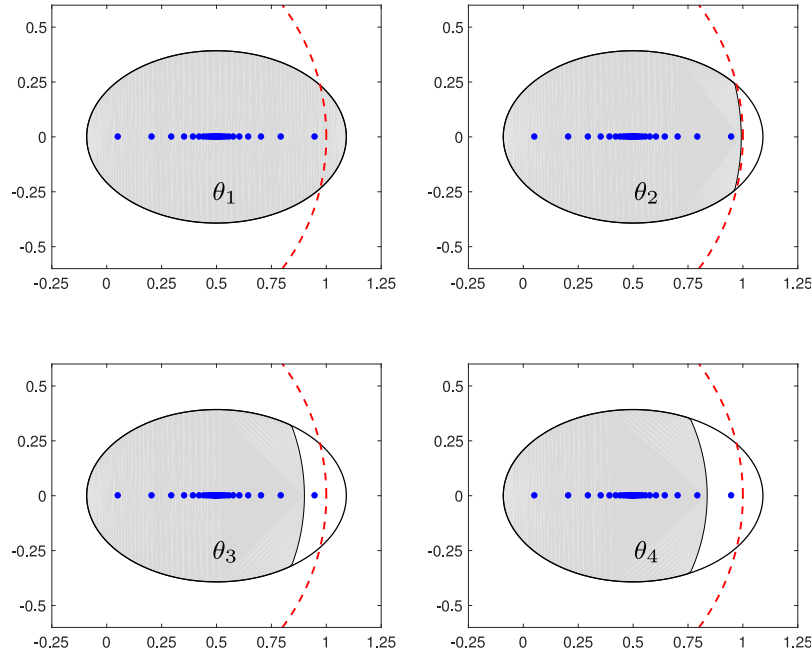$$\sigma(\mathbf{A}) = \Sigma = \{\lambda_1, \ldots, \lambda_n\}$$

Fig. 2. Illustration of Theorem 3 and Corollary 2 applied to the matrix (17). In each plot, the black oval shows the boundary of $W(\mathbf{A})$; the gray region shows Ω_j, which must contain θ_j; the blue dots show the eigenvalues of $\mathbf{A}$, and the red dashed line shows the boundary of the unit disk. Since Ω_j is contained in the unit disk for $j > 1$, Corollary 2 ensures $\mathbf{V}^*\mathbf{A}\mathbf{V}$ has at most one unstable mode.

and, for $k = 1, \ldots, n-1$,

$$\sigma(\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k) = \Sigma_k = \{\theta_1^{(k)}, \ldots, \theta_k^{(k)}\},$$

where the columns of $\mathbf{V}_k$ form an orthonormal basis for the Krylov subspace range$(\mathbf{V}_k) = \mathcal{K}_k(\mathbf{A}, \mathbf{b})$.

We can use this theorem (and its constructive proof) to build stable $\mathbf{A}$ and corresponding $\mathbf{b}$ for which *all eigenvalues of the orthogonal projection ROMs $\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$ from the Krylov subspace $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ fall at any desired location in the right half-plane*, for $k = 1, \ldots, n-1$, despite the fact that these ROMs all match k moments. The next example illustrates this point.

Example 3. Using the construction described by Duintjer Tebbens and Meurant [15, Proposition 2.1], we form

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & -362880 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & -1451520 \\ & 1 & 3 & 0 & 0 & 0 & 0 & -1693440 \\ & & 1 & 4 & 0 & 0 & 0 & -846720 \\ & & & 1 & 5 & 0 & 0 & -211680 \\ & & & & 1 & 6 & 0 & -28224 \\ & & & & & 1 & 7 & -2016 \\ & & & & & & 1 & -64 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The matrix $\mathbf{A}$ was constructed to have the stable spectrum

$$\sigma(\mathbf{A}) = \{-1, -2, -3, -4, -5, -6, -7, -8\}.$$

The form of $\mathbf{A}$ and $\mathbf{b}$ makes it easy to write down the associated Krylov subspace,

range$(\mathbf{V}_k) = \mathcal{K}_k(\mathbf{A}, \mathbf{b}) = \text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$,

so that $\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$ is the $k \times k$ principal submatrix of $\mathbf{A}$. Since these submatrices are lower triangular, one can easily read off their eigenvalues:

$$\sigma(\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k) = \{1, \ldots, k\}.$$

Though the kth order ROM matches k moments of the stable system, all k modes are unstable. Fig. 3 contrasts these unstable modes with the stable eigenvalues of $\mathbf{A}$.
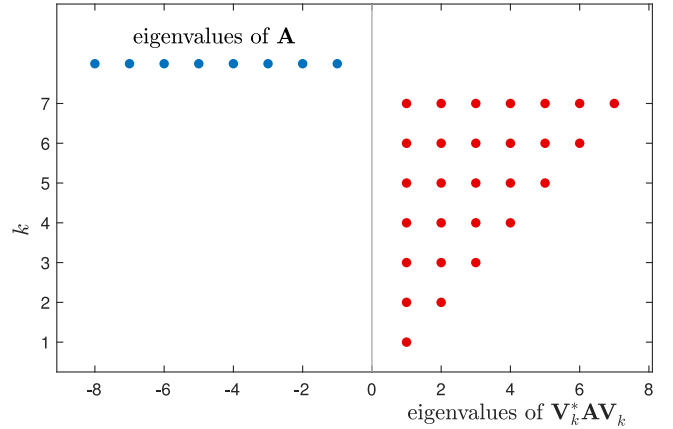

Fig. 3. Comparison of the eigenvalues of $\mathbf{A}$ (blue dots, in the left half-plane) to those of $\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$ of order $k = 1, \ldots, 7$ for the system in Example 3, generated using orthogonal projection onto Krylov subspaces. These ROMs are entirely unstable, even through the kth order ROM matches k moments of the original system.

This construction can deliver such startling results because *we only specify the eigenvalues of $\mathbf{A}$*. One might suspect that the $\mathbf{A}$ this construction produces might have a significant departure from normality, corresponding to transient growth of the dynamical system and eigenvalue instability. Indeed, Fig. 4 confirms this departure from normality. The numerical range $W(\mathbf{A})$ extends beyond 10^6 into the right half-plane ($\omega(\mathbf{A}) \approx 1.211 \times 10^6$), signaling rapid growth of $\|e^{t\mathbf{A}}\|$ for small t; see (11). The ε-pseudospectra reveal that $\mathbf{A}$ is close to an unstable system: since $\sigma_\varepsilon(\mathbf{A})$ extends into the right half-plane for $\varepsilon = 10^{-4}$, there exist matrices $\mathbf{E} \in \mathbb{C}^{8 \times 8}$ with $\|\mathbf{E}\| < 10^{-4}$ that make $\mathbf{A} + \mathbf{E}$ unstable.

In short, this pathological example corresponds to a special $\mathbf{A}$ with unusual dynamics and a fragile spectrum. (Our Σ and Σ_k are inspired by an ill-conditioned pole placement example of Mehrmann and Xu [19, Example 2].)
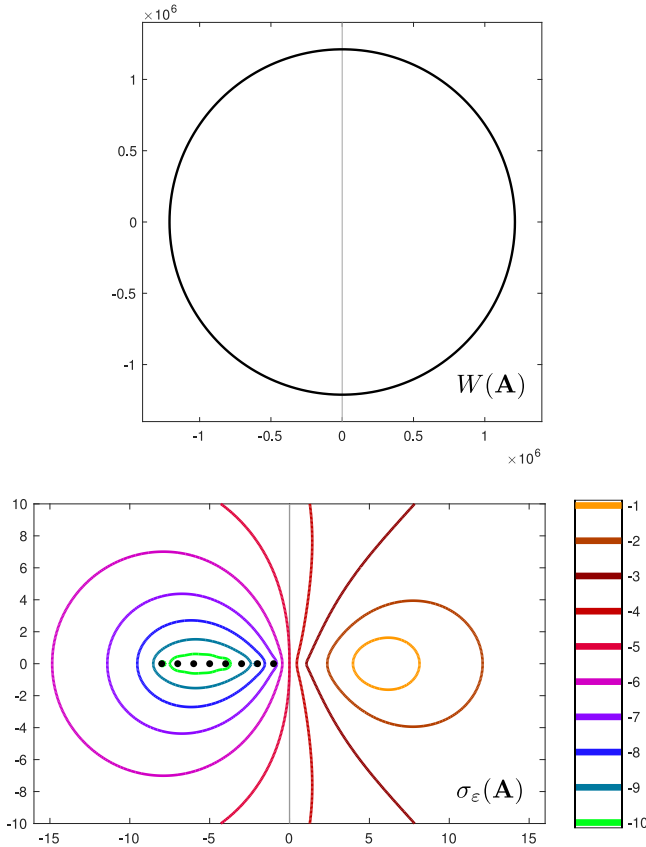
Fig. 4. The numerical range $W(\mathbf{A})$ (top) and ε-pseudospectra $\sigma_\varepsilon(\mathbf{A})$ (bottom) for the stable $\mathbf{A}$ from Example 3. Since $W(\mathbf{A})$ extends far into the right half-plane, solutions to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ can exhibit significant transient growth before asymptotic decay. In the bottom plot (computed using EigTool [18]), the lines show the boundaries of $\sigma_\varepsilon(\mathbf{A})$, with colors corresponding to $\log_{10}(\varepsilon)$. For $\varepsilon = 10^{-4}$, $\sigma_\varepsilon(\mathbf{A})$ clearly extends into the right half-plane: there exist nearby matrices $\mathbf{A} + \mathbf{E}$ that are unstable, with $\|\mathbf{E}\| < 10^{-4}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Oblique projection: adversarial construction

Like the orthogonal projection methods addressed in the previous sections, oblique projection methods approximate the true state vector $\mathbf{x}(t) \approx \mathbf{V}\widehat{\mathbf{x}}(t) \in \mathcal{V}$, for some k-dimensional subspace range$(\mathbf{V}) = \mathcal{V}$. Oblique projection methods replace the orthogonality constraint (3) with the *Petrov–Galerkin condition*

$$\mathbf{W}^*\left(\mathbf{V}\dot{\widehat{\mathbf{x}}}(t) - \left(\mathbf{A}\mathbf{V}\widehat{\mathbf{x}}(t) + \mathbf{b}u(t)\right)\right) = \mathbf{0}, \tag{18}$$

where range$(\mathbf{W}) = \mathcal{W}$ is some (generally different) k-dimensional subspace. The bases for $\mathcal{V}$ and $\mathcal{W}$ stored in the columns of $\mathbf{V}$ and $\mathbf{W}$ are now constructed to be *biorthogonal*: $\mathbf{W}^*\mathbf{V} = \mathbf{I}$. The Petrov–Galerkin constraint (18) gives the reduced system

$$\dot{\widehat{\mathbf{x}}}(t) = (\mathbf{W}^*\mathbf{A}\mathbf{V})\widehat{\mathbf{x}}(t) + (\mathbf{W}^*\mathbf{b})u(t). \tag{19}$$

The balanced truncation method (see, e.g., [1, Chapter 7], [20, Chapter 7]) fits this template, and generates models that are guaranteed to preserve stability. However, to compute the balancing biorthogonal bases $\mathbf{V}$ and $\mathbf{W}$ one must solve two Lyapunov matrix equations (typically at considerable computation expense, though algorithmic improvements make this increasingly tractable for large-scale problems; see, e.g., [21–24]).

The bi-Lanczos algorithm provides an inexpensive alternative for constructing oblique projection models. This method has the advantage that the biorthogonal bases can be computed using three-term vector recurrences that only require multiplication by $\mathbf{A}$ and $\mathbf{A}^*$ to construct biorthogonal bases for $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ and $\mathcal{K}_k(\mathbf{A}^*, \mathbf{c})$ (that form the columns of $\mathbf{V}$ and $\mathbf{W}$).

The resulting order-k bi-Lanczos ROM *matches $2k$ moments* of the transfer function, while orthogonal projection onto the Krylov subspace $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ produces a ROM that only matches k moments [1, Section 11.2]. Moreover, the three-term recurrence behind bi-Lanczos makes the bases $\mathbf{V}$ and $\mathbf{W}$ quicker to compute than the orthonormal basis for $\mathcal{V}$ required by the orthogonal projection method (which uses long recurrences in the Gram–Schmidt process).

Despite these advantages, the bi-Lanczos method often suffers from significant numerical instability. While the bases that form the columns of $\mathbf{V}$ and $\mathbf{W}$ are biorthogonal, the columns of these two matrices might themselves be quite ill-conditioned bases for $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ and $\mathcal{K}_k(\mathbf{A}^*, \mathbf{c})$. In extreme cases, the method can *break down*. More often the iterations come close to failure, exhibiting numerical instabilities; see, e.g., [25]. The problem is apparent even when $k = 1$. Suppose that $\mathbf{c}^*\mathbf{b} = 0$, as could easily occur in a physical system where the input occurs at a point far from the output measurement. In this case there exist no biorthogonal bases for $\mathcal{V} = \mathcal{K}_1(\mathbf{A}, \mathbf{b}) = \text{span}(\mathbf{b})$ and $\mathcal{W} = \mathcal{K}_1(\mathbf{A}^*, \mathbf{c}) = \text{span}(\mathbf{c})$.

When the bi-Lanczos procedure succeeds without breakdown, it produces the factorizations

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{T}_k + \gamma_k\mathbf{v}_{k+1}\mathbf{e}_k^*, \tag{20a}$$

$$\mathbf{A}^*\mathbf{W}_k = \mathbf{W}_k\mathbf{T}_k^* + \overline{\beta_k}\mathbf{w}_{k+1}\mathbf{e}_k^*; \tag{20b}$$

premultiplying (20a) by $\mathbf{W}_k^*$ and using the biorthogonality of the bases yields

$$\mathbf{W}_k^*\mathbf{A}\mathbf{V}_k = \mathbf{T}_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \gamma_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \gamma_{k-1} & \alpha_k \end{bmatrix} \in \mathbb{C}^{k \times k}.$$

The *tridiagonal* structure is inherited from the three-term recurrence relations at the heart of the bi-Lanczos process. Indeed, the jth columns of Eqs. (20a)–(20b) give

$$\gamma_j\mathbf{v}_{j+1} = \mathbf{A}\mathbf{v}_j - \alpha_j\mathbf{v}_j - \beta_{j-1}\mathbf{v}_{j-1}, \tag{21a}$$

$$\overline{\beta_j}\mathbf{w}_{j+1} = \mathbf{A}^*\mathbf{w}_j - \overline{\alpha_j}\mathbf{w}_j - \gamma_{j-1}\mathbf{w}_{j-1} \tag{21b}$$

for $j = 1, \ldots, k$, with $\beta_0 = \gamma_0 = 0$ and $\mathbf{v}_0 = \mathbf{w}_0 = \mathbf{0}$. (We assume $\gamma_j \in \mathbb{R}$, a natural choice in bi-Lanczos codes.)

6.1. Greenbaum's theorem

Few concrete results are known about the spectra of ROMs generated using the bi-Lanczos process. The most substantial insight comes from Greenbaum [26, Theorem 3] (motivated by the study of bi-Lanczos-based iterative methods for solving $\mathbf{A}\mathbf{x} = \mathbf{b}$). We will interpret this result in the context of moment-matching model reduction.

Suppose we are given a system of order n,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t),$$

for which we seek a ROM of order $k \leq n/2$. (The output vector $\mathbf{c}$ will be constructed later.)

Choose *any* parameters

$$\alpha_1, \ldots, \alpha_k, \tag{22}$$
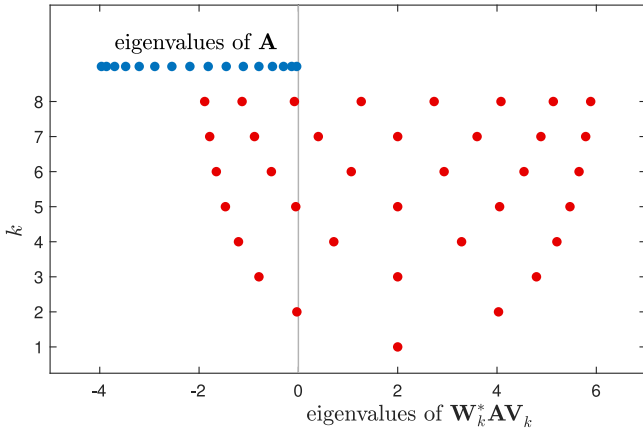
$$\beta_1, \ldots, \beta_{k-1}, \tag{23}$$

Fig. 5. Comparison of the eigenvalues of **A** (blue dots) in the left half-plane to those of the ROMs of order $k = 1, \ldots, 8$ from Example 4, generated using the bi-Lanczos algorithm that matches $2k$ moments. Despite the fact that **A** is stable and Hermitian, all of these ROMs are highly unstable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and construct $\gamma_1, \ldots, \gamma_{k-1}$ by running the recurrence

$$\widehat{\mathbf{v}}_{j+1} := \mathbf{A}\mathbf{v}_j - \alpha_j \mathbf{v}_j - \beta_{j-1}\mathbf{v}_{j-1} \tag{24a}$$

$$\gamma_j := \|\widehat{\mathbf{v}}_{j+1}\| \tag{24b}$$

$$\mathbf{v}_{j+1} := \widehat{\mathbf{v}}_{j+1}/\gamma_j, \tag{24c}$$

for $j = 1, \ldots, k$, with $\mathbf{v}_1 = \mathbf{b}/\|\mathbf{b}\|$, $\mathbf{v}_0 = \mathbf{0}$, and $\beta_0 = 0$; this recurrence is obviously meant to mimic (21a).

Theorem 5. *Consider the parameters $\{\alpha_j\}_{j=1}^k$, $\{\beta_j\}_{j=1}^{k-1}$, and $\{\gamma_j\}_{j=1}^k$ from (22), (23), and (24b), and the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$ from (24), with $k \le n/2$. Suppose $\mathbf{c} \in \mathbb{C}^n$ satisfies*

$$\mathbf{c} \perp \mathrm{span}\{\mathbf{v}_2, \ldots, \mathbf{v}_{k+1}, \mathbf{A}\mathbf{v}_{k+1}, \ldots, \mathbf{A}^{k-1}\mathbf{v}_{k+1}\}. \tag{25}$$

Then either the bi-Lanczos process breaks down, or it runs to completion and creates the ROM

$$\mathbf{W}_k^* \mathbf{A} \mathbf{V}_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \gamma_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \gamma_{k-1} & \alpha_k \end{bmatrix} \in \mathbb{C}^{k \times k}.$$

The cases of breakdown in Theorem 5 correspond, for example, to scenarios where $\beta_j = 0$ or $\gamma_j = 0$, since these values are used to normalize $\mathbf{w}_{j+1}$ and $\mathbf{v}_{j+1}$ in the Lanczos algorithm. (We shall not concern ourselves with look-ahead procedures (see, e.g., [25]) here, which can provide a work-around to many instances of breakdown.)

As Greenbaum evocatively describes it, Theorem 5 essentially says that the bi-Lanczos process can be viewed as executing an arbitrary recurrence for the first $n/2$ steps, provided one is free to select an appropriate left-starting vector $\mathbf{c}$. Since these recurrence coefficients are essential elements of the resulting ROM, Theorem 5 suggests a way to design cases where benign choices of **A** (even stable, normal or Hermitian matrices) lead to unstable ROMs. The next example gives an extreme illustration.

Example 4. Consider a stable continuous-time system with Hermitian matrix

$$\mathbf{A} = \begin{bmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix} \in \mathbb{C}^{16 \times 16},$$

and input vector $\mathbf{b} = [1, 0, \ldots, 0]^T \in \mathbb{C}^{16}$. The eigenvalues of **A** are all negative real numbers:

$$\sigma(\mathbf{A}) = \left\{ -2 + 2\cos\left(\frac{k\pi}{17}\right) : k = 1, \ldots, 16 \right\}.$$

Since **A** is Hermitian, $W(\mathbf{A})$ is the convex hull of the spectrum, and *any orthogonal projection method must produce a stable ROM.* Greenbaum's theorem shows that oblique projection methods can produce much more exotic results.

Suppose we seek a ROM of order $k = 8$, and specify the bi-Lanczos recurrence parameters

$$\alpha_1 = \cdots = \alpha_8 = 2,$$
$$\beta_1 = \cdots = \beta_7 = 1.$$

(These parameters were selected to give a reduction that was likely to be unstable.) From the three-term recurrence (24) we compute (to five digits)

$$\begin{aligned} \gamma_1 &= 4.12311, & \gamma_2 &= 3.68474, \\ \gamma_3 &= 4.12603, & \gamma_4 &= 4.31536, \\ \gamma_5 &= 4.43571, & \gamma_6 &= 4.52257, \\ \gamma_7 &= 4.58628. \end{aligned}$$

The vector $\mathbf{c}$ is then constructed, consistent with Theorem 5, by orthogonally projecting $\mathbf{Ab}$ onto the orthogonal complement of the span in (25).

Fig. 5 shows the eigenvalues of the matrix $\mathbf{W}_k^* \mathbf{A} \mathbf{V}_k$ from the resulting ROM for $k = 1, \ldots, 8$. Despite the fact that **A** is Hermitian and stable, each of the models of order $k = 1, \ldots, 8$ is *unstable*; e.g., the ROMs of order $k = 7$ and $k = 8$ both have five unstable modes.

In closing this section, we emphasize a fundamental distinction between the adversarial construction in Theorem 4 for orthogonal projection methods, and its counterpart in Theorem 5 for oblique projection. In the orthogonal case, *the user can only specify the eigenvalues of* **A**; the construction obtains pathological results by building **A** with large departure from normality, along with a corresponding **b**. In the oblique case, *the user supplies the entire matrix* **A** *and input vector* **b**. The construction in no way influences the departure of **A** from normality; it achieves its ends only by designing the output vector **c**. Thus the oblique construction is more troubling, as it can produce startling results for apparently benign **A**.

7. The unsung merits of unstable ROMs

Conventionally, an unstable ROM seems to be a poor approximation to a stable dynamical system, regardless of its other virtues (e.g., transfer functions that match moments). Missing the asymptotic character of the model is a fundamental shortcoming. Before dismissing unstable ROMs entirely, one should note that they can still give insight into the original system, especially regarding transient dynamics. We warn that efforts to suppress the instability can have the unintended consequence of compromising transient accuracy to obtain long-term qualitative agreement. We illustrate these points with two examples.

Example 5.

This example begins with a benchmark problem for the control of a flutter condition in a Boeing B-767 aircraft, contributed by Anderson, Ly, and Liu to the collection [28]. This 55×55 matrix, call it $\mathbf{A}_0$, is unstable, having a complex-conjugate pair of eigenvalues in the right half-plane. Burke, Lewis, and Overton [29] used an eigenvalue optimization algorithm to design a low-rank perturbation that stabilizes $\mathbf{A}_0$. It is this stable **A** that we investigate
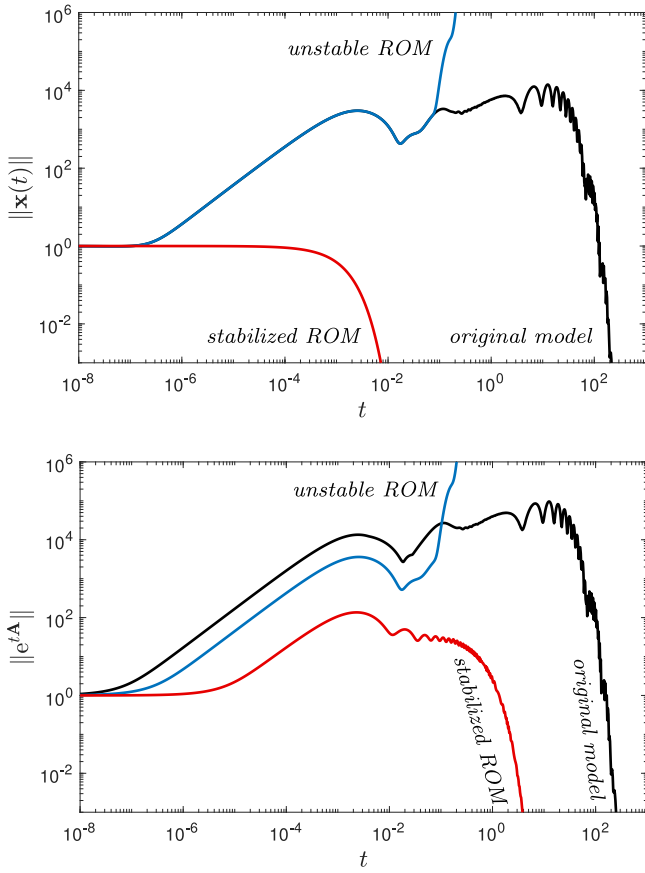
Fig. 6. Evolution of a solution $\mathbf{x}(t)$ to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ for the Boeing B-767 matrix in Example 5 (top), and the analogous plot for the solution operator $e^{t\mathbf{A}}$ (bottom). The unstable ROM provides a much more accurate impression of the dynamics at early t than does the stabilized model (both of order $k = 20$).

here; we will refer to it as the "original model" when comparing it to the ROMs we derive from it.

This **A** has spectral abscissa $\alpha(\mathbf{A}) \approx -0.07877$, but the numerical range $W(\mathbf{A})$ extends far into the right half-plane, with numerical abscissa

$$\omega(\mathbf{A}) \approx 8.4560 \times 10^6.$$

As evident from (11), solutions to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ must initially exhibit strong growth, although the system is asymptotically stable; Fig. 6 provides confirmation. The top plot in Fig. 7 shows ε-pseudospectra of **A**, which were previously investigated in [4, Chapter 15].

It suffices to simply consider $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ here, rather than the full input–output system (1)–(2). We let the initial condition

$$\mathbf{x}(0) := \mathbf{x}_0 = [1, 1, \ldots, 1]^T / \sqrt{55} \in \mathbb{C}^{55}$$

play the role that the input vector **b** normally does when constructing Krylov-based ROMs.

We select the dimension $k = 20$ for our ROMs. (The qualitative results shown here are not particularly sensitive to the choice of k and $\mathbf{x}_0$.) First we construct an orthogonal projection model using the Krylov subspace $\mathcal{V} = \mathcal{K}_k(\mathbf{A}, \mathbf{x}_0)$. The resulting $\mathbf{V}^*\mathbf{A}\mathbf{V}$ is unstable: it has 5 eigenvalues in the right half-plane, as can be seen in the bottom-left plot in Fig. 7. Of course, as Fig. 6 shows, the unstable model diverges from the stable model as $t \to \infty$, *but at earlier times, it does an excellent job of signaling the system's transient growth on the scale of* 10^3. Such growth could be significant for a motivating application, e.g., warning that a linearized model might be a poor approximation of an underlying nonlinear system.
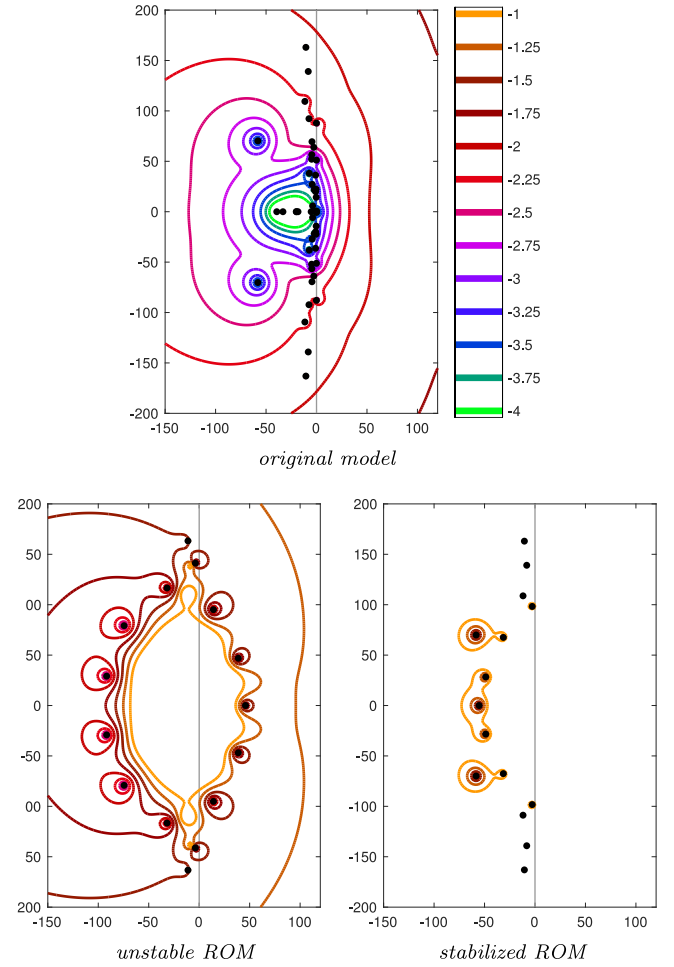

Fig. 7. For the Boeing B-767 model, $\sigma_\varepsilon(\mathbf{A})$ for the full order model (top), compared to the unstable ROM (bottom left) and its stabilized variant (bottom right), both with $k = 20$. (All plots use $\varepsilon = 10^{-1}, 10^{-1.25}, \ldots, 10^{-4}$, and were computed with Eigtool [18].) The stabilization procedure repels eigenvalues near the origin and reduces the departure from normality. (These models have a few eigenvalues beyond these axes, which do not have a major influence on the transient dynamics.) For a general investigation of the use of orthogonal Krylov projection to approximate pseudospectra, see [27].

Still, the instability in $\mathbf{V}^*\mathbf{A}\mathbf{V}$ is unappealing, and one might naturally prefer a stable ROM. To obtain stability, we follow a general approach of Grimme, Sorensen, and van Dooren (in the context of the bi-Lanczos algorithm [30]). To build an orthonormal basis for $\mathcal{K}_k(\mathbf{A}, \mathbf{x}_0)$ we use the Arnoldi algorithm [31]. After constructing the order-20 unstable model, we *restart* the Arnoldi algorithm [32], replacing the starting vector $\mathbf{x}_0$ with a "filtered" starting vector $\phi_1(\mathbf{A})\mathbf{x}_0$, where ϕ_1 is the degree-5 monic polynomial with roots at the unstable eigenvalues of $\mathbf{V}^*\mathbf{A}\mathbf{V}$. We then use $\mathcal{V} = \mathcal{K}_k(\mathbf{A}, \phi_1(\mathbf{A})\mathbf{x}_0)$ to construct a new ROM, which now only has only 3 unstable modes. Repeat the process: let the cubic polynomial ϕ_2 have roots at the 3 new unstable modes, and add $\phi_2(\mathbf{A})$ to the filter. The next space $\mathcal{V} = \mathcal{K}_k(\mathbf{A}, \phi_2(\mathbf{A})\phi_1(\mathbf{A})\mathbf{x}_0)$ has just 1 unstable mode, which we make the root of the linear polynomial ϕ_3. Finally, $\mathcal{V} = \mathcal{K}_k(\mathbf{A}, \Phi(\mathbf{A})\mathbf{x}_0)$ delivers a stable order-20 ROM, where $\Phi(z) := \phi_3(z)\phi_2(z)\phi_1(z)$ has roots at all the previously encountered unstable modes.

The relevant eigenvalues of this new stabilized ROM can be seen in the bottom-right plot of Fig. 7. Notice that the polynomial filter, with its roots at the unstable eigenvalues (including the 5 unstable modes in the bottom-left plot of Fig. 7), effectively deters the ROM from having modes near the origin. Moreover, the stabilization
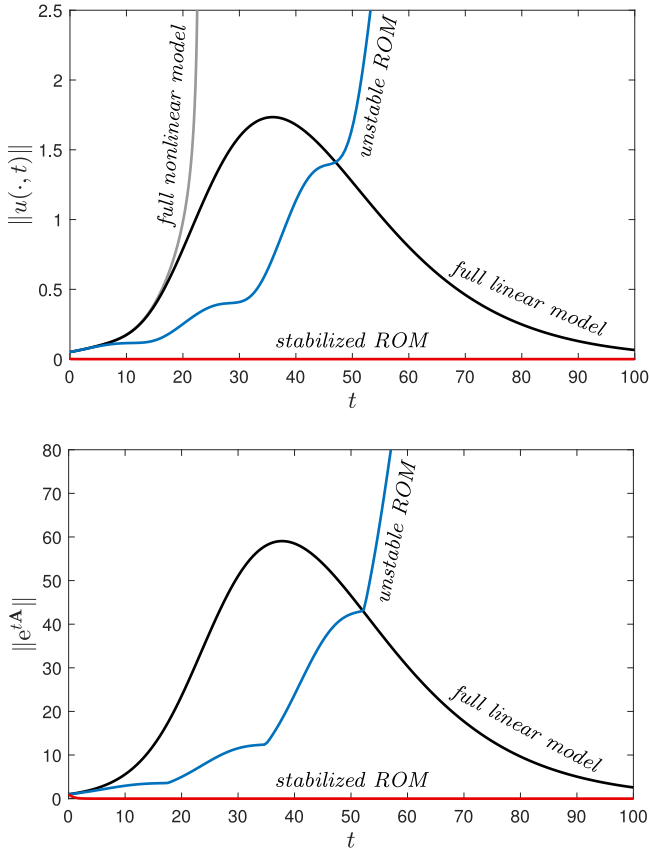
Fig. 8. Evolution of a solution $u(x, t)$ to the nonlinear heat model and its linearization from Example 6 (top), and the analogous plot for the solution operator e^{tA} for the linearized operator and two linear ROMs derived from it (bottom). Again, the unstable ROM provides a more accurate impression of the dynamics at early t than does the stabilized model (both of order $k = 40$).

process has suppressed the departure from normality, as reflected in a diminished numerical abscissa:

$$
\begin{aligned}
\text{original model:} &\quad \omega(\mathbf{A}) \approx 8.45603 \times 10^6; \\
\text{unstable ROM:} &\quad \omega(\mathbf{V}^*\mathbf{A}\mathbf{V}) \approx 2.24872 \times 10^6; \\
\text{stabilized ROM:} &\quad \omega(\mathbf{V}^*\mathbf{A}\mathbf{V}) \approx 8.90589 \times 10^4.
\end{aligned}
$$

The stabilized system still exhibits transient growth for some initial conditions (see the bottom plot of Fig. 6), but far less than the original model.

Example 6. Consider the nonlinear heat equation

$$
u_t(x, t) = u_{xx}(x, t) + u_x(x, t) + \tfrac{1}{8}u(x, t) + u^3(x, t), \tag{26}
$$

posed on the domain $x \in (0, \ell) \subset \mathbb{R}$ and $t \geq 0$, with homogeneous Dirichlet boundary conditions $u(0, t) = u(\ell, t) = 0$. We are interested in initial conditions $u_0(x) = u(x, 0) \in H_0^1(0, \ell)$ that are small in norm. Eq. (26) was studied by Sandstede and Scheel [33], who analyzed stability properties of the trivial solution $u \equiv 0$ on the finite domain $x \in (0, \ell)$ versus the infinite domains $x \in (0, \infty)$ and $x \in (-\infty, \infty)$. (On finite domains, sufficiently small initial conditions u_0 give solutions $u(x, t)$ for which $\|u(\cdot, t)\|_{H_0^1(0, \ell)} \to 0$ as $t \to \infty$; in contrast, $u \equiv 0$ is unstable on the infinite domains $x \in (0, \infty)$ and $x \in (-\infty, \infty)$. This discrepancy suggests that moderately small values of u_0 can exhibit interesting behavior on finite domains. Galkowski generalized this model, drawing a connection between the dynamics and pseudospectra of the linear part of the model [34].)
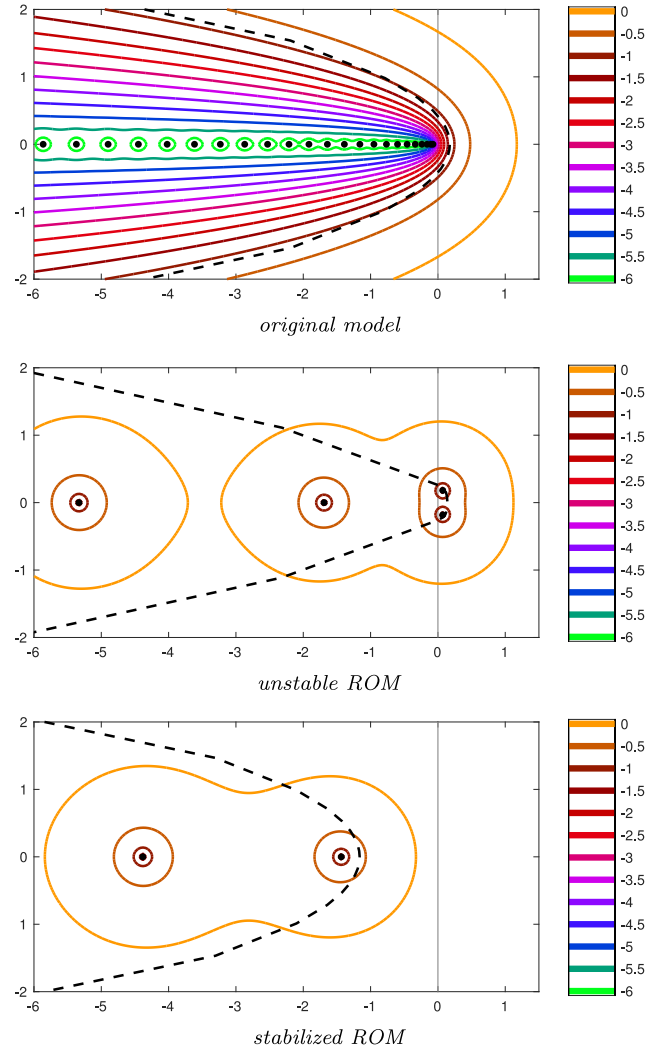

Fig. 9. For the linearized heat model, the rightmost part of $\sigma_\varepsilon(\mathbf{A})$ for the full order discretization (top) and order $k = 40$ ROMs (middle and bottom). (All plots use the same scale $\varepsilon = 10^0, 10^{-0.5}, \dots, 10^{-6}$, and were computed with EigTool [18].) The dashed black curves denote the boundaries of the numerical range; the stabilization procedure drives both eigenvalues and $W(\mathbf{V}^*\mathbf{A}\mathbf{V})$ too far to the left.

We take a domain of length $\ell = 30$ and discretize the system (26) using a Chebyshev pseudospectral collocation method, based on codes and techniques from Trefethen [35]. Fig. 8 shows the evolution of $\|u(\cdot, t)\|_{L^2(0, \ell)}$ for the initial condition

$$
u_0(x, t) = 10^{-5}x(x - \ell)(x - \ell/2),
$$

along with the solution of the analogous linear problem that omits the u^3 term in (26). The linear model is stable, but experiences transient growth; this growth gradually increases the contribution of the u^3 term in the nonlinear model, leading to apparent divergence. Fig. 9 shows $\sigma_\varepsilon(\mathbf{A})$ (in the $L^2(0, \ell)$ norm) for this discretized, linearized operator of order $n = 127$. To obtain a ROM for this linearized part of the problem, we transform coordinates so the vector Euclidean norm approximates the $L^2(0, \ell)$ norm, then compute the associated Krylov orthogonal projection ROM. The resulting order $k = 40$ ROM has a complex conjugate pair of unstable eigenvalues. Following the stabilization procedure described in Example 5, one iteration of the restarted Arnoldi method with a filtered starting vector yields a stabilized model. As with Example 5, the unstable ROM does a better job of capturing the transient growth of the linear system. While the unstable ROM does not qualitatively match the asymptotic behavior of the *linear system*, it is more consistent with the apparent divergence of the true *nonlinear system*.

8. Conclusion

This note has collected a number of results that give insight into the unstable modes that can arise in projection-based reduced-order models. The illustrative examples have been intentionally small in scale to make simple points, but the implications for large-scale systems are evident. While the bounds in Theorems 1 and 3 limit the location and number of unstable modes, these bounds can undoubtedly be sharpened with further analysis.

Acknowledgments

References

[1] A.C. Antoulas, Approximation of Large-Scale Dynamical Systems, SIAM, Philadelphia, 2005.

[2] K. Kunisch, S. Volkwein, Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics, SIAM J. Numer. Anal. 40 (2002) 492–515.

[3] M. Rathinam, L.R. Petzold, A new look at proper orthogonal decomposition, SIAM J. Numer. Anal. 41 (2003) 1893–1925.

[4] L.N. Trefethen, M. Embree, Spectra and Pseudospectra: The Behavior of Non-normal Matrices and Operators, Princeton University Press, Princeton, NJ, 2005.

[5] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1991.

[6] D. Hinrichsen, B. Kelb, Spectral value sets: A graphical tool for robustness analysis, Systems Control Lett. 21 (1993) 127–136.

[7] M. Karow, Geometry of spectral value sets (Ph.D. thesis), Universität Bremen, 2003.

[8] S.M. Rump, Eigenvalues, pseudospectrum and structured perturbations, Numer. Linear Algebra Appl. 413 (2006) 567–593.

[9] R. Byers, D. Kressner, On the condition of a complex eigenvalue under real perturbations, BIT 44 (2004) 209–214.

[10] T.A. Manteuffel, G. Starke, On hybrid iterative methods for nonsymmetric systems of linear equations, Numer. Math. 73 (1996) 489–506.

[11] M. Embree, B. Keeler, Pseudospectra of matrix pencils for transient analysis of differential–algebraic equations, SIAM J. Matrix Anal. Appl. 38 (2017) 1028–1054.

[12] R.A. Horn, C.R. Johnson, Matrix Analysis, second ed., Cambridge University Press, Cambridge, 2013.

[13] R.L. Carden, M. Embree, Ritz value localization for non-Hermitian matrices, SIAM J. Matrix Anal. Appl. 33 (2012) 1320–1338.

[14] G.D. Smith, Numerical Solution of Partial Differential Equations: Finite Difference Methods, third ed., Oxford University Press, Oxford, 1985.

[15] J. Duintjer Tebbens, G. Meurant, Any Ritz value behavior is possible for Arnoldi and GMRES, SIAM J. Matrix Anal. Appl. 33 (2012) 958–978.

[16] A. Greenbaum, Z. Strakoš, Matrices that generate the same Krylov residual spaces, in: G. Golub, A. Greenbaum, M. Luskin (Eds.), Recent Advances in Iterative Methods, Springer-Verlag, New York, 1994, pp. 95–118.

[17] A. Greenbaum, V. Pták, Z. Strakoš, Any nonincreasing convergence curve is possible for GMRES, SIAM J. Matrix Anal. Appl. 17 (1996) 465–469.

[18] T.G. Wright, EigTool, 2002, software available at https://github.com/eigtool.

[19] V. Mehrmann, H. Xu, An analysis of the pole placement problem. I. The single input case, Electron. Trans. Numer. Anal. 4 (1996) 89–105.

[20] K. Zhou, Robust and Optimal Control, Prentice Hall, Upper Saddle River, NJ, 1995, with John C. Doyle and Keith Glover.

[21] P. Benner, J. Saak, Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: A state of the art survey, GAMM-Mitteilungen 36 (2013) 32–52.

[22] T. Penzl, A cyclic low-rank Smith method for large sparse Lyapunov equations, SIAM J. Sci. Comput. 21 (2000) 1401–1418.

[23] J. Sabino, Solution of Large-Scale Lyapunov Equations Via the Block Modified Smith Method (Ph.D. thesis), Rice University, 2006.

[24] V. Simoncini, Computational methods for linear matrix equations, SIAM Rev. 58 (2016) 377–441.

[25] B.N. Parlett, D.R. Taylor, Z.A. Liu, A look-ahead Lanczos algorithm for unsymmetric matrices, Math. Comp. 44 (1985) 105–124.

[26] A. Greenbaum, On the role of the left starting vector in the two-sided Lanczos algorithm and nonsymmetric linear system solvers, in: D.F. Griffiths, D.J. Higham, G.A. Watson (Eds.), Numerical Analysis 1997, Addison Wesley Longman, Harlow, Essex, UK, 1998, pp. 124–132.

[27] K.-C. Toh, L.N. Trefethen, Calculation of pseudospectra by the Arnoldi iteration, SIAM J. Sci. Comput. 17 (1996) 1–15.

[28] E.J. Davison (Ed.), Benchmark Problems for Control System Design: Report of the IFAC Theory Committee, International Federation of Automatic Control, 1990.

[29] J.V. Burke, A.S. Lewis, M.L. Overton, A nonsmooth, nonconvex optimization approach to robust stabilization by static output feedback and low-order controllers, in: Proceedings of ROCOND 2003, Milan, 2003.

[30] E.J. Grimme, D.C. Sorensen, P. van Dooren, Model reduction of state space systems via an implicitly restarted Lanczos method, Numer. Algorithms 12 (1995) 1–31.

[31] Y. Saad, Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices, Numer. Linear Algebra Appl. 34 (1980) 269–295.

[32] D.C. Sorensen, Implicit application of polynomial filters in a k-step Arnoldi method, SIAM J. Matrix Anal. Appl. 13 (1992) 357–385.

[33] B. Sandstede, A. Scheel, Basin boundaries and bifurcations near convective instabilities: A case study, J. Differential Equations 208 (2005) 176–193.

[34] J. Galkowski, Nonlinear instability in a semiclassical problem, Comm. Math. Phys. 316 (2012) 705–722.

[35] L.N. Trefethen, Spectral Methods in MATLAB, SIAM, Philadelphia, 2000.